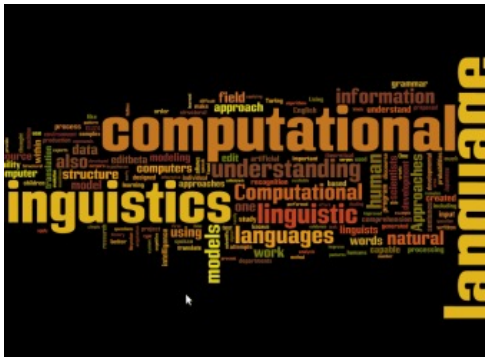
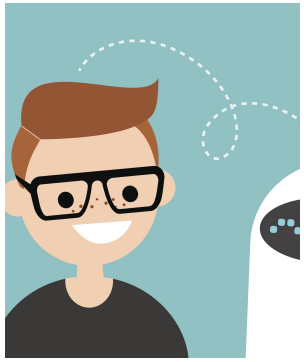




Lecture 8

Corpus analysis and computational linguistics

- © **Kushneruk Svetlana Leonidovna**

- Doctor of Philology, Professor of
- Chelyabinsk State University

1. Corpus analyses of vocabulary

a set of lexical forms having the **same stem** and belonging to the **same** major **word class**, differing only in inflection and/or spelling

a base form of a word + its inflected forms + derivatives

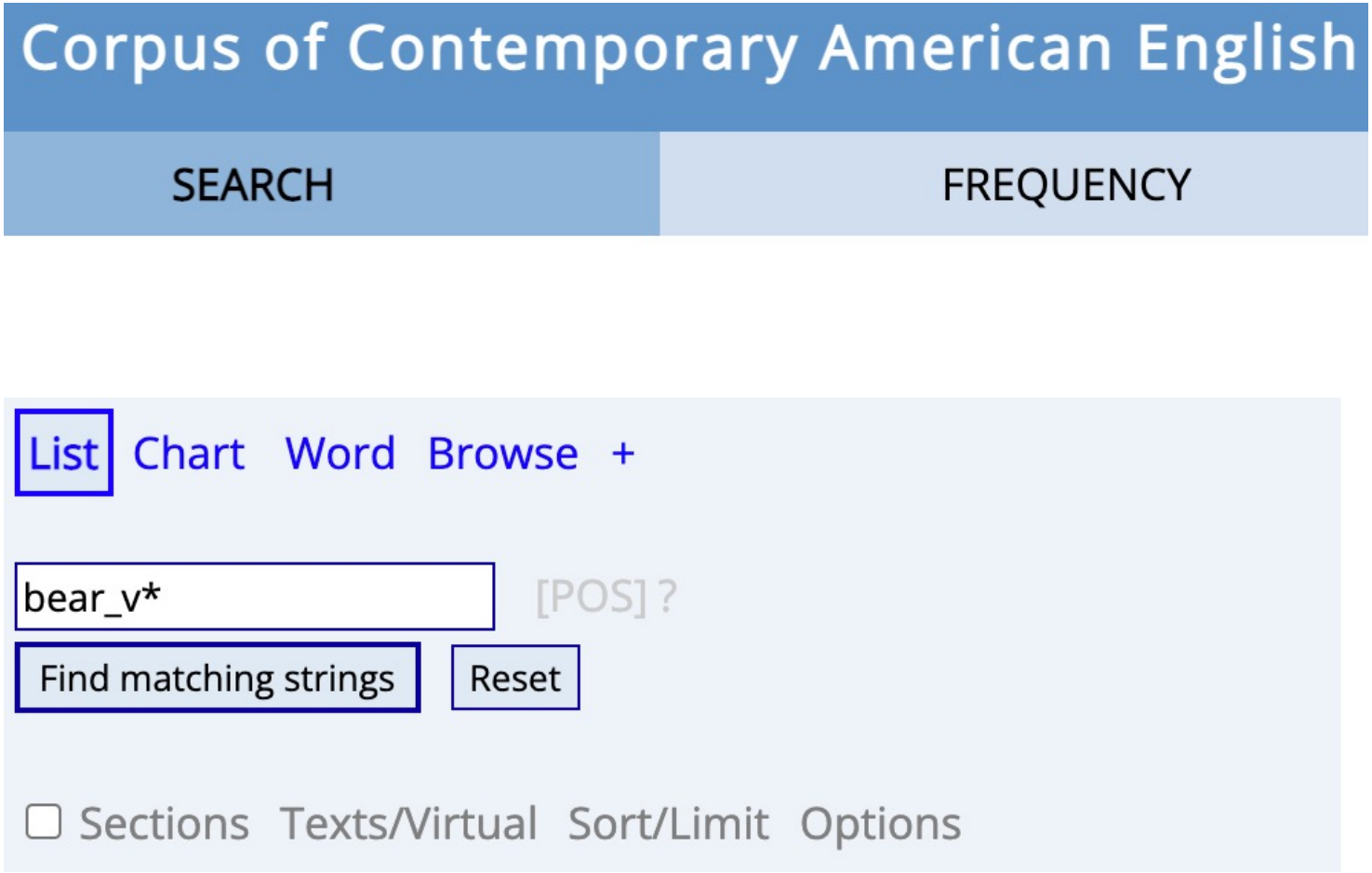
- LEMMA : inflected forms of the verb
  - *develops, developed, developing*
- WORD FAMILY additionally includes representatives of other word classes
  - adjectives: *undeveloped, underdeveloped*
  - nouns: *development, developments, developer, developers*

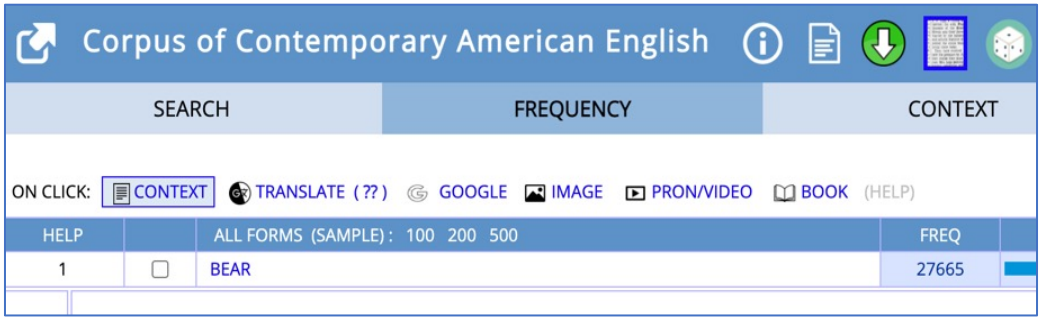
1.1. Polysemy

Polysemy – the existence of several meanings in an individual word

The more frequent a word is, the more meanings it is likely to have

the polysemy of the word *bear*





Corpus of Contemporary American English

SEARCH | FREQUENCY | CONTEXT | ACCOUNT

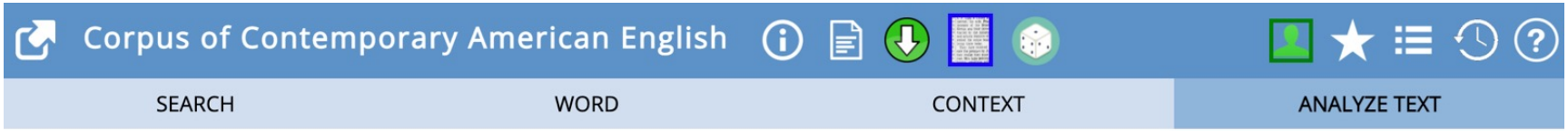
CLICK FOR MORE CONTEXT | NEW | SAVE | TRANSLATE | ANALYZE

1	2012	BLOG	addictinginfo.org		, my friends, if you are (wo) man enough to grin and bear it, can be very enlightening. # *rimshot* Try the veal, I'll
2	2012	BLOG	dailypaul.com		Here's a little help researching the Bronfmans: http: **35;1818;TOOLONG... # And bear in mind, many of the really big names go to great effort to make
3	2012	BLOG	...tionalgeographic.com		, we generally take precautions to avoid interactions or attacks (hiking in groups, bear spray, a gun, etc). However, in other places in the
4	2012	BLOG	...tionalgeographic.com		comparison, during the 20th century there have been 71 fatal grizzly (brown) bear attacks in North America. Each year in the United States, 16-18 people d
5	2012	BLOG	...uality.wordpress.com		pointless discrimination hurts all of us, from those impacted directly to those helping to bear the burden of the societal costs. # As much as Mitt Romney a
6	2012	BLOG	freethoughtblogs.com		God's. Men and women are made in the image of God, they bear his image... so therefore they owe a something to God. That's the
8	2012	BLOG	core77.com (1)		probably for legal reasons, so I will not quote anyone directly. Also bear in mind I'm going off of an anecdote I heard only once, about
9	2012	BLOG	...mbreen.wordpress.com		out concerns, even if we do not see the risks ourselves. We must bear in mind that perpetrators are often charming, sociable characters and just because
10	2012	BLOG	...ithchip.blogspot.com		are " traders, like everyone else in a free society, and they should bear that title proudly, considering the crucial importance of the services they offer. "
11	2012	BLOG	...ocetah.wordpress.com		something like " Yanaweti " or " Yonaweti, " from " yonah ", bear, and " uweti, " old. The actual pronunciation would have been something
12	2012	BLOG	...esternjournalism.com		# We have an Ace-in-the-Hole. Just one. He has no credentials that can bear scrutiny. His " sealed " past has to be Un-Sealed by a military force
13	2012	BLOG	...nstructionlitmag.com		Knicks. If his early returns -- his signature three-pointers, a positive attitude -- bear out over the course of the season, he may end up usurping the title
14	2012	BLOG	blackwindmetal.com		feelings I got from each album; The Time Of The Oath just seems to bear the mark of a confident band. Again, it's not all too different
15	2012	BLOG	blog.adw.org		adjustment. # Some of us who got married were simply unable to conceive and bear children. I don't think that absolved us from the sacramental respons
16	2012	BLOG	blog.adw.org		their love can be a sacrament of Christ's love for the Church which can bear fruit in other ways. I know you said May's argument was directed at
17	2012	BLOG	blog.adw.org		the form it does does not mean that couples who for some reason can not bear a child are excluded. It only means that they can not purposefully, or
18	2012	BLOG	...le-east.blogspot.com		US State Department should ask the Congress to allocate funds so that these efforts could bear fruition. While the term has been overused, the need to w
19	2012	BLOG	joyfullyjobless.com		the home office gallery gathered by Judy Heminsley. You'll see wonderful environments that bear no resemblance to a cubicle. 3 Responses to A Sense of
20	2012	BLOG	...yabsoff.blogspot.com		interest in kissing the un-cute boy in front of me. # i couldn't bear to let the opportunity pass me by altogether, so i turned to my best

1.2. Synonymy as a lexical relation

Synonymy – a bilateral or symmetrical sense relation in which two or more linguistic forms have similar meaning: *issue* and *problem*

no two words can be considered perfect synonyms as **corpus data** reveal *important differences* in the phraseological patterns



-SAMPLES- MY TEXTS

SEARCH CLEAR

HELP	Compare to previous WordAndPhrase (PDF)
	Overview
	Word-oriented functions
	Phrase-oriented functions

You can always get to this page by clicking this icon above.

