Syntactic-Ngrams over Time from a Very Large Corpus of English Books

Yoav Goldberg and Jon Orwant

Bar-Ilan University

Presented at *SEM 2013, Atlanta, GA

Many thanks to Google's parsing team

Ryan

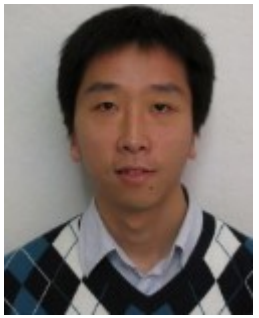
Keith

Slav

Kuzman

Dipanjan

Fernando



Hao

Michael

Joakim
(at the time)

Terry

Disclaimer:

I'm a syntax and parsing guy

I don't know much about semantics

I'm not even sure I know what semantics mean
(I do know what tensors are, though,
and some of you seem to like them)

however, I am pretty sure you will find this useful

A lexical/syntactic resource

Based on 350 billion parsed words

Time-indexed

Available for download

distributed on the web under a
Creative commons non-commercial share-alike license

commondatastorage.googleapis.com/books/syntactic-ngrams/index.html

Google books Syntactic N-grams

Content: These datasets contain counted syntactic ngrams (dependency tree fragments) extracted from the English portion
datasets are described in the following publication. A more popular description is available here. The dataset format and orga
file.

Usage: This release is licensed under the terms and conditions of the Creative Commons Attribution-Non Commercial Share

Version 20130501

English All

Nodes 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89

Arcs 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 3
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89

Biarcs 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89

Triarcs 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89

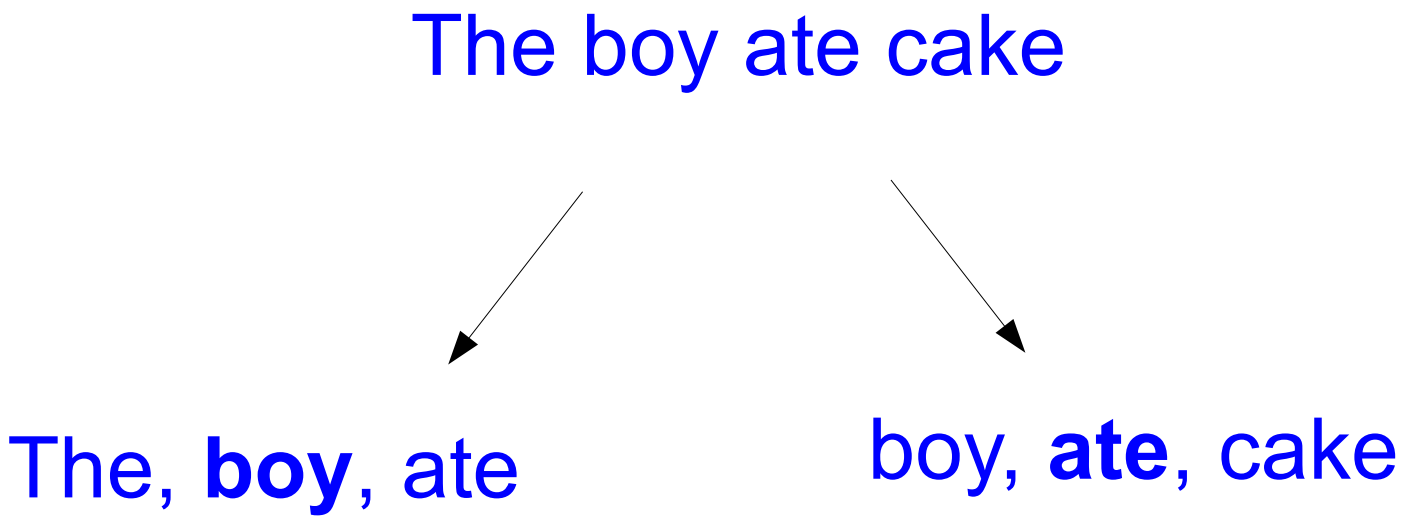
Quadarcs 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88

Lexical/syntactic resource

“You shall know a word by the company it keeps”
- Firth

what is the company of a word?

sequential context is widely used



...but sequential context is only a proxy

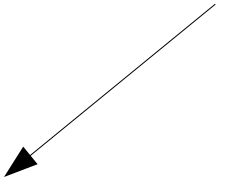
(often misleading)

The boy with the brown eyes ate the cake

...but sequential context is only a proxy

(often misleading)

The boy with the brown eyes ate the cake

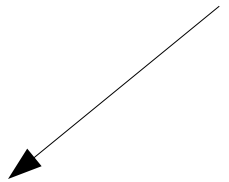


eyes, **ate**, the

...but sequential context is only a proxy

(often misleading)