You can enter any text that you would like in the form at the left -- for example, a paper that you've written, or a newspaper article that you've copied from another website. After inputting the text, you can then see useful information about words and phrases in that text, based on data from COCA.

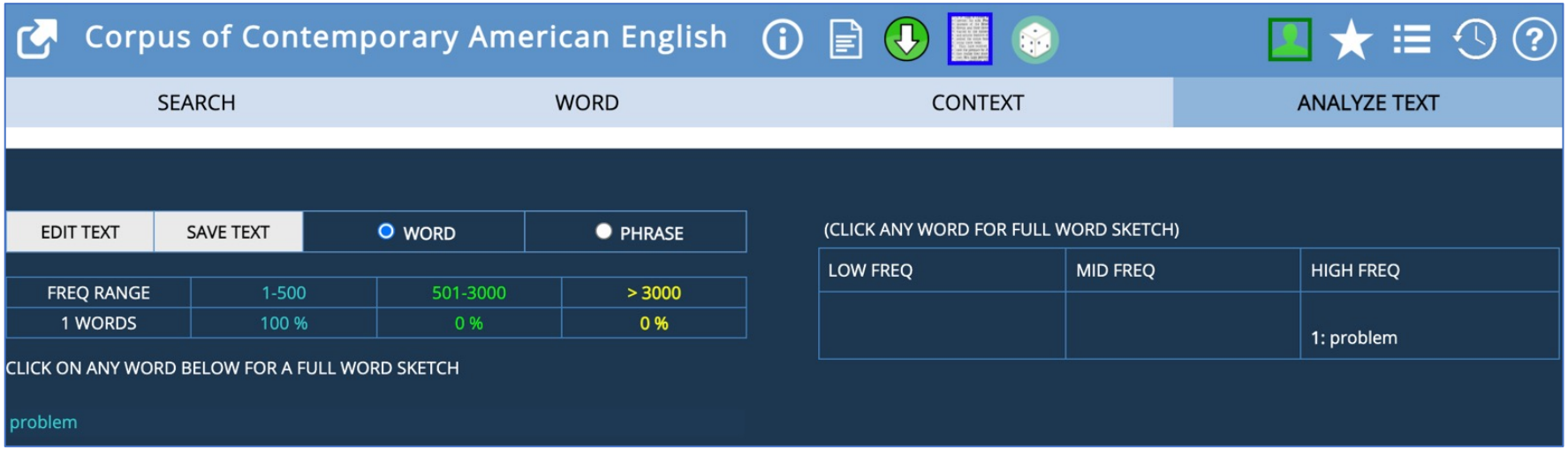
First, it will highlight all of the medium and lower-frequency words in your text and create lists of these words that you can use offline. This frequency data can help language learners focus on new words, and it can allow you to see "what the text is about" (i.e. text-specific words).

Second, you can click through the words in the text to see a detailed "word sketch" of any of the words -- showing their definition and their translation (in more than 100 languages); links to pronunciation, images, and videos; related topics, collocates, "clusters" (2, 3, and 4-word phrases); and concordance lines.

Finally, you can do powerful searches on selected phrases in your text, to show related phrases in COCA. In this way, this resource is like a "collocational thesaurus" to see what related phrases are most likely in different styles of English.

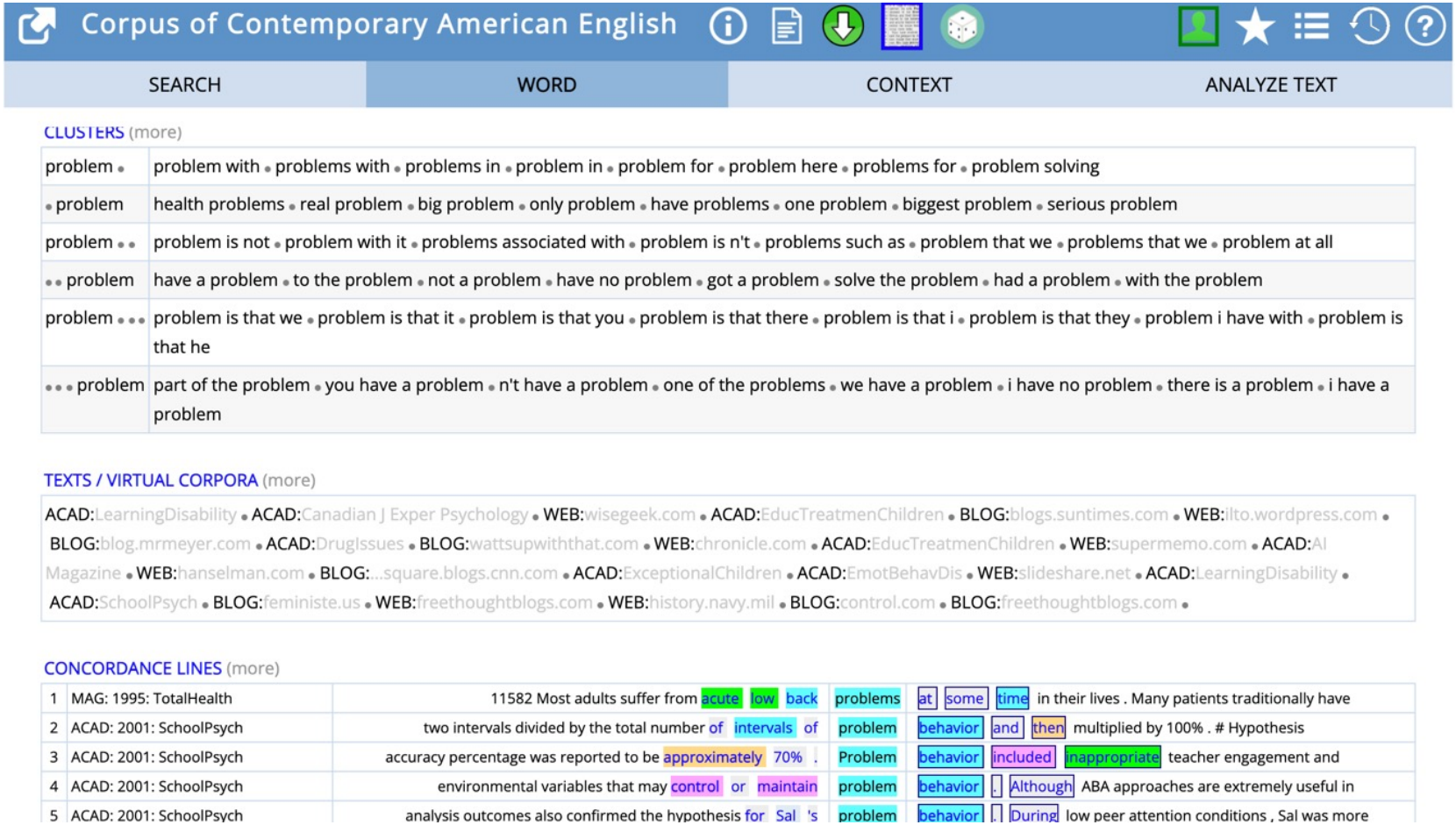
Just enter some text (or analyze a random text via SAMPLES), and there will be more help files on the next page.

TRUMP inaugural address


Corpus of Contemporary American English
SEARCH
WORD
CONTEXT
ANALYZE TEXT
EDIT TEXT
SAVE TEXT
WORD
PHRASE
FREQ RANGE
1-500
501-3000
> 3000
1 WORDS
100 %
0 %
0 %
CLICK ON ANY WORD BELOW FOR A FULL WORD SKETCH
problem
(CLICK ANY WORD FOR FULL WORD SKETCH)
LOW FREQ
MID FREQ
HIGH FREQ
1: problem


Corpus of Contemporary American English
SEARCH
WORD
CONTEXT
ANALYZE TEXT
See in iWeb
Collocates
Clusters
Topics
Texts
KWIC
HELP
problem
(NOUN)
#187
BLOG
WEB
TV/M
SPOK
FIC
MAG
NEWS
ACAD
1. a state of difficulty that needs to be resolved 2. a source of difficulty 3. a question raised for consideration or solution
D M O C G E
YouGlish
PlayPhrase
Yarn
Translate: choose language
SYNONYMS
NEW: DEFIN +SPEC +GENL
difficulty difficulty, drawback, glitch, hindrance, hitch, obstacle, obstruction, problem, setback, snag puzzle conundrum, enigma, poser, problem, puzzle, question, riddle
TOPICS (more)
solve, solution, eg, student, solving, fix, ie, symptom, treatment, cause, variable, difficulty, mathematical, computer, correlate, behavior, condition, finding, classroom, score
COLLOCATES (more)
NOUN solution, solving, america, us, alcohol, drinking, obama, pollution
VERB solve, cause, face, address, fix, resolve, arise, pose
ADJ serious, environmental, mental, severe, fundamental, behavioral, psychological, underlying
ADV eg, ie, therein, head-on, twofold
RELATED WORDS
problematic, problem-solving, problem-solver, problem-based, problem-solve, problematical, problem-free, unproblematic, problematically, problematic, sub-problem

Information about clusters, text types and concordance lines



SYNONYMS

	RANK	FREQ	Word	PoS	Audio	Video	Image	LANG?
1	187	504175	problem	NOUN				
2	206	457301	question	NOUN				
3	972	99455	challenge	NOUN				
4	2039	43423	difficulty	NOUN				
5	4099	16342	obstacle	NOUN				
6	4501	14181	catch	NOUN				
7	4652	13476	puzzle	NOUN				
8	8052	5407	setback	NOUN				
9	8292	5126	obstruction	NOUN				
10	10043	3689	drawback	NOUN				
11	11643	2850	glitch	NOUN				
12	12058	2688	riddle	NOUN				
13	14198	1969	conundrum	NOUN				
14	14667	1841	hitch	NOUN				
15	15413	1669	enigma	NOUN				
16	18376	1171	snag	NOUN				
17	18646	1133	hindrance	NOUN				
18	29329	392	poser	NOUN				

issue

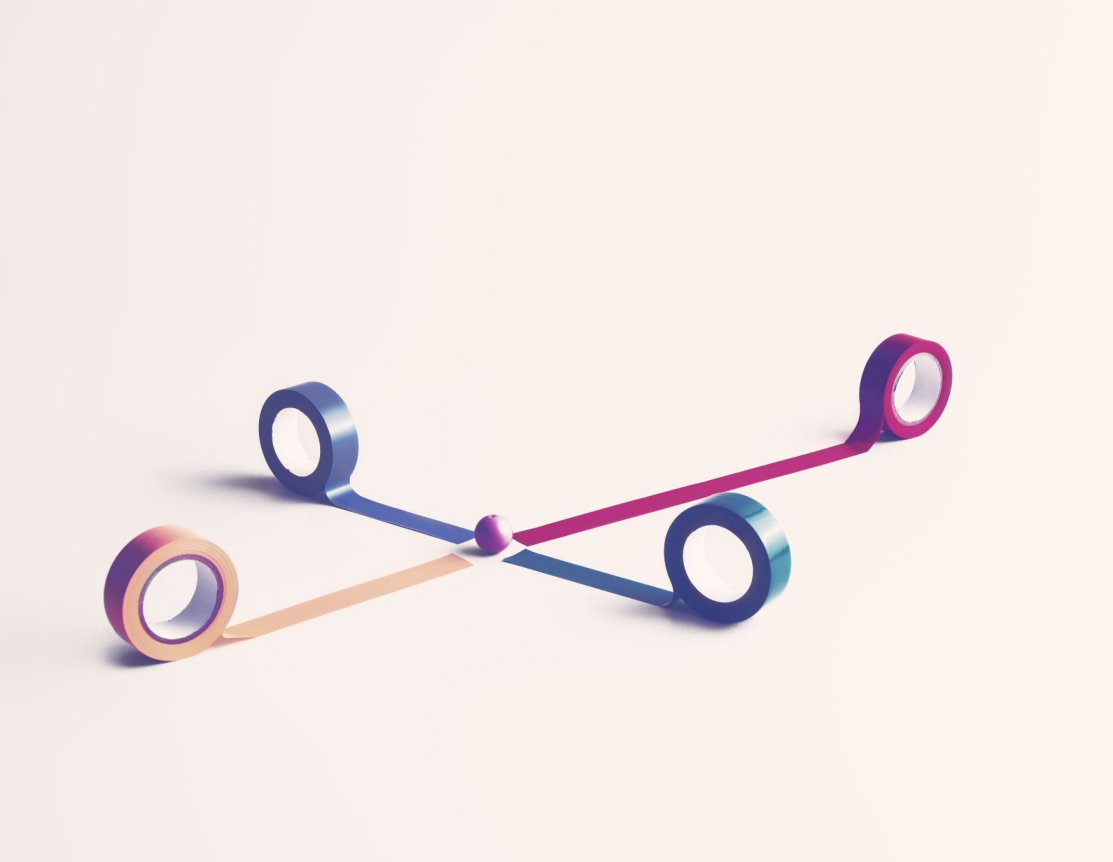
1	187	504175	problem	NOUN				
2	206	457301	question	NOUN				
3	211	444412	number	NOUN				
4	261	362137	issue	NOUN				
5	575	167351	matter	NOUN				
6	803	120773	subject	NOUN				
7	876	110976	concern	NOUN				
8	1037	93507	production	NOUN				
9	1610	58410	topic	NOUN				
10	1865	48365	copy	NOUN				
11	1878	47967	release	NOUN				
12	2322	36338	distribution	NOUN				
13	2394	34656	publication	NOUN				
14	2494	32842	edition	NOUN				
15	2750	28903	delivery	NOUN				
16	3369	22027	dispute	NOUN				
17	5891	9267	circulation	NOUN				
18	6980	6884	offspring	NOUN				
19	9082	4422	installment	NOUN				
20	16233	1512	issuance	NOUN				

problem

	RANK	FREQ	Word
1	187	504175	problem
2	206	457301	question
3	972	99455	challenge
4	2039	43423	difficulty
5	4099	16342	obstacle
6	4501	14181	catch
7	4652	13476	puzzle
8	8052	5407	setback
9	8292	5126	obstruction
10	10043	3689	drawback
11	11643	2850	glitch
12	12058	2688	riddle
13	14198	1969	conundrum
14	14667	1841	hitch
15	15413	1669	enigma
16	18376	1171	snag
17	18646	1133	hindrance
18	29329	392	poser

issue

1	187	504175	problem
2	206	457301	question
3	211	444412	number
4	261	362137	issue
5	575	167351	matter
6	803	120773	subject
7	876	110976	concern
8	1037	93507	production
9	1610	58410	topic
10	1865	48365	copy
11	1878	47967	release
12	2322	36338	distribution
13	2394	34656	publication
14	2494	32842	edition
15	2750	28903	delivery
16	3369	22027	dispute
17	5891	9267	circulation
18	6980	6884	offspring
19	9082	4422	installment
20	16233	1512	issuance

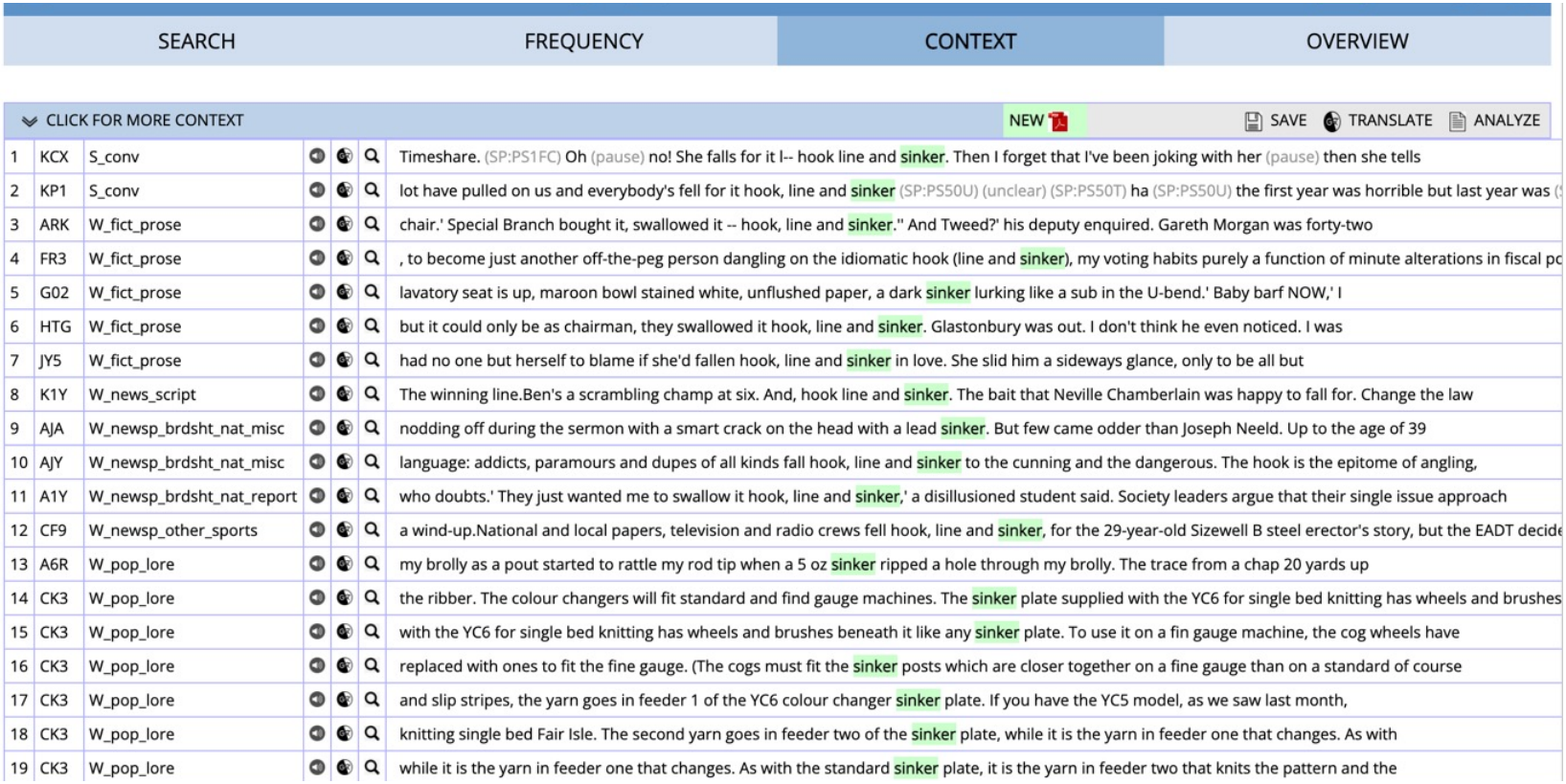


1.3. Metaphoricity and idiomaticity of words

Metaphoric meaning is a type of extended meaning when words are used in an extended non-literal way to explain something in a comparative relationship

hook, line and sinker

I fell for his story hook, line and sinker.