The boy with the brown eyes ate the cake

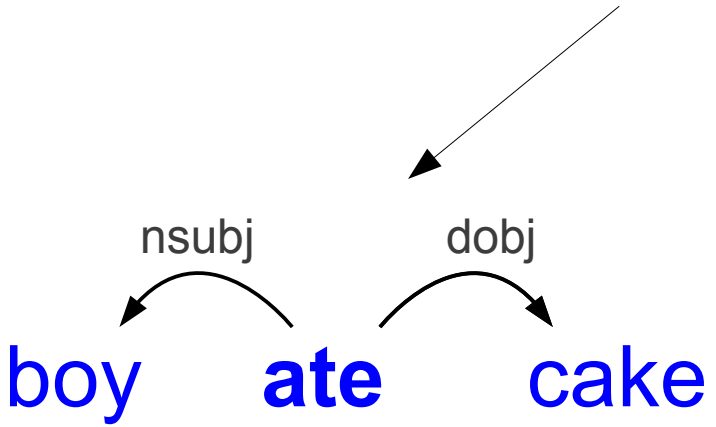


eyes, **ate**, the

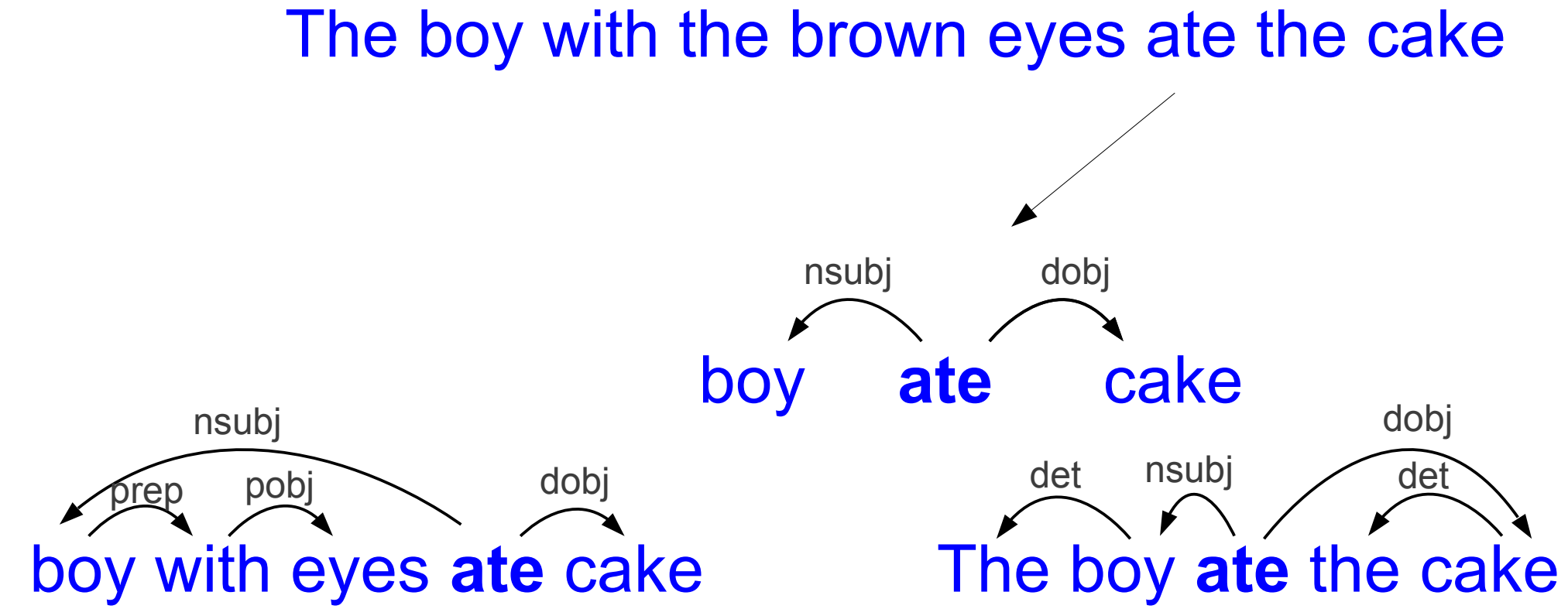
brown, eyes, **ate**, the, cake

what we really care for is the **syntactic context**

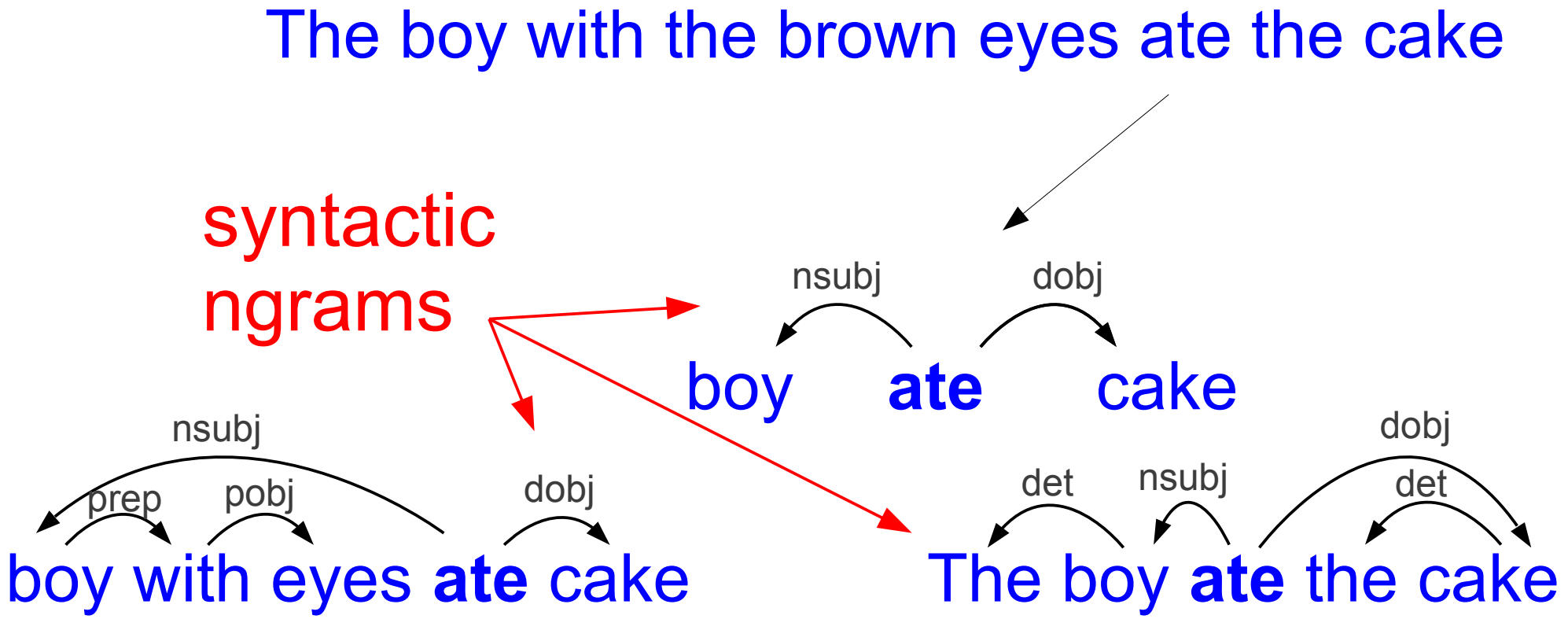
The boy with the brown eyes ate the cake



what we really care for is the **syntactic context**



what we really care for is the **syntactic context**



We took a large corpus covering many years

We parsed it with a good parser

We extracted and counted syntactic-ngrams

We took a large corpus covering many years

English Google Books

~3.5M books
published between 1520 to 2008
(most after 1800)

~350B words
~x100 times larger than prev efforts

We took a large corpus covering many years

We parsed it with a good parser

We extracted and counted syntactic-ngrams

CRF tagger **with cluster features**
induced from the books corpus

s

We parsed it with a good parser

We extracted and counted syntactic-ngrams

CRF tagger **with cluster features**
induced from the books corpus

We parsed it with a good parser

arc-eager transition parser
beam of size 8
features of Zhang and Nivre (2011)
state of the art

CRF tagger **with cluster features**
induced from the books corpus

Trained on
WSJ + Brown + Question-Treebank

arc-eager transition parser
beam of size 8
features of Zhang and Nivre (2011)
state of the art++

We took a large corpus covering many years

counting at this scale is not trivial
(luckily, Google has great infrastructure)

We extracted and counted syntactic-ngrams

We took a large corpus covering many years

counting at this scale is not trivial
(luckily, Google has great infrastructure)

We extracted and counted syntactic-ngrams

How do these look like?

We provide several datasets,
each with a different kind of syntactic-ngrams.

they have names:
arcs, biarcs, triarcs, quadarcs, …

I will describe them shortly
(more details in the paper and website)

content words

vs.

functional markers*

*Defined based on dependency labels

content words

said, dog, beautiful, quickly, he, John, 59, hundreds, increasing, jumped, ...

vs.

functional markers*

*Defined based on dependency labels

content words

said, dog, beautiful, quickly, he, John, 59, hundreds, increasing, jumped, ...

vs.

functional markers*

to, will, his, the, not, did, your, has, some, ...

*Defined based on dependency labels

content words

said, dog, beautiful, quickly, he, John, 59, hundreds, increasing, jumped, ...

vs.

to, will, his, the, not, did, your, has, some, ...

functional markers*

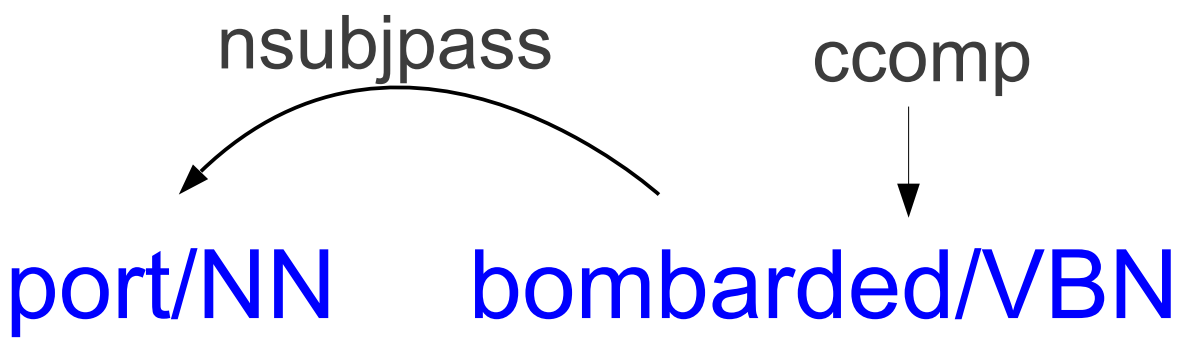
focus on relations between content words

but retain information about functional markers

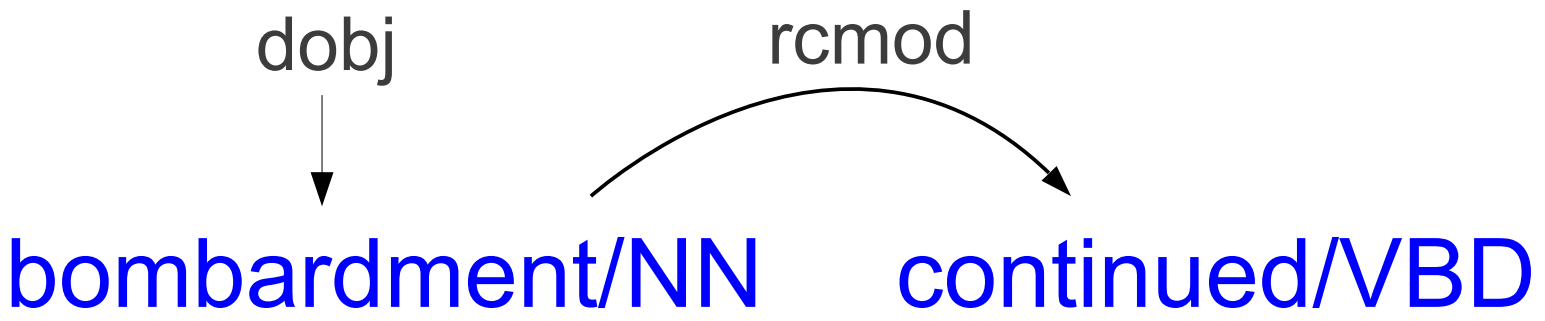
*Defined based on dependency labels

arcs: two content words

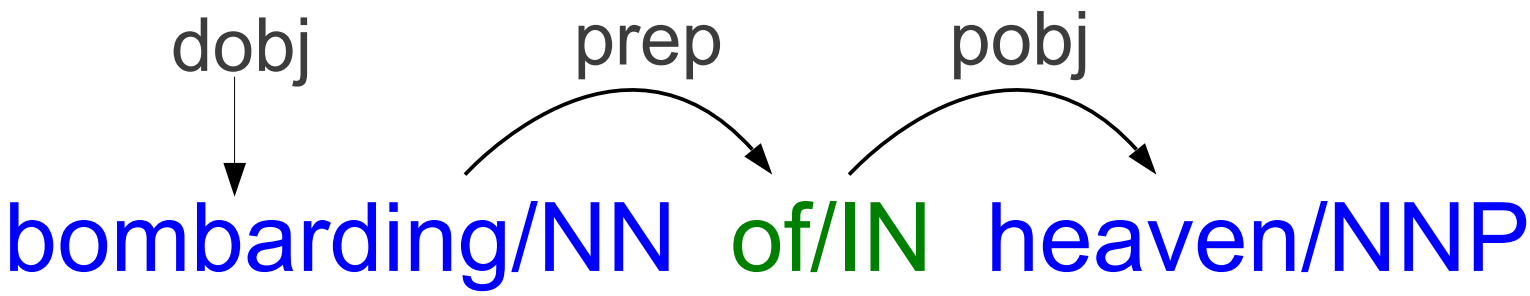
arcs: two content words



arcs: two content words

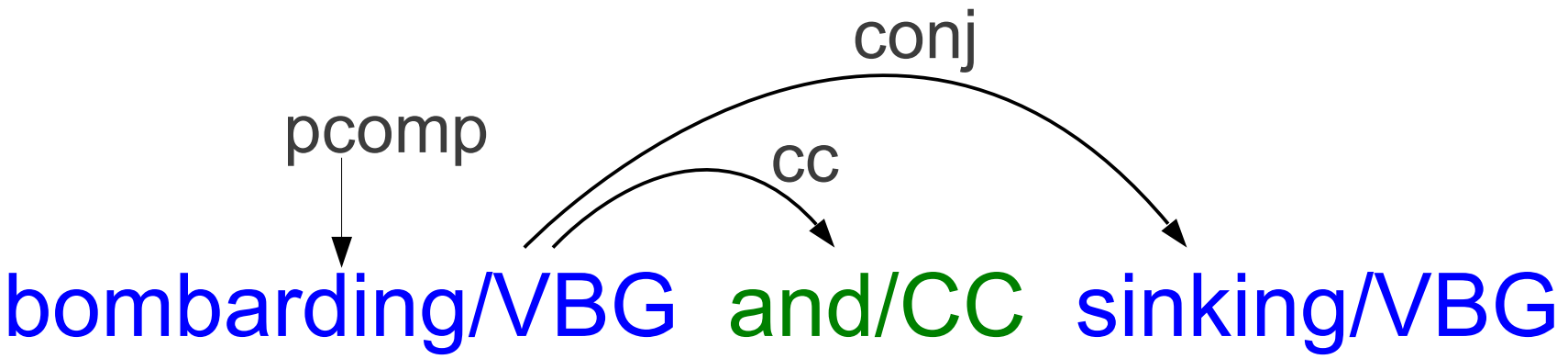


arcs: two content words



prepositions not counted as content words

arcs: two content words



coordinators not counted as content words

arcs: two content words

“arcs” ngrams are very useful

arcs: two content words

"arcs" ngrams are very useful

can answer many natural queries:

- subjects/objects of a given verb
- adjectivial modifiers of a noun
- things coordinated with a word
- …

arcs: two content words

“arcs” ngrams are very useful

most work in syntactic vector-space models can be replicated using this set

arcs: two content words

"arcs" ngrams are very useful

most work in syntactic vector-space models can be replicated using this set

My student tried using it
with our current model and got
a very nice boost in accuracy!

C. Biemann, PhD, a few days ago

arcs: two content words

"arcs" ngrams are very useful

I can inspect the modification patterns of gradable adjectives!
This is sooo interesting for me :-)

G. Weidman Sassoon, PhD, **a real semantician**

nearly/RB/advmod/2 **tall**/JJ/acomp/0	6707
unusually/RB/advmod/2 **tall**/JJ/acomp/0	4444
extremely/RB/advmod/2 **tall**/JJ/amod/0	4419
unusually/RB/advmod/2 **tall**/JJ/amod/0	4331
fairly/RB/advmod/2 **tall**/JJ/acomp/0	3466
extremely/RB/advmod/2 **tall**/JJ/acomp/0	3267
fairly/RB/advmod/2 **tall**/JJ/amod/0	3218
immensely/RB/advmod/2 **tall**/JJ/amod/0	2806
exceptionally/RB/advmod/2 **tall**/JJ/amod/0	2623
generally/RB/advmod/2 **tall**/JJ/acomp/0	2470
relatively/RB/advmod/2 **tall**/JJ/amod/0	2253
exceptionally/RB/advmod/2 **tall**/JJ/acomp/0	1929
enormously/RB/advmod/2 **tall**/JJ/amod/0	1567
nearly/RB/advmod/2 **tall**/JJ/amod/0	1550
really/RB/advmod/2 **tall**/JJ/acomp/0	1532
remarkably/RB/advmod/2 **tall**/JJ/acomp/0	1523
really/RB/advmod/2 **tall**/JJ/amod/0	1474
immensely/RB/advmod/2 **tall**/JJ/acomp/0	1452
relatively/RB/advmod/2 **tall**/JJ/acomp/0	1427
particularly/RB/advmod/2 **tall**/JJ/amod/0	1422
particularly/RB/advmod/2 **tall**/JJ/acomp/0	1379
moderately/RB/advmod/2 **tall**/JJ/amod/0	1360

tterns of

This is sooo interesting for me :-)

G. Weidman Sassoon, PhD, **a real semantician**

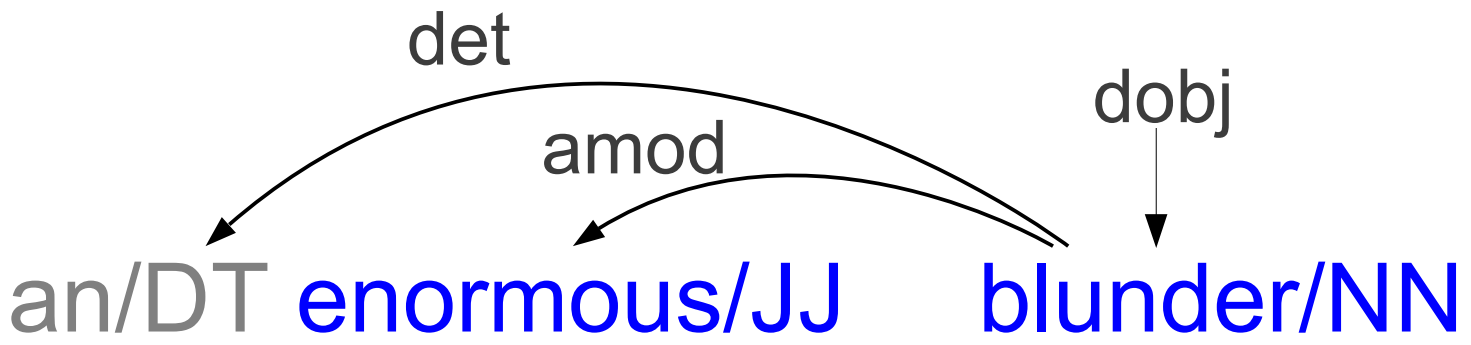
arcs: two content words

“arcs” ngrams are very useful

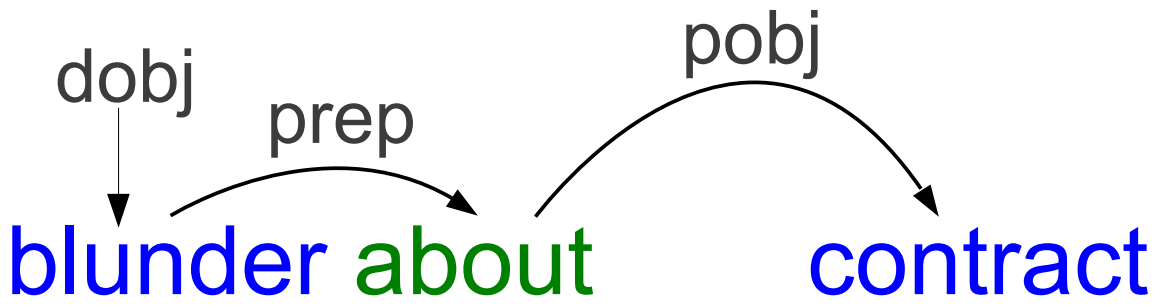
there's much more available

extended-arcs: two content words + all functional markers

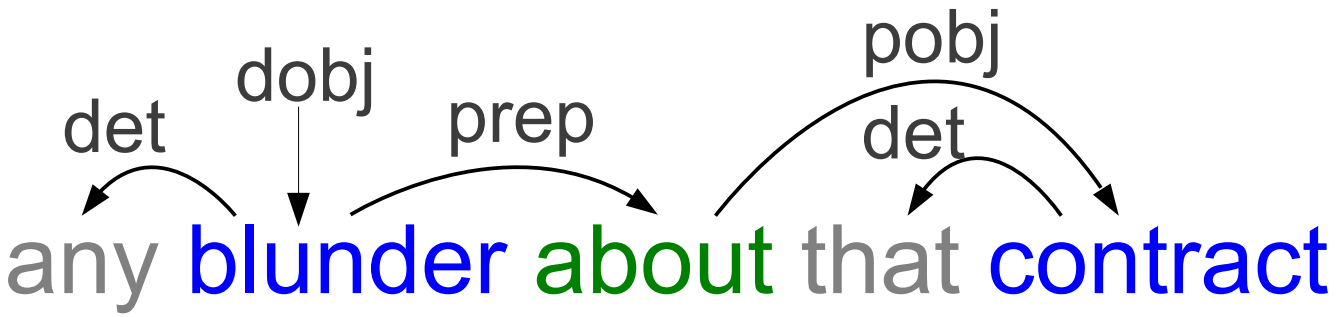
extended-arcs: two content words + all functional markers



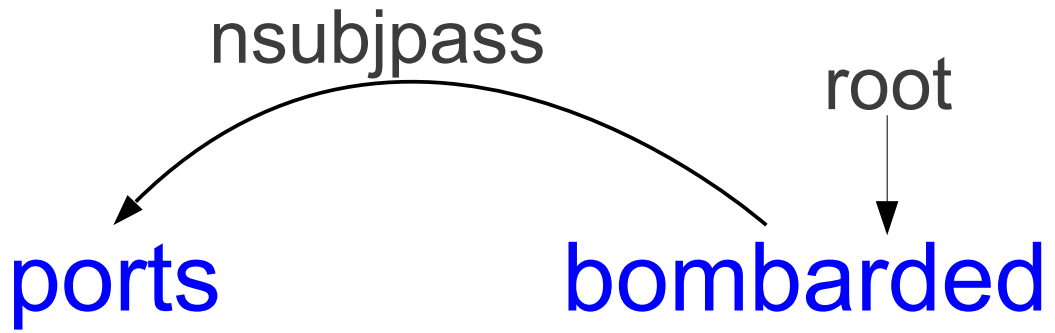
extended-arcs: two content words + all functional markers