SEARCH | FREQUENCY | CONTEXT | OVERVIEW

CLICK FOR MORE CONTEXT — NEW — SAVE — TRANSLATE — ANALYZE

1	KCX	S_conv	Timeshare. (SP:PS1FC) Oh (pause) no! She falls for it l-- hook line and sinker. Then I forget that I've been joking with her (pause) then she tells
2	KP1	S_conv	lot have pulled on us and everybody's fell for it hook, line and sinker (SP:PS50U) (unclear) (SP:PS50T) ha (SP:PS50U) the first year was horrible but last year was (
3	ARK	W_fict_prose	chair.' Special Branch bought it, swallowed it -- hook, line and sinker." And Tweed?' his deputy enquired. Gareth Morgan was forty-two
4	FR3	W_fict_prose	, to become just another off-the-peg person dangling on the idiomatic hook (line and sinker), my voting habits purely a function of minute alterations in fiscal pc
5	G02	W_fict_prose	lavatory seat is up, maroon bowl stained white, unflushed paper, a dark sinker lurking like a sub in the U-bend.' Baby barf NOW,' I
6	HTG	W_fict_prose	but it could only be as chairman, they swallowed it hook, line and sinker. Glastonbury was out. I don't think he even noticed. I was
7	JY5	W_fict_prose	had no one but herself to blame if she'd fallen hook, line and sinker in love. She slid him a sideways glance, only to be all but
8	K1Y	W_news_script	The winning line.Ben's a scrambling champ at six. And, hook line and sinker. The bait that Neville Chamberlain was happy to fall for. Change the law
9	AJA	W_newsp_brdsht_nat_misc	nodding off during the sermon with a smart crack on the head with a lead sinker. But few came odder than Joseph Neeld. Up to the age of 39
10	AJY	W_newsp_brdsht_nat_misc	language: addicts, paramours and dupes of all kinds fall hook, line and sinker to the cunning and the dangerous. The hook is the epitome of angling,
11	A1Y	W_newsp_brdsht_nat_report	who doubts.' They just wanted me to swallow it hook, line and sinker,' a disillusioned student said. Society leaders argue that their single issue approach
12	CF9	W_newsp_other_sports	a wind-up.National and local papers, television and radio crews fell hook, line and sinker, for the 29-year-old Sizewell B steel erector's story, but the EADT decide
13	A6R	W_pop_lore	my brolly as a pout started to rattle my rod tip when a 5 oz sinker ripped a hole through my brolly. The trace from a chap 20 yards up
14	CK3	W_pop_lore	the ribber. The colour changers will fit standard and find gauge machines. The sinker plate supplied with the YC6 for single bed knitting has wheels and brushes
15	CK3	W_pop_lore	with the YC6 for single bed knitting has wheels and brushes beneath it like any sinker plate. To use it on a fin gauge machine, the cog wheels have
16	CK3	W_pop_lore	replaced with ones to fit the fine gauge. (The cogs must fit the sinker posts which are closer together on a fine gauge than on a standard of course
17	CK3	W_pop_lore	and slip stripes, the yarn goes in feeder 1 of the YC6 colour changer sinker plate. If you have the YC5 model, as we saw last month,
18	CK3	W_pop_lore	knitting single bed Fair Isle. The second yarn goes in feeder two of the sinker plate, while it is the yarn in feeder one that changes. As with
19	CK3	W_pop_lore	while it is the yarn in feeder one that changes. As with the standard sinker plate, it is the yarn in feeder two that knits the pattern and the

- British National Corpus (BNC)
- concordances for the word '*sinker*'

quantitative analysis of frequencies of words + interpretative discourse-oriented analysis

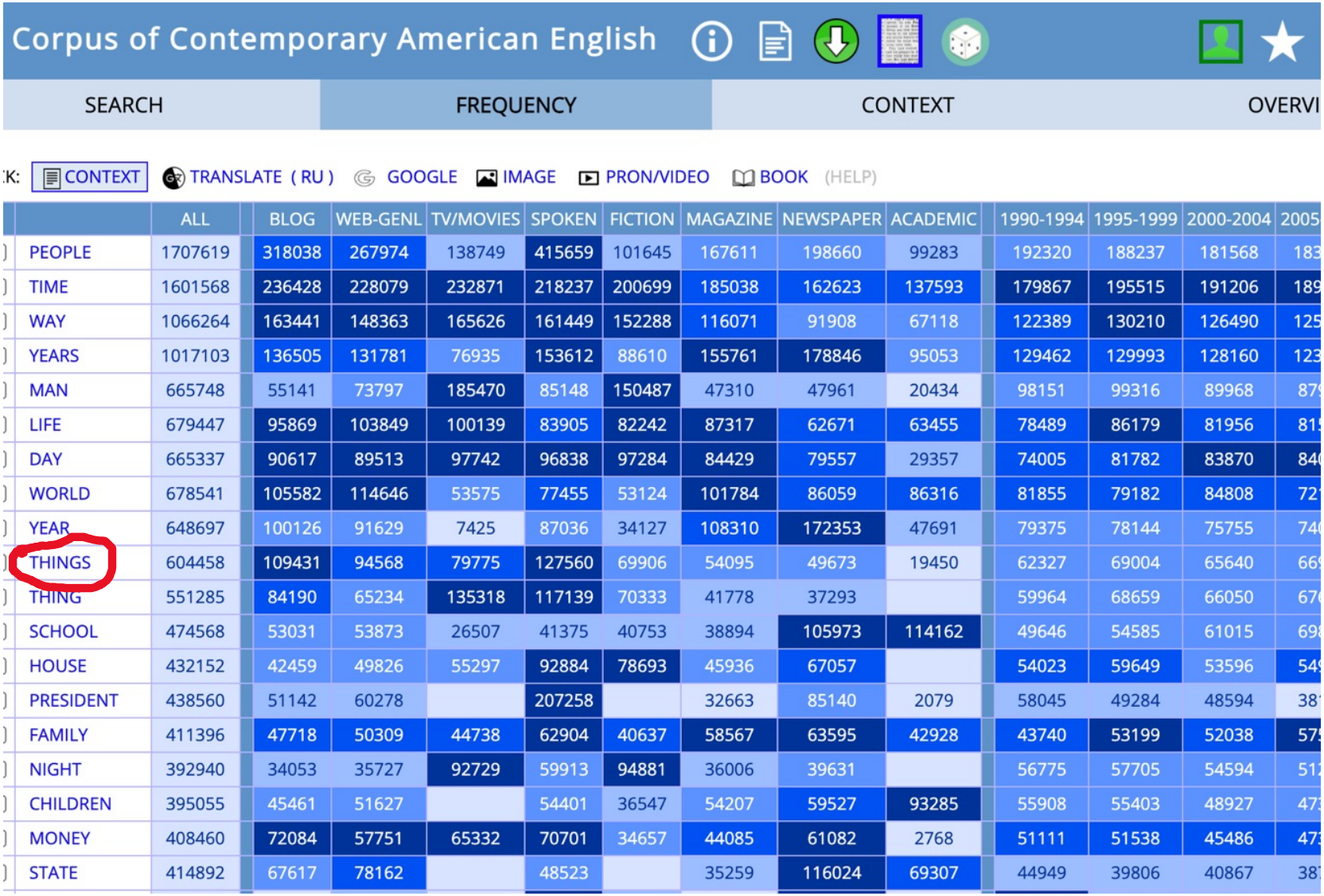
Computers and concordancers are not able to identify idiomatic meanings of words and therefore ***automatic searches need to be combined with qualitative forms of analysis*** based on the reading of concordance lines

1.4. Register variation: jargon, slang and appropriateness

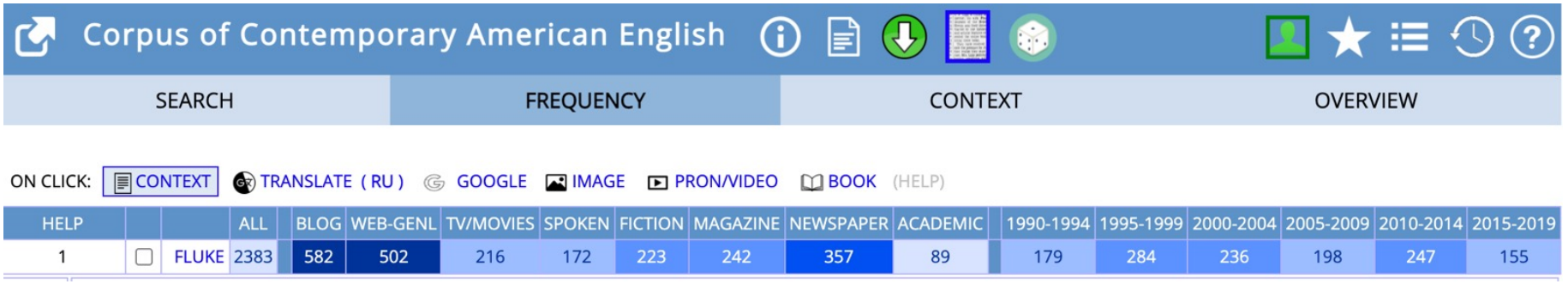
Register

it focuses on the varied use of language as dependent on factors such as communicative situation, formality level, the age and gender of language users and relationships between them