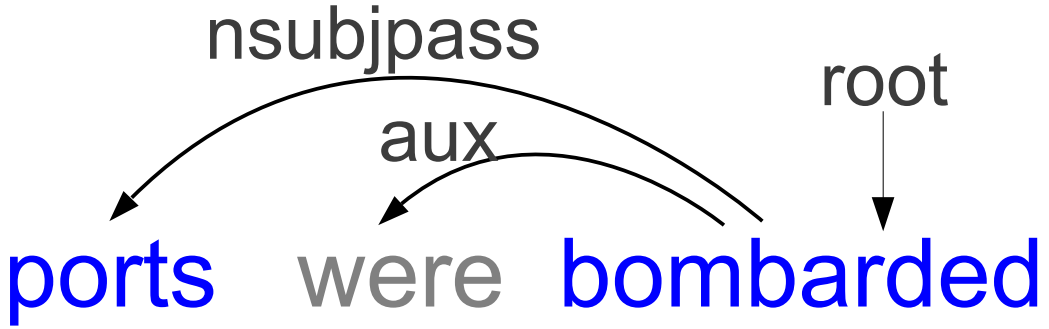
extended-arcs: two content words + all functional markers



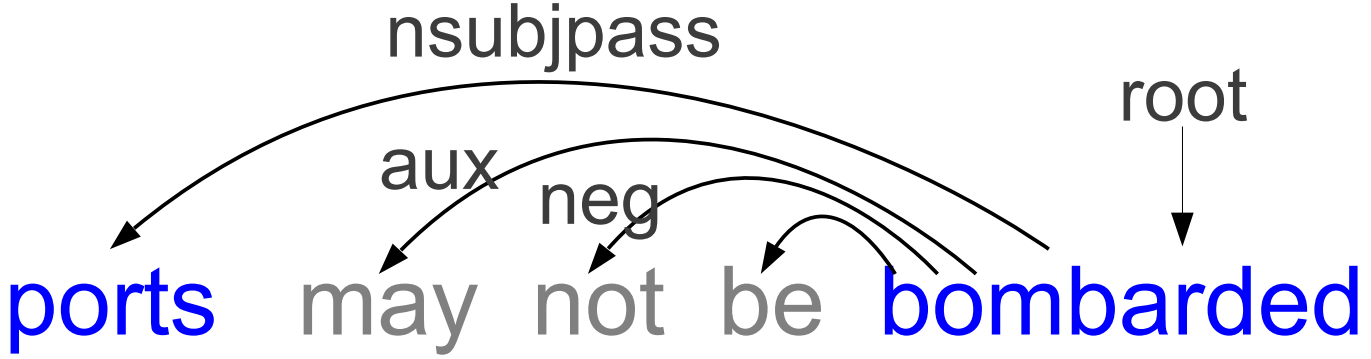
extended-arcs: two content words + all functional markers



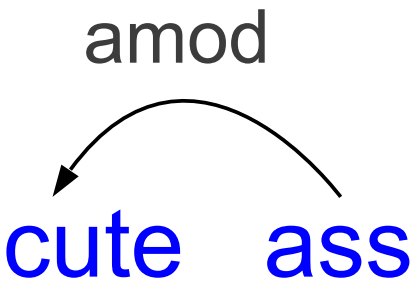
extended-arcs: two content words + all functional markers



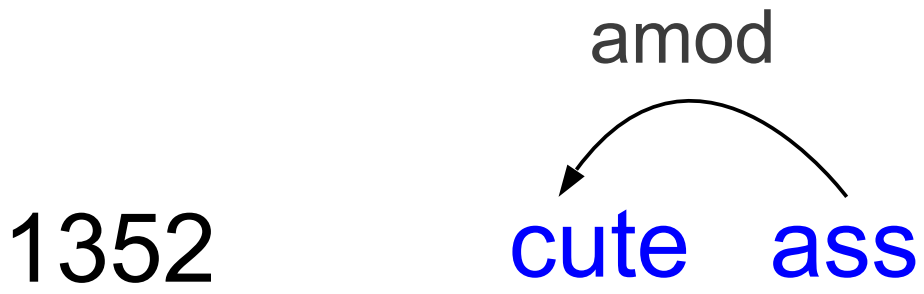
extended-arcs: two content words + all functional markers



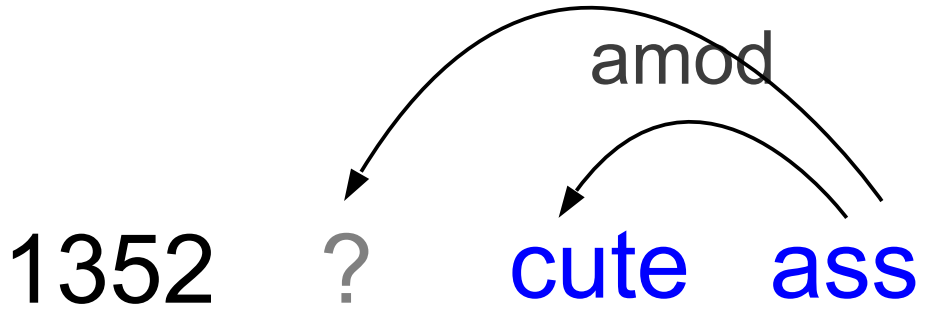
extended-arcs: two content words + all functional markers



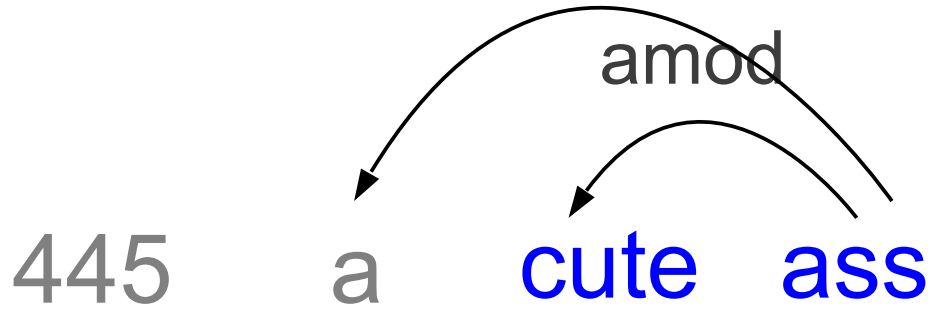
extended-arcs: two content words + all functional markers



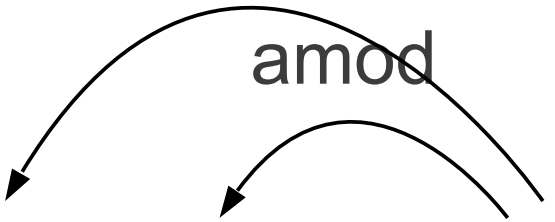
extended-arcs: two content words + all functional markers



extended-arcs: two content words + all functional markers



extended-arcs: two content words + all functional markers

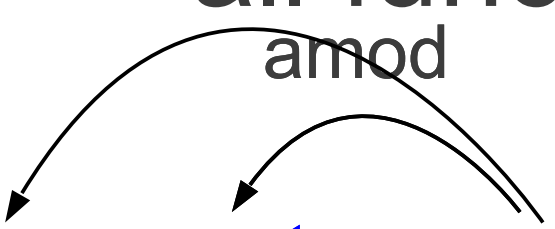


445	a	cute	ass
235	her		
221	your		
167			
103	the		
82	that		
71	his		
28	my		

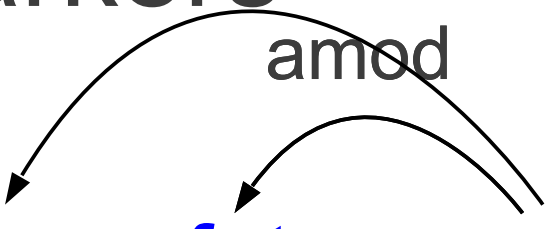
extended-arcs: two content words + all functional markers

445	a
235	her
221	your
167	
103	the
82	that
71	his
28	my

extended-arcs: two content words + all functional markers

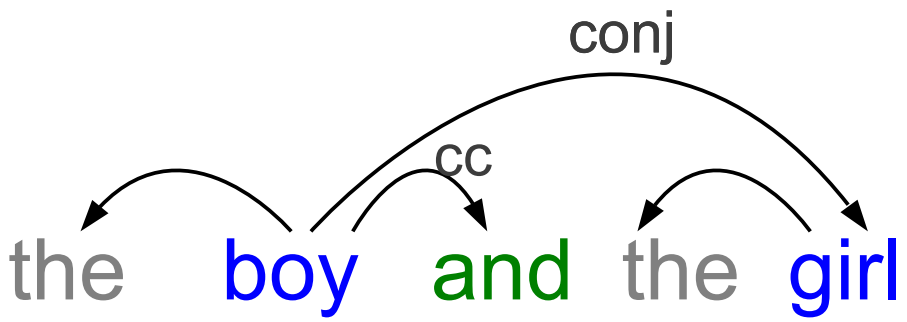


445	a	cute	ass
235	her		
221	your		
167			
103	the		
82	that		
71	his		
28	my		

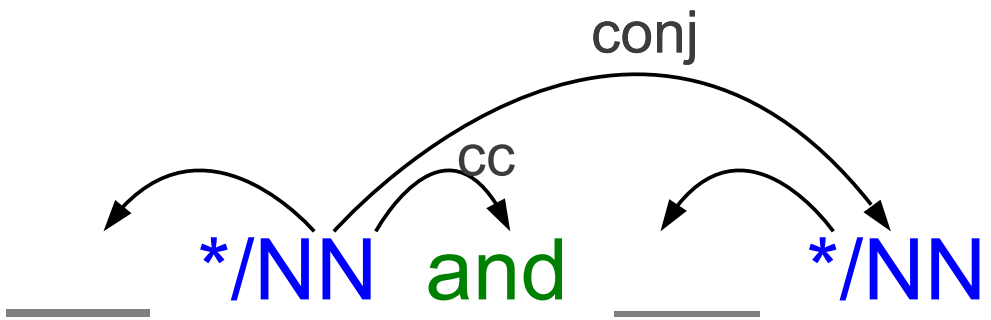


2768	your	fat	ass
1504	his		
1044			
993	a		
764	her		
640	my		
359	that		
173	the		
	…		
11	a frog 's		

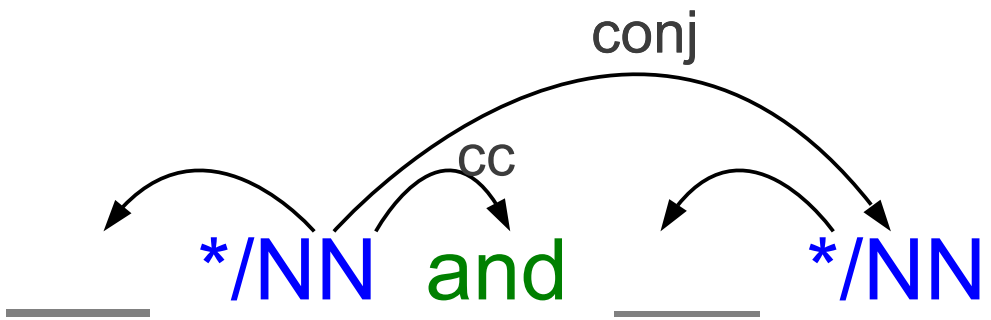
Functional modifiers of coordinated nouns



Functional modifiers of coordinated nouns

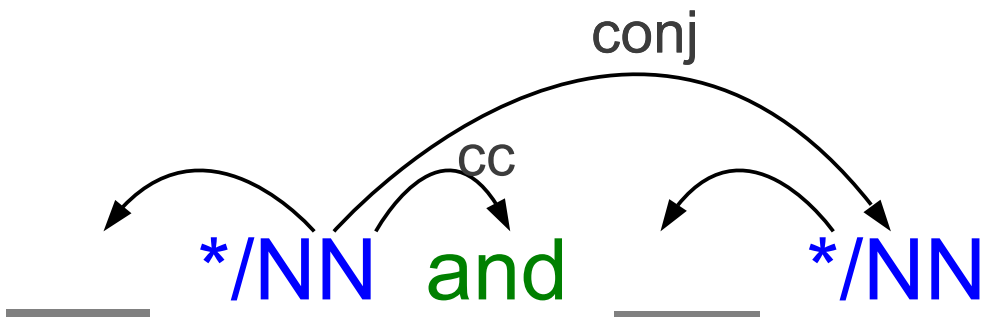


Functional modifiers of coordinated nouns



parallelism?

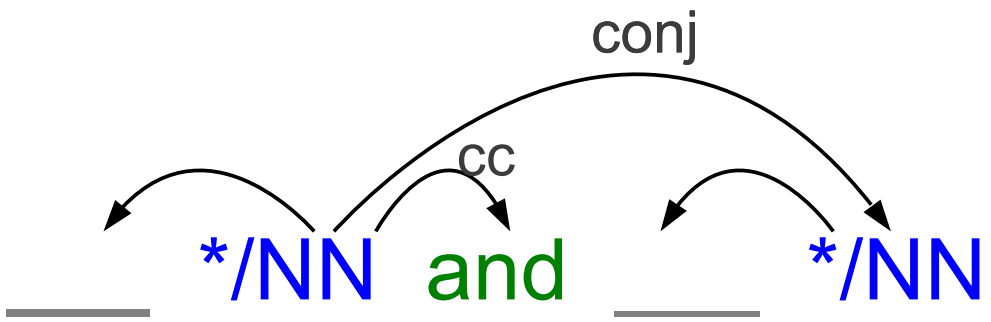
Functional modifiers of coordinated nouns



parallelism?

79250839	the	and	the
15031401	a	and	a
3820439	the	and	its
2614562	the	and	his
2467965	his	and	his
2242856	a	and	the
2133545	the	and	a
2030446	the	and	their
1856827	an	and	a
1686133	a	and	an
1020169	their	and	their
892783	his	and	the
750079	my	and	my
714221	her	and	her
658563	its	and	its
475910	an	and	an
467310	our	and	our
459989	the	and	her

Functional modifiers of coordinated nouns

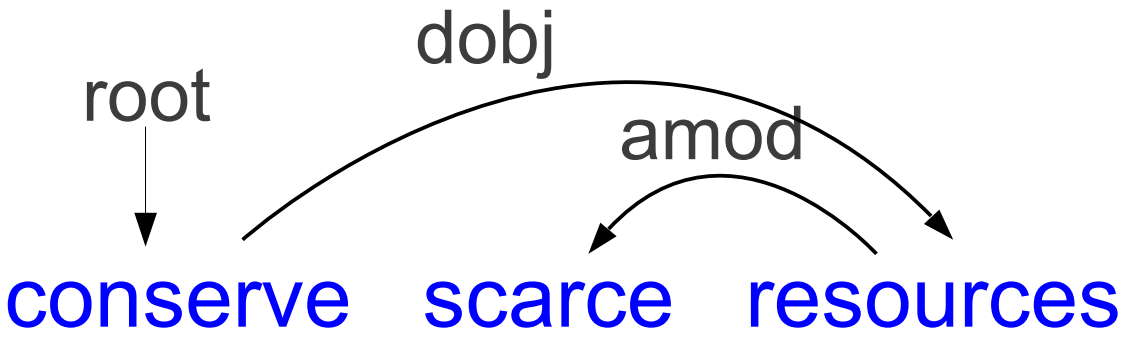


parallelism?

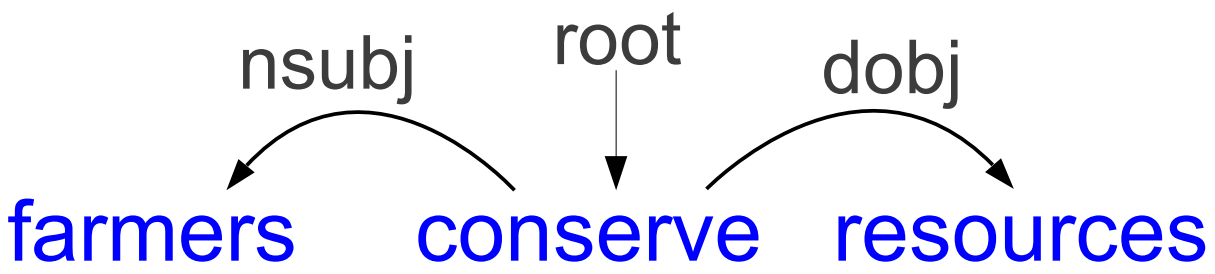
79250839	the	and	the
15031401	a	and	a
3820439	the	and	its
2614562	the	and	his
2467965	his	and	his
2242856	a	and	the
2133545	the	and	a
2030446	the	and	their
1856827	an	and	a
1686133	a	and	an
1020169	their	and	their
892783	his	and	the
750079	my	and	my
714221	her	and	her
658563	its	and	its
475910	an	and	an
467310	our	and	our
459989	the	and	her