The most frequent nouns across different registers in **COCA corpus**

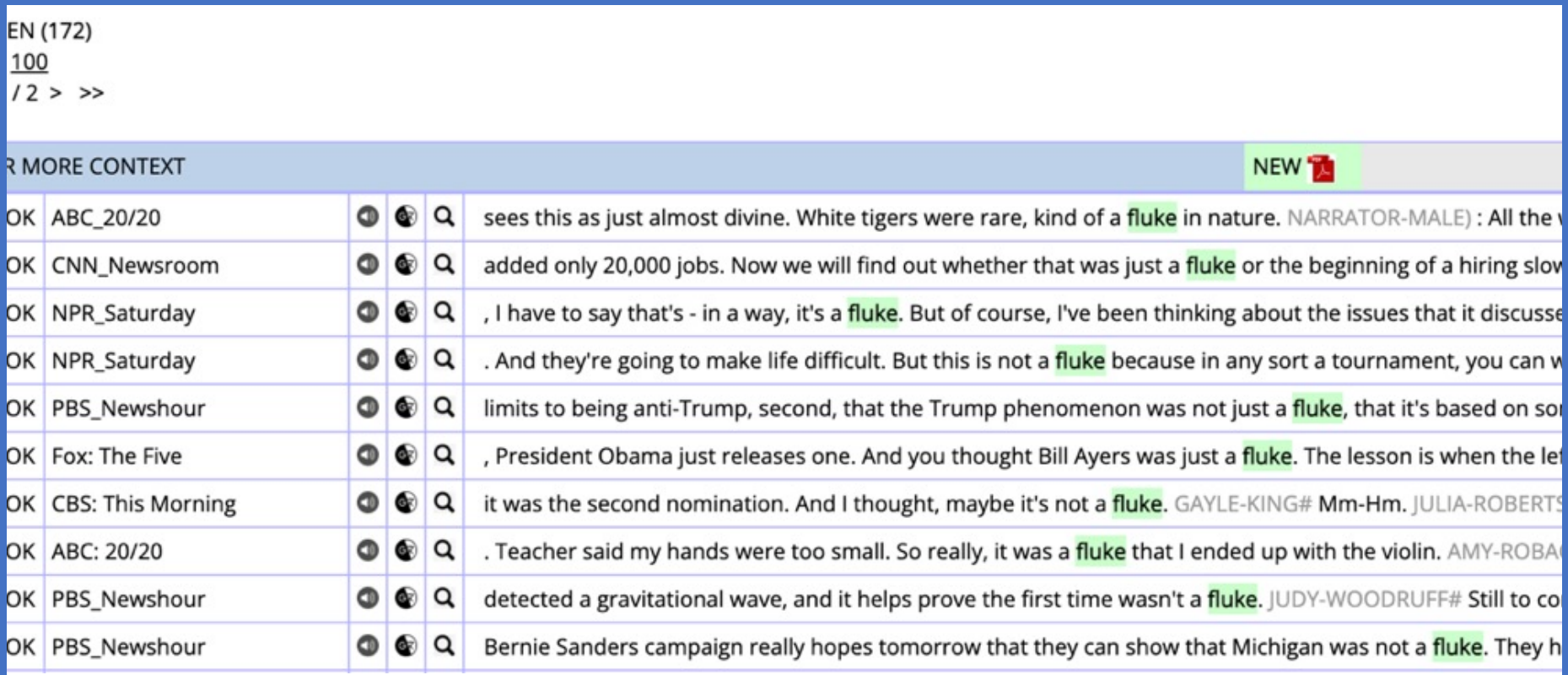


	ALL	BLOG	WEB-GENL	TV/MOVIES	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005
PEOPLE	1707619	318038	267974	138749	415659	101645	167611	198660	99283	192320	188237	181568	183
TIME	1601568	236428	228079	232871	218237	200699	185038	162623	137593	179867	195515	191206	189
WAY	1066264	163441	148363	165626	161449	152288	116071	91908	67118	122389	130210	126490	125
YEARS	1017103	136505	131781	76935	153612	88610	155761	178846	95053	129462	129993	128160	123
MAN	665748	55141	73797	185470	85148	150487	47310	47961	20434	98151	99316	89968	87
LIFE	679447	95869	103849	100139	83905	82242	87317	62671	63455	78489	86179	81956	81
DAY	665337	90617	89513	97742	96838	97284	84429	79557	29357	74005	81782	83870	84
WORLD	678541	105582	114646	53575	77455	53124	101784	86059	86316	81855	79182	84808	72
YEAR	648697	100126	91629	7425	87036	34127	108310	172353	47691	79375	78144	75755	74
THINGS	604458	109431	94568	79775	127560	69906	54095	49673	19450	62327	69004	65640	66
THING	551285	84190	65234	135318	117139	70333	41778	37293		59964	68659	66050	67
SCHOOL	474568	53031	53873	26507	41375	40753	38894	105973	114162	49646	54585	61015	69
HOUSE	432152	42459	49826	55297	92884	78693	45936	67057		54023	59649	53596	54
PRESIDENT	438560	51142	60278		207258		32663	85140	2079	58045	49284	48594	38
FAMILY	411396	47718	50309	44738	62904	40637	58567	63595	42928	43740	53199	52038	57
NIGHT	392940	34053	35727	92729	59913	94881	36006	39631		56775	57705	54594	51
CHILDREN	395055	45461	51627		54401	36547	54207	59527	93285	55908	55403	48927	47
MONEY	408460	72084	57751	65332	70701	34657	44085	61082	2768	51111	51538	45486	47
STATE	414892	67617	78162		48523		35259	116024	69307	44949	39806	40867	38

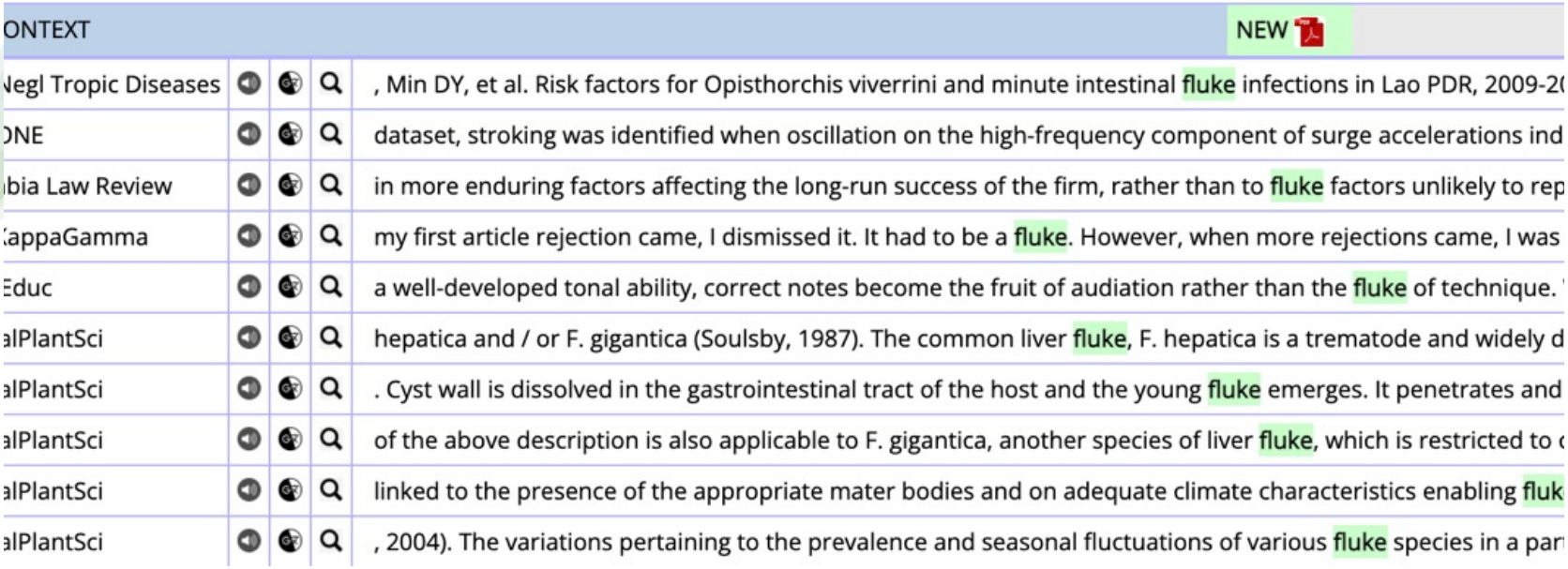


Informal words in a corpus

fluke – an unlikely chance occurrence, especially a surprising piece of luck

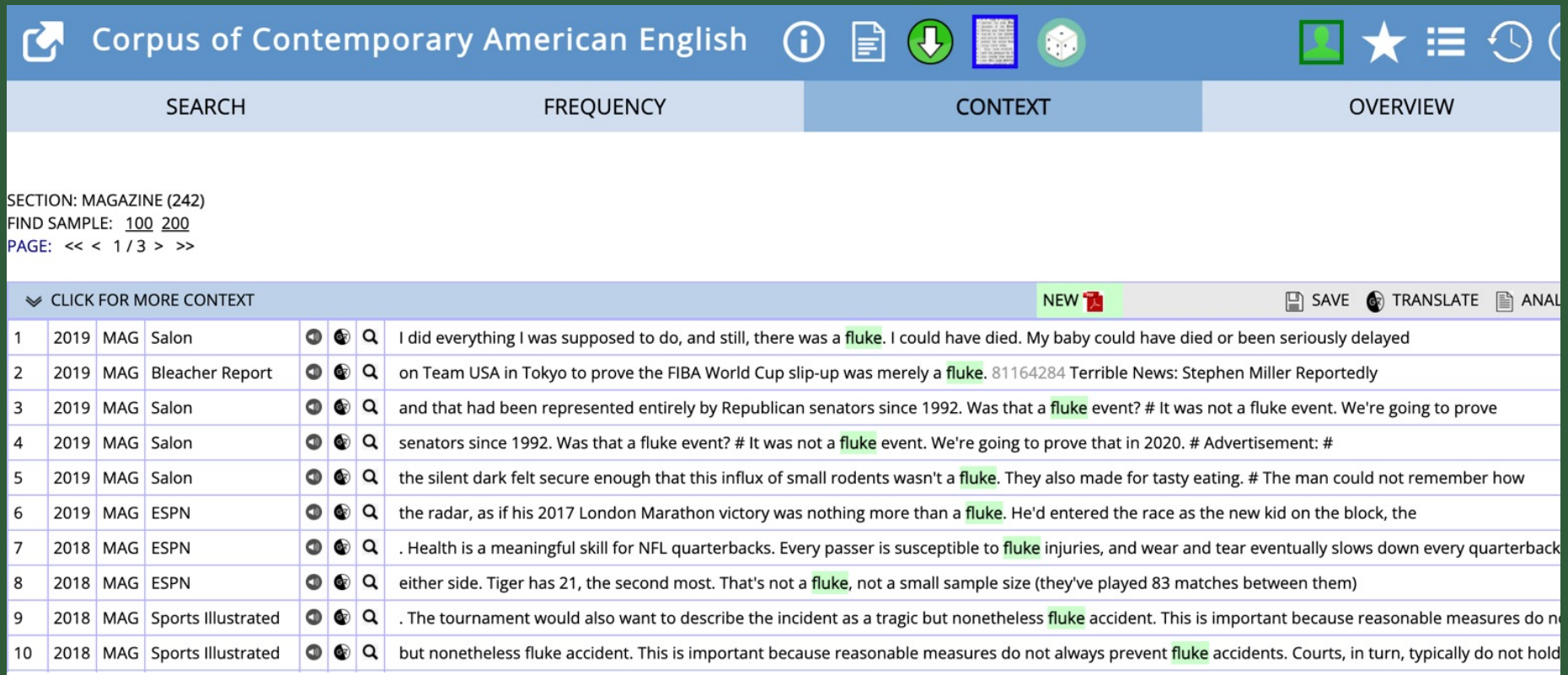


The spoken section – 172 examples



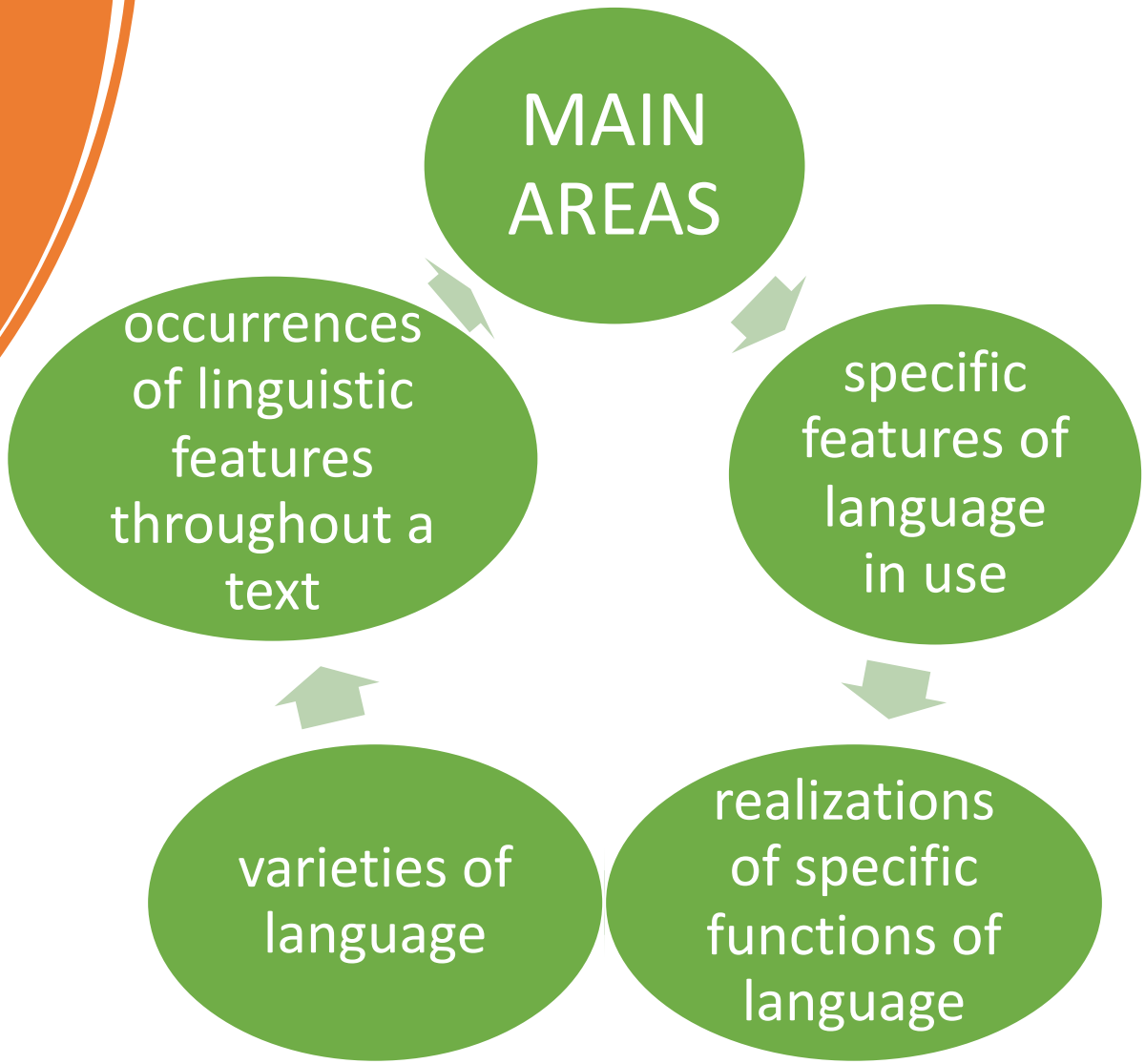
ONTEXT				NEW
Negl Tropic Diseases				, Min DY, et al. Risk factors for Opisthorchis viverrini and minute intestinal fluke infections in Lao PDR, 2009-2(
ONE				dataset, stroking was identified when oscillation on the high-frequency component of surge accelerations ind
bia Law Review				in more enduring factors affecting the long-run success of the firm, rather than to fluke factors unlikely to rep
KappaGamma				my first article rejection came, I dismissed it. It had to be a fluke. However, when more rejections came, I was
Educ				a well-developed tonal ability, correct notes become the fruit of audiation rather than the fluke of technique.
alPlantSci				hepatica and / or F. gigantica (Soulsby, 1987). The common liver fluke, F. hepatica is a trematode and widely d
alPlantSci				. Cyst wall is dissolved in the gastrointestinal tract of the host and the young fluke emerges. It penetrates and
alPlantSci				of the above description is also applicable to F. gigantica, another species of liver fluke, which is restricted to
alPlantSci				linked to the presence of the appropriate mater bodies and on adequate climate characteristics enabling fluk
alPlantSci				, 2004). The variations pertaining to the prevalence and seasonal fluctuations of various fluke species in a par

The **academic section** – 89 examples



The **magazine section** – 242 examples

2. Corpus analysis of discourse



2.1. Lexical features of discourse

The notion of **lexical cohesion** is of central importance to the study of discourse. It can be defined as the ways *'in which the components of the surface texts are mutually connected within a sequence*'.

a. REITERATION

Reiteration involves either the direct repetition of a word or the use of a related item that carries the same meaning

synonym, near-synonym or superordinate

‘furniture’ can be replaced by superordinates

‘item’, ‘object’, ‘thing’

[In a shoe-shop: <$1> is the customer, <$2> is the assistant]
<$2> Probably needs adjusting but I'll check that.
<$1> Oh right. That's **lovely**.
<$2> | Okay.
<$1> | Yeah | that's **nice**.
<$2> | They're **nice** aren't they.
<$1> Yeah they are **nice**.
<$2> **Very very nice**.
<$1> Thank you.
<$2> They feel **right**?
<$1> Yeah.
<$2> Does it?
<$1> That feels **pretty good** actually.
<$2> Yes **smashing fit**.
<$1> Yeah. (CANCODE © Cambridge University Press)

b. RELEXICALIZATION – a specific form of paraphrasing in which speakers take up one another's vocabulary as they participate in conversations

c. INTERTEXTUALITY

explores how a discourse makes reference to prior and future discourses

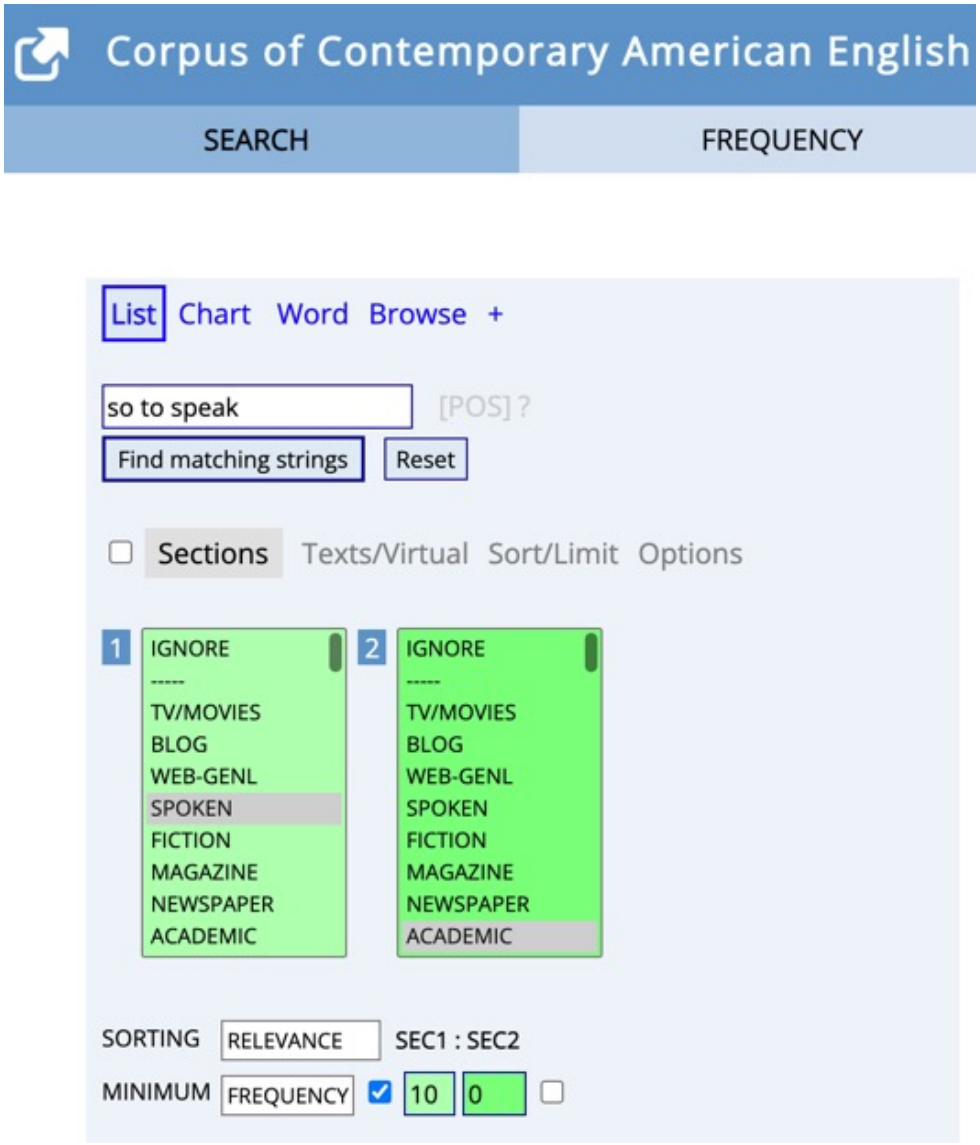
d. DISCOURSE MARKERS

'fine', 'good', 'like', 'now', 'okay', 'right', 'so', 'actually', 'well', 'I mean', 'you know', 'so to speak', 'in other words'