biarcs: three content words

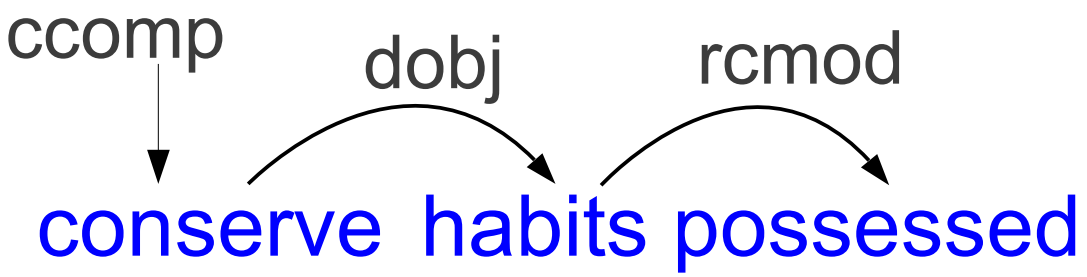
biarcs: three content words



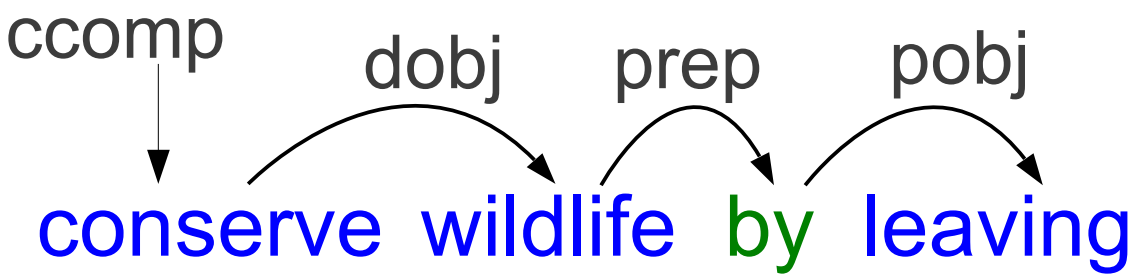
biarcs: three content words



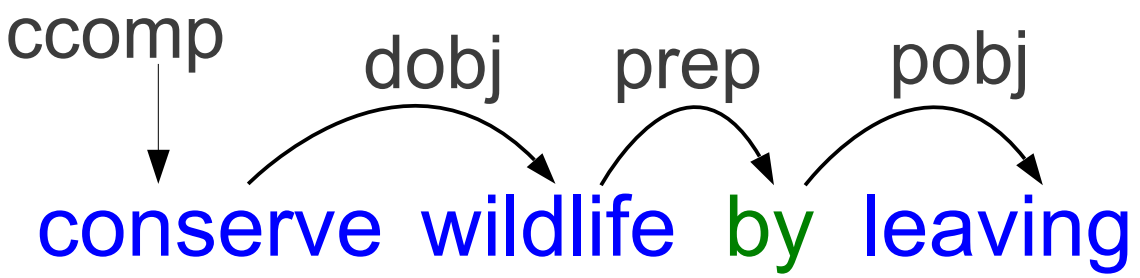
biarcs: three content words



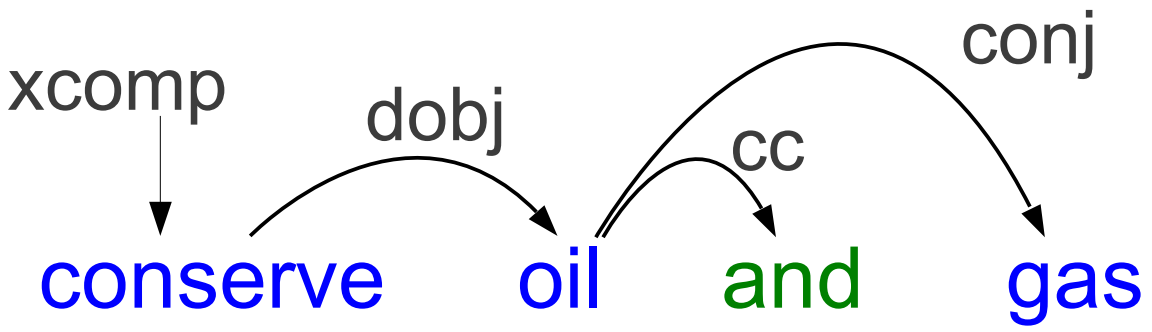
biarcs: three content words



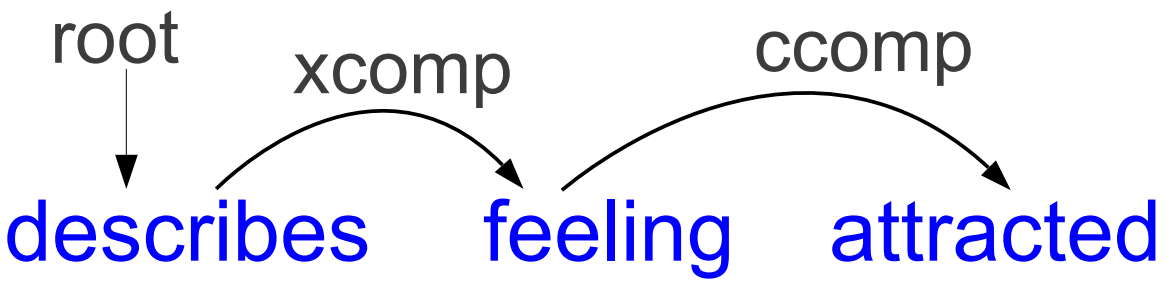
biarcs: three content words



biarcs: three content words



biarcs: three content words



biarcs: three content words

capture interactions between
subject, verb and object

biarcs: three content words

capture interactions between
two adjectives of a noun

biarcs: three content words

capture interactions between
verb, adverb and subject

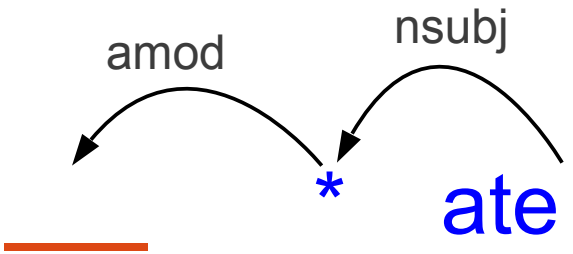
biarcs: three content words

VSM's not covered by “arcs” dataset
are probably covered by this one

biarcs: three content words

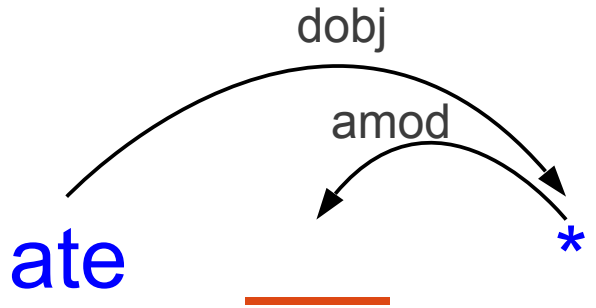
second-order questions

adjectives of things that eat



old, young, little, other, most, many, first, poor, whole, white, ancient, average, obese, few, hungry, primitive, native, condemned, human, large, wild, black, great, small, starving, american, neotropical, rich, entire, ordinary, pregnant, thin, lean, normal, prehistoric, overweight, elder, fat, grave, wicked, local, holy, wealthy, working, unfortunate, miserable, sick, indian, cannibalistic, indigenous, savage, persian, maori, southern, primate, female, aboriginal, skinny, austrelian, ...

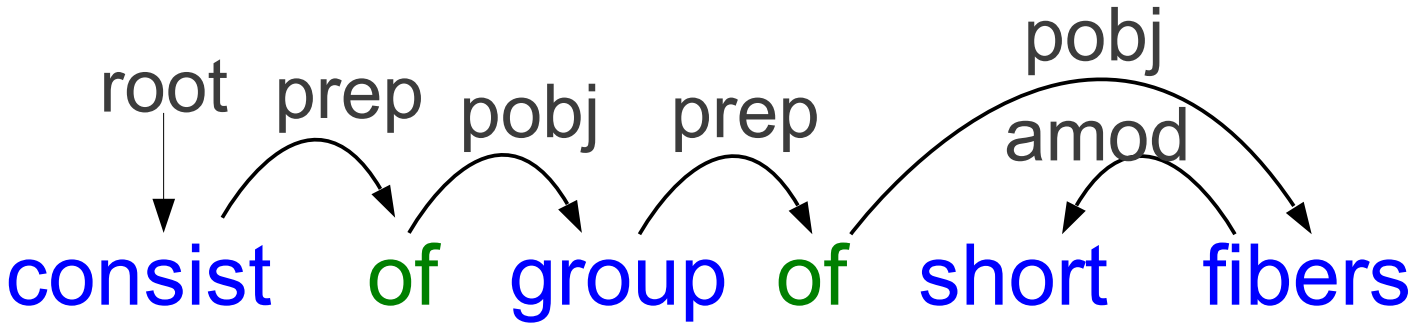
adjectives of things being eaten



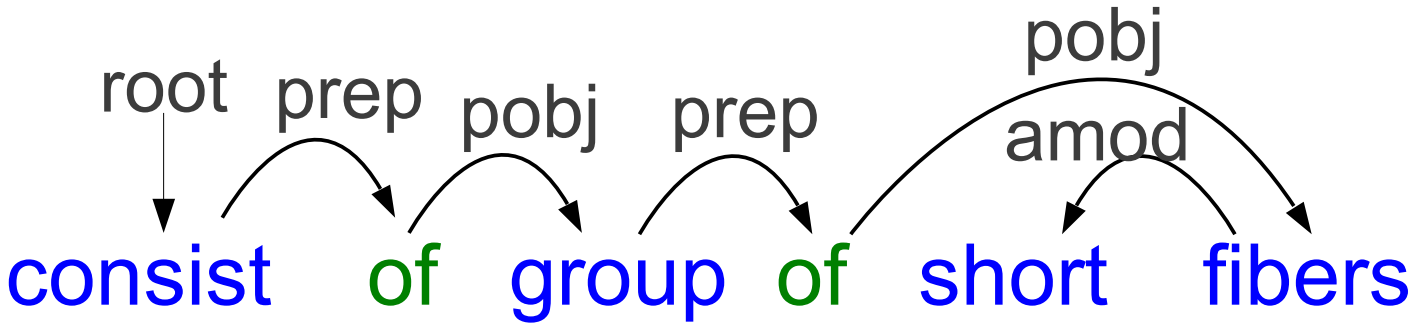
last, little, good, same, hearty, more, cold, whole, few, large, much, raw, small, great, hot, human, many, own, first, boiled, only, forbidden, big, other, light, simple, lobe, wild, fresh, green, roast, sweet, several, huge, delicious, quick, enormous, late, boiled, dry, white, frugal, early, next, fried, hasty, different, black, dried, red, fried, stale, canned, chinese, sour, cooked, french, vegetarian, mexican, baked, wonderful, poisoned, scrambled, roasted, enough, broiled, soft, kosher, ...