A **discourse marker** is a word or phrase – *a* ***conjunction, adverbial, comment clause, interjection*** – that is uttered with the primary function of ***bringing the listener's attention to a particular kind of linkage*** of the upcoming utterance with the immediate discourse context

discourse marker

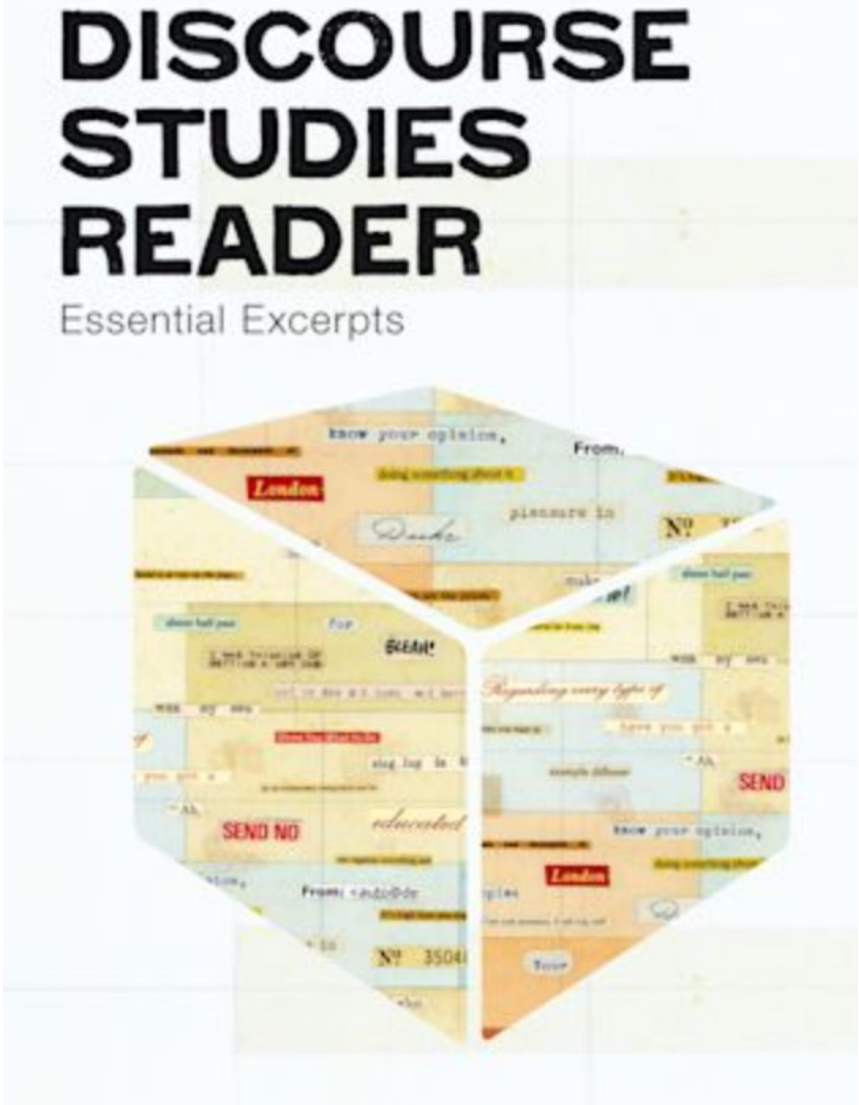
‘so to speak’



Exploring how different **forms of vocabulary are configured** over longer stretches of discourse contributes to a better understanding of their role in the creation and maintenance of ***discourse flow***

2.2. Exemplary study

- Baker, P. (2013). Corpora and discourse analysis, in K. Hyland (ed.) *Discourse Studies Reader: Essential excerpts.* London: Bloomsbury Academic, 11–34.
- **12 holiday leaflets** (17,865 words)



Using WordSmith tools, Baker started his analysis by producing a **list of the most frequent words** from his holiday corpus and compared them with data from the BNC

Table 9.1 Examples of the most frequent words in the holiday corpus and BNC expressed in percentages (adapted from Baker 2013)

	Word	*% frequency in the holiday corpus*	*% frequency in the BNC*
1	The	5.55	6.20
2	And	3.62	2.68
3	To	2.64	2.66
4	A	2.44	2.21
5	Of	1.96	3.12
6	You	1.95	0.68
7	For	1.38	0.90
8	In	1.37	1.97
9	On	1.15	0.74
10	All	1.04	0.28

Observations

1. the personal pronoun 'you' is typically regarded as a feature of spoken language

- the high frequency of 'you' suggests a 'personal style' of writing, where the writer is directly addressing the reader

2. the frequency of function words

- tend to have a high frequency, irrespective of the type of texts analyzed

the most frequent lexical items from

the holiday corpus

- ‘beach’, ‘pool’, ‘studios’

➢ represent the type of language that is used when people describe their holidays

➢ analysis of clusters (n-grams) formed around words such as ‘bar’, ‘club’

➢ explores the occurrence of verbs as another source of information about the ***discourse of tourism***

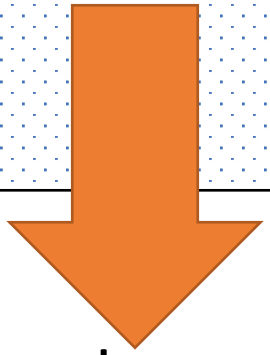
3. Computational linguistics and corpus linguistics

- **Computational linguistics** is the field of computer science that looks at how computer systems can be created that work with language in some way

DATA MINING

identification of patterns and extraction of information across very large datasets

TEXT MINING

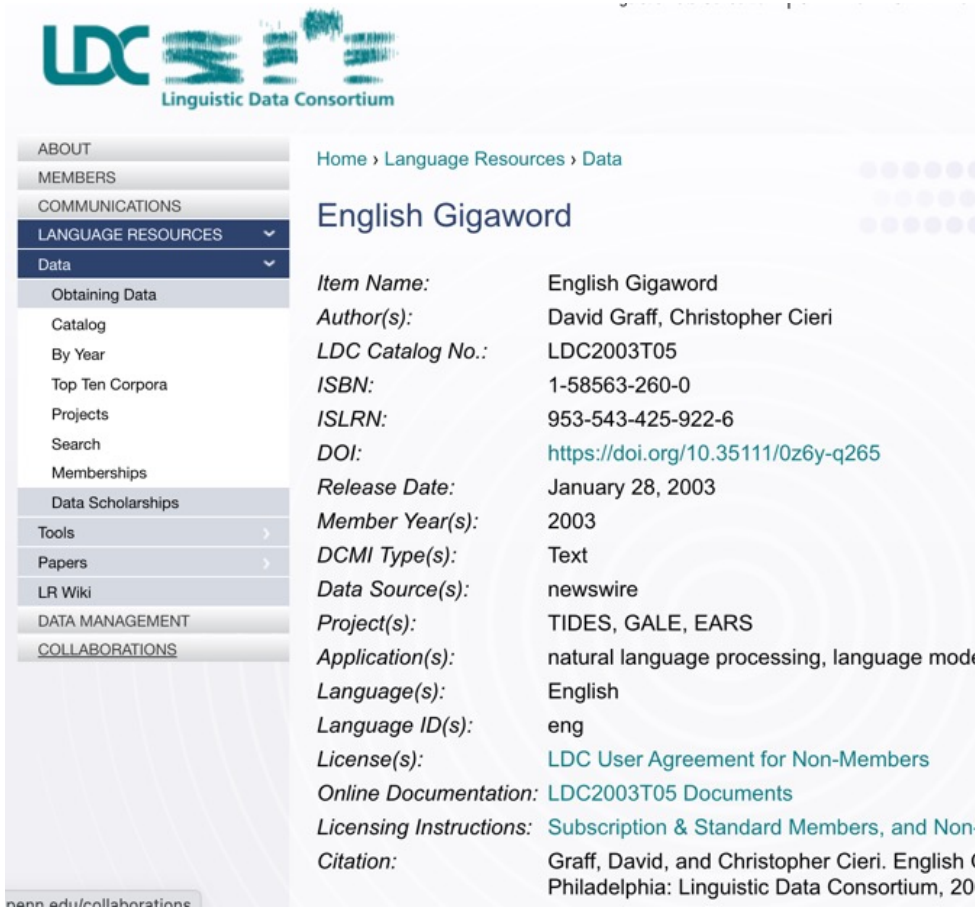


- programs for orthography and grammar correction
- information retrieval from document databases
- translation from one natural language into another

Computational linguistics might be considered as a synonym of **automatic processing of natural language**, since the main task of computational linguistics is just the **construction of computer programs to process words and texts in natural language**

Overlaps

- **Corpus linguistics** is ultimately about **finding out about the nature and usage of language**.
- **Computational linguistics** may also be concerned with **modelling the nature of language computationally**, it is *in addition* focused on **solving technical problems involving language** [McEnery and Hardie 2012: 228].



- Computational linguistics
  - ❖ makes extensive use of corpora and other sorts of digital ‘language resources’
  - ❖ many large corpora are constructed mainly for use in computational linguistics

English Gigaword corpus

Computational linguistics

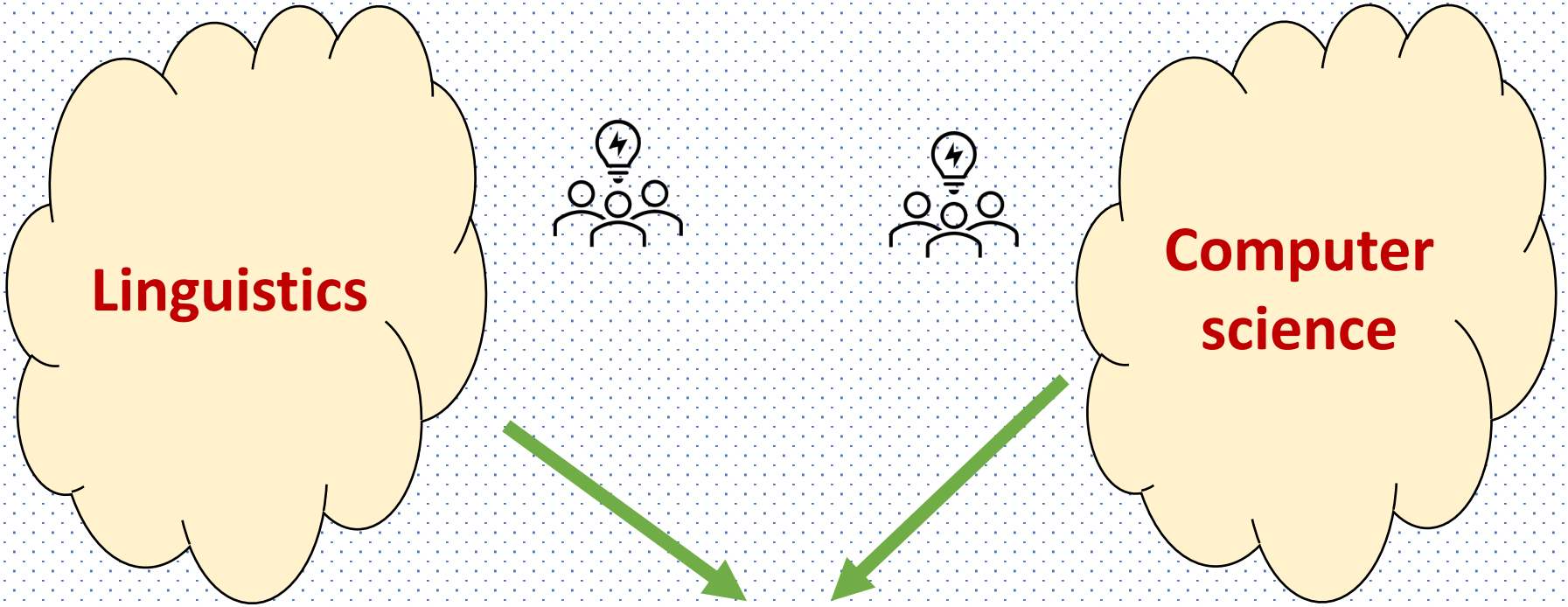
seeks to develop the computational machinery needed for an agent to exhibit various forms of linguistic behavior [Mani 2013: 466].