triarcs: four content words

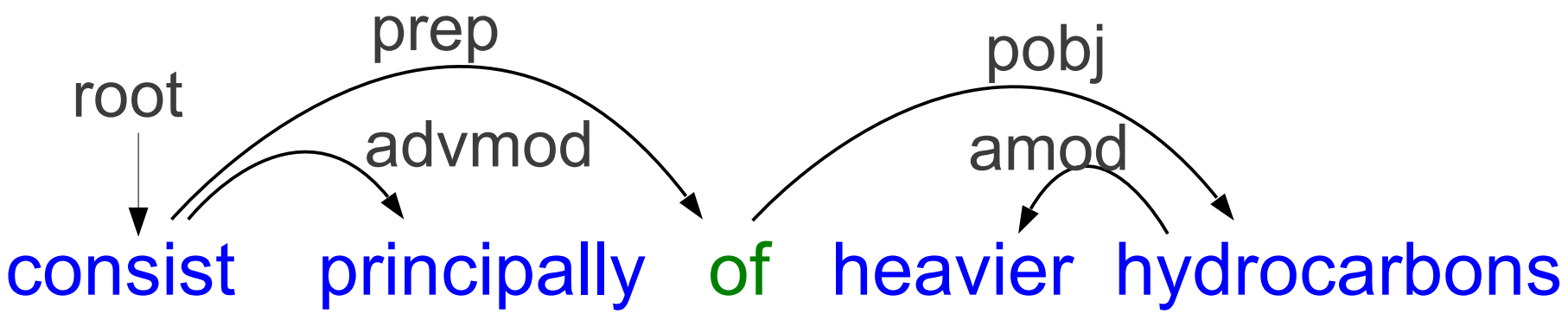
triarcs: four content words



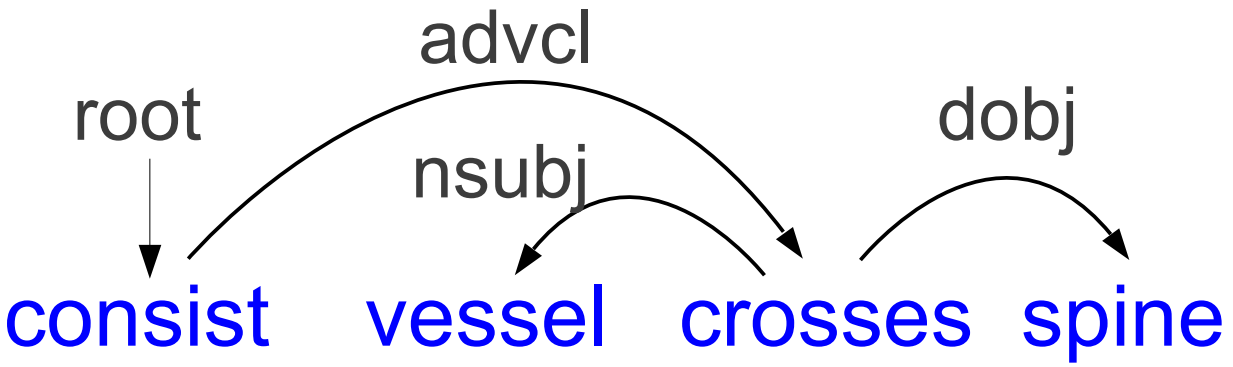
triarcs: four content words



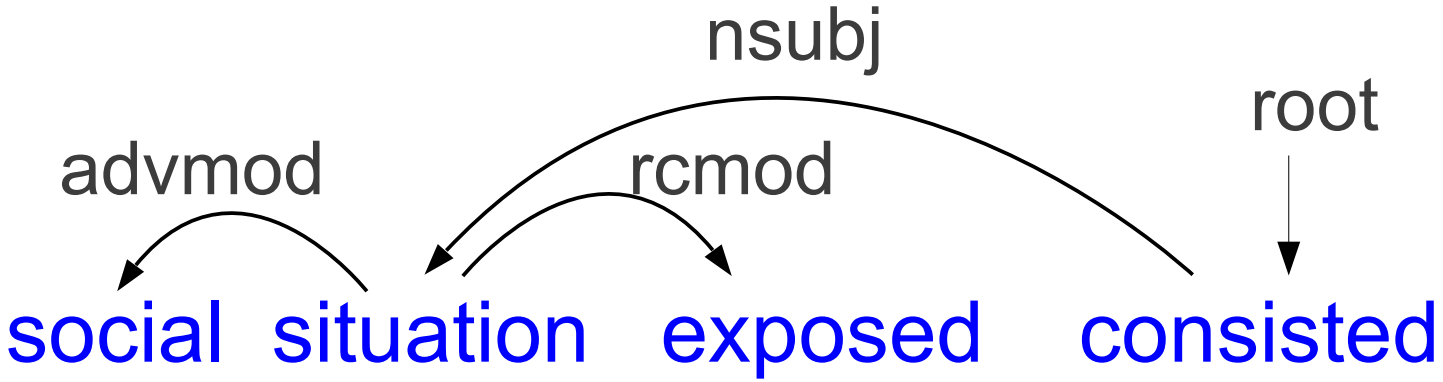
triarcs: four content words



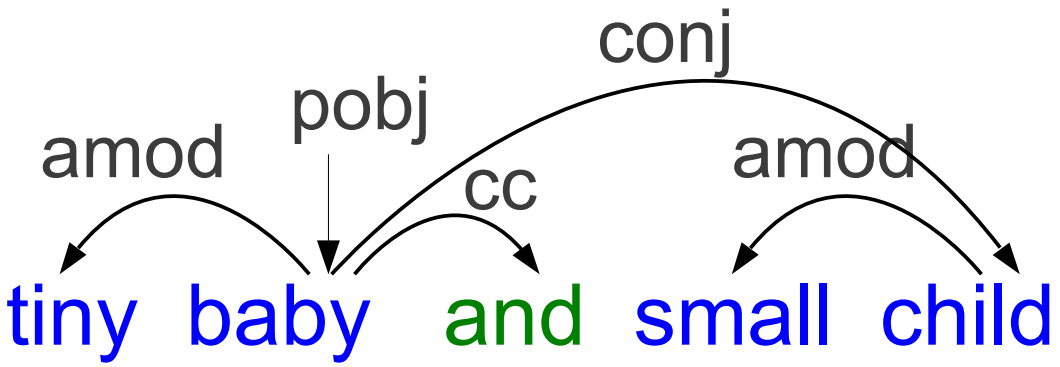
triarcs: four content words



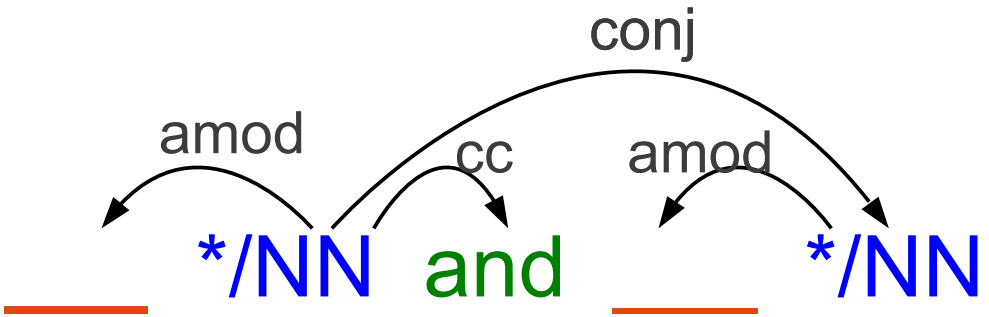
triarcs: four content words



triarcs: four content words

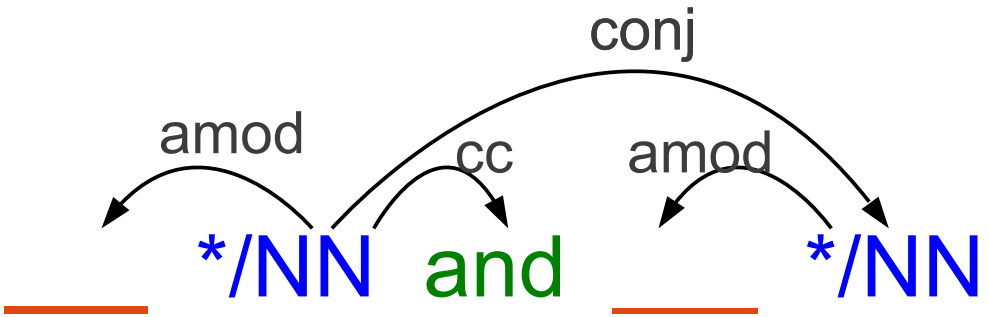


Adjectivial modifiers of coordinated nouns



parallelism?

Adjectivial modifiers of coordinated nouns



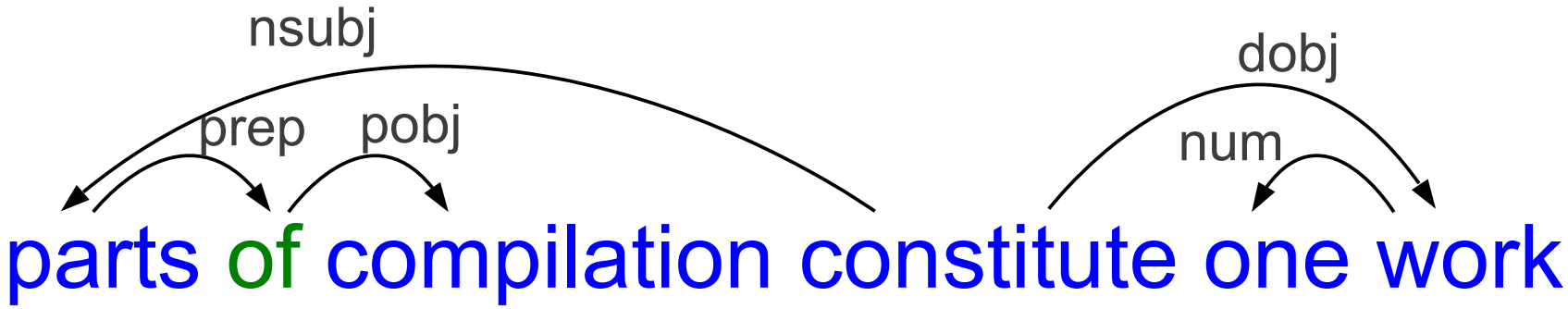
parallelism!!