Agent: both human beings and artificial agents such as computer programs

Machinery: computer programs as well as the linguistic knowledge that they contain

Computational linguistics

involves designing and developing programs to carry out linguistic tasks



COMPUTATIONAL LINGUISTICS

?

?

- how language is represented and processed by the human mind
- what is the role of statistics in language understanding

3.2. Classification of applied linguistic systems

Text preparation

– *Automatic hyphenation* of words in natural language texts

– *Spell checking*

detection and correction of typographic and spelling errors

– *Grammar checking*

detection and correction of grammatical errors

– *Style checking*

detection and correction of stylistic errors

– *Referencing* specific words, word combinations, and semantic links between them

- **Information retrieval** in scientific, technical, and business document databases
- **Automatic translation** from one natural language to another
- **Natural language interfaces** to databases and other systems
- **Extraction of factual data** from business or scientific texts
- **Text generation** from pictures and formal specifications
- **Natural language understanding**
- **Optical character recognition, speech recognition**

4. Morphological processing

Morphology is the study of the structure of words

The task of an automatic morphological analyzer is to take a word in a language and break it down into its **stem form** along with any **affixes** that it may have attached to that stem

Lily reads well

- *Lily* as a proper name
- *reads* as the 3d p.s. present form of the verb *read* (*read + s*)
- *well* as either an adverb or a singular noun

4.1. Tokenization

the first step in morphological analysis

it is to identify separate words

4.2. Morphological analysis and synthesis

- **a**. An *automatic morphological analyzer* takes a word and breaks it down into its component morphemes (stems and affixes)
- Sometimes, instead of a full morphological analysis, a simple **stemming** algorithm is used which strips off suffixes to arrive at a stem form

b. Another strategy is to use a fully inflected lexicon, which includes all the possible affixed forms of every word in the language

The analyzer simply looks up the word in the list. Such lists are almost inevitably **incomplete**, and they can become too large and unwieldy for computers to handle

5. Syntactic processing

- Given a set of linguistic rules that describe how elements of a sentence can be put together, a computer program called a *syntactic parser* will try to find the best grammatical analysis of a sentence

5.1. Context-free grammars

I can fish.

- *phrase structure rules* break up a sentence into its constituent parts, consisting of syntactic phrases or words

S → NP VP	|	V → can
VP → Aux V	|	V → fish
VP → V NP	|	V → dance
VP → V	|	Aux → can
VP → Aux V NP	|	D → the
NP → D N	|	N → fish
NP → N	|	N → dance
NP → Pronoun	|	Pronoun → I

“Toy grammar” abbreviations

Aux	Auxiliary (can)
D	determiner (the)
N	noun
NP	noun phrase
S	sentence
V	verb
VP	verb phrase

5.2. Parsing

I can fish.

How can we get a computer to analyze the syntactic structure of a sentence?

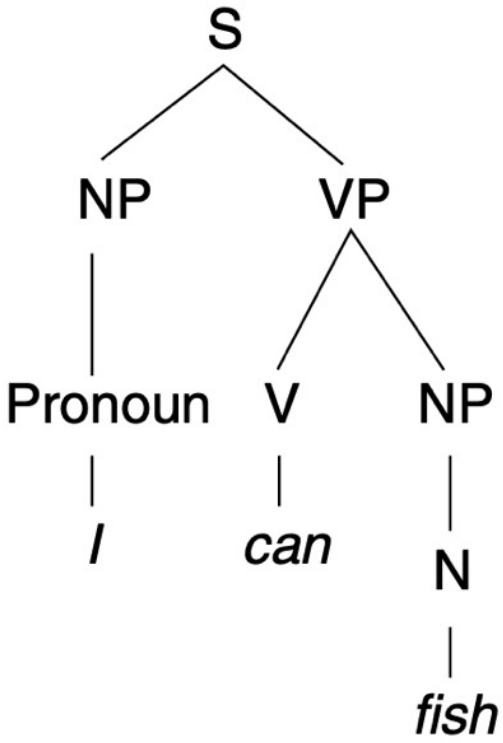
A **parser** takes an input sentence and produces one or more syntactic representations of it

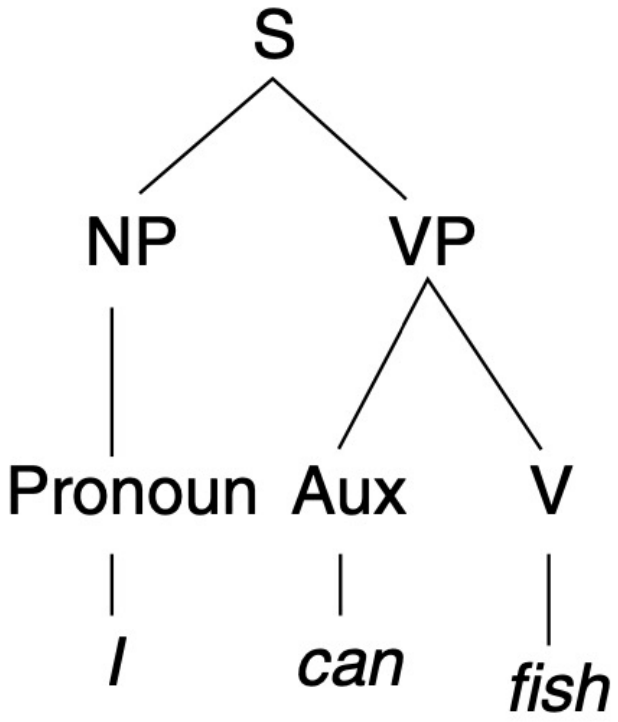
It produces a single representation if the sentence is syntactically **unambiguous**, but more than one representation if there is syntactic **ambiguity**

I can fish.

PARSE TREE

top-down parser builds the parse trees from the top-down





5.3. Part-of-speech tagging

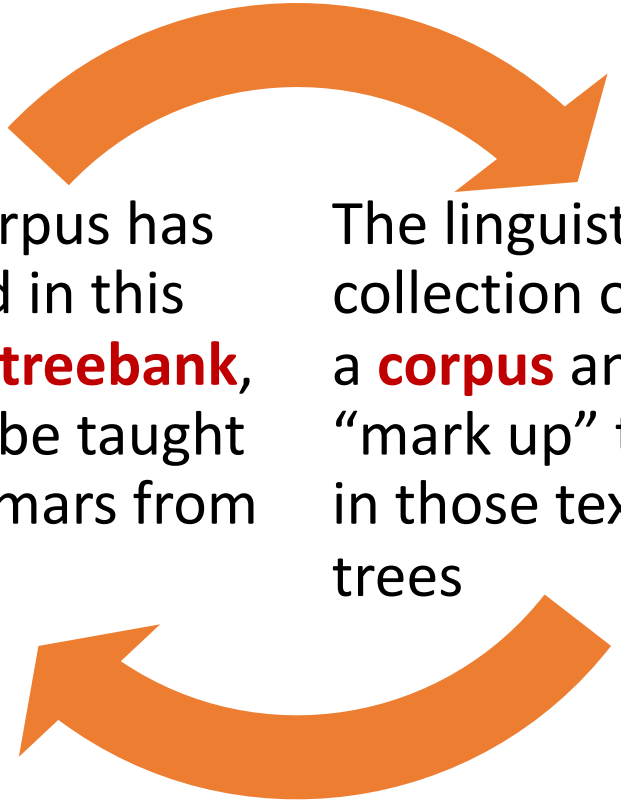
To each word that has more than one part of speech, the tagger assigns the **most likely part-of-speech**

This is based on context-based rules derived by

human intuition (after ***the***, ***fish*** is likely to be a **noun**)

machines that learn from a collection of example sentences that have been tagged already with parts of speech

5.4. Statistical parsing



Once a large corpus has been annotated in this way, creating a **treebank**, computers can be taught to induce grammars from it

The linguists take a collection of texts called a **corpus** and analyze and “mark up” the sentences in those texts with parse trees

6. Semantic processing

How can a computer decide which meaning is intended?

- **word-sense disambiguator**
- It can use the **context of neighboring words** in the sentence as well as other words in the document to figure out which meaning of a given word is most likely.

the verb *cook*

- A verb's meaning includes
- syntactic *subcategorization*

⇒

the syntactic elements, or "arguments," it combines with

- *thematic roles*

⇒

- the semantic relations between a verb and its arguments

She cooked meatloaf

She cooked Mary a great dinner

a. COOK (Theme $_{NP}$)

b. COOK (Recipient $_{NP}$, Theme $_{NP}$)

c. COOK (Theme $_{NP}$, Recipient $_{for\text{-}PP}$)

She cooked a great dinner for Mary

Sentence meaning

How can a computer combine meanings?

- One way to approach this is to have **semantic rules** that accompany syntactic rules in the grammar.
- The syntactic rule (a) below is augmented with a semantic rule (b) that says the VP's meaning is constructed by applying the V's meaning to the NP's meaning

a. $\text{VP} \rightarrow \text{V NP}$

b. $\text{VP}.\textit{meaning} = \text{Apply}(\text{V}.\textit{meaning}, \text{NP}.\textit{meaning})$