347380	late	and	early
318353	new	and	new
143298	good	and	good
123184	high	and	low
119851	social	and	social
87337	high	and	high
83516	%	and	%
82964	human	and	human
78980	low	and	high
74488	different		different
72617	same	and	same
68260	great	and	great
67055	good	and	bad
62282	many	and	many
61822	other	and	other
61126	own	and	own
58781	more	and	more
57556	young	and	young
57392	black	and	white
54690	white	and	black

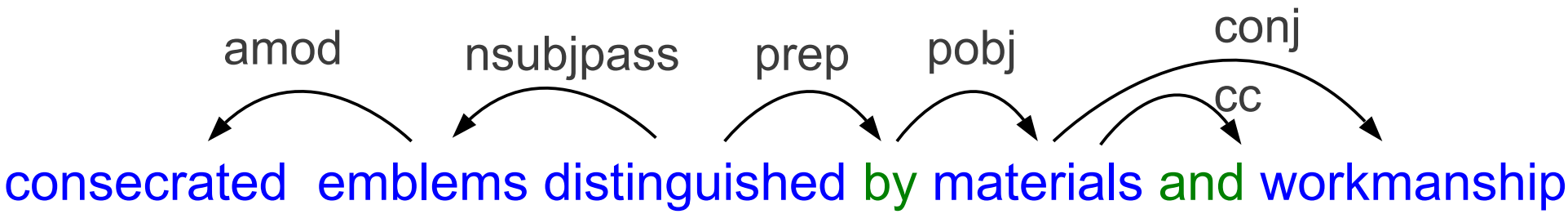
quadarcs: five content words
(but restricted to specific patterns)

quadarcs: five content words

(but restricted to specific patterns)

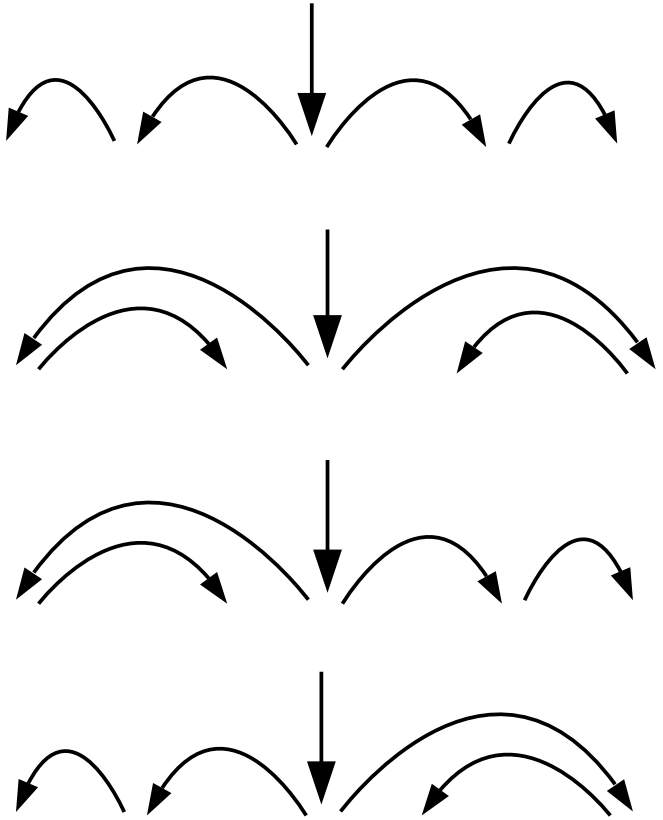


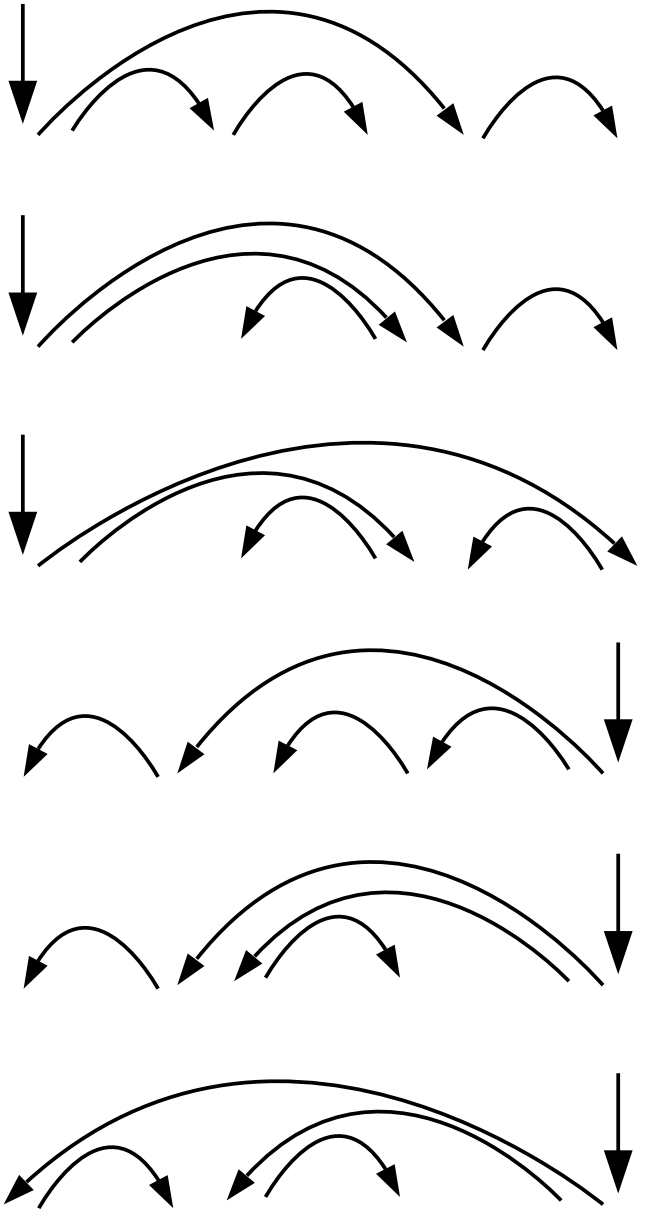
quadarcs: five content words
(but restricted to specific patterns)



quadarcs: five content words

(but restricted to specific patterns)



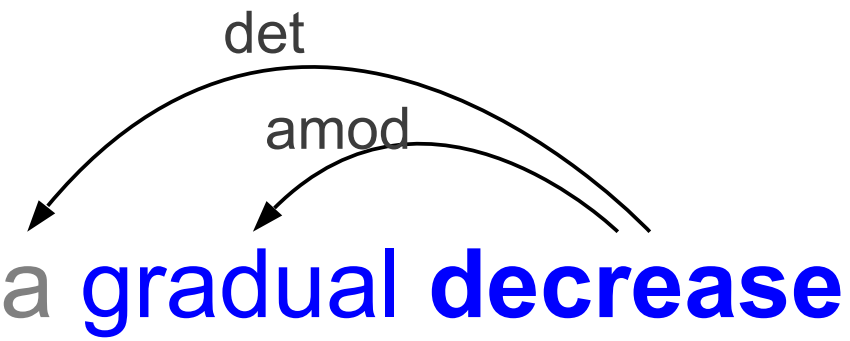


A content-word root,
with two chains of
two content-words each

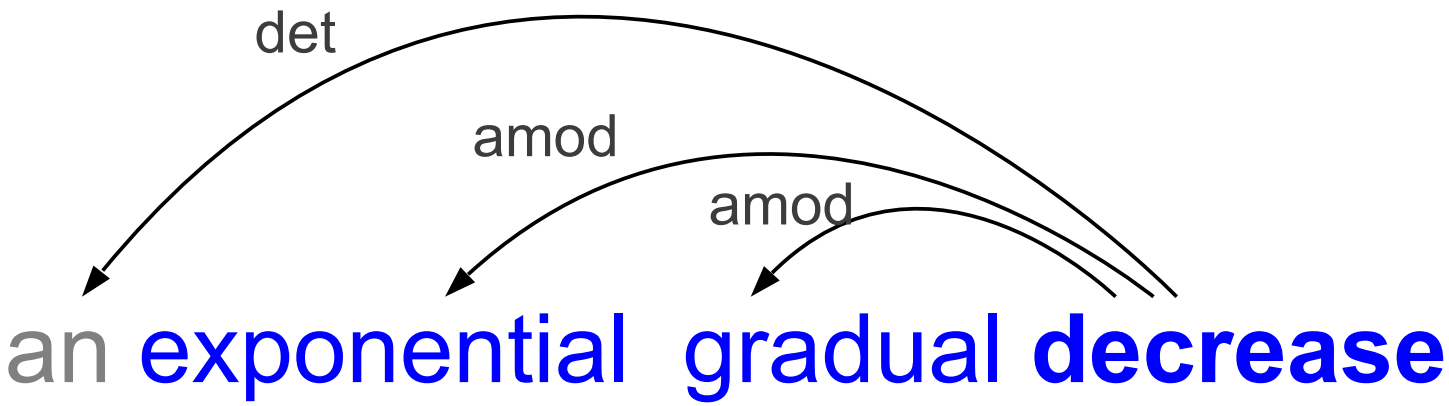
There are also the *extended* versions
(with functional markers)
of biarcs, triarcs and quadarcs

nounargs: noun and all its modifiers
(+ all functional markers)

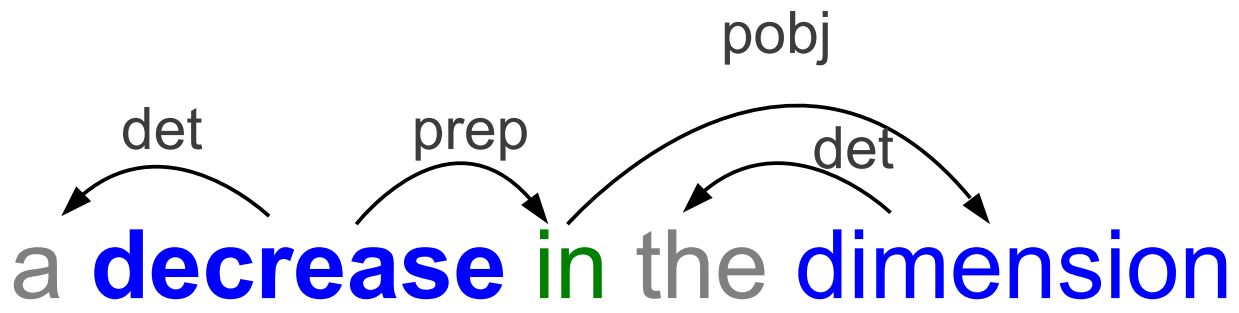
nounargs: noun and all its modifiers
(+ all functional markers)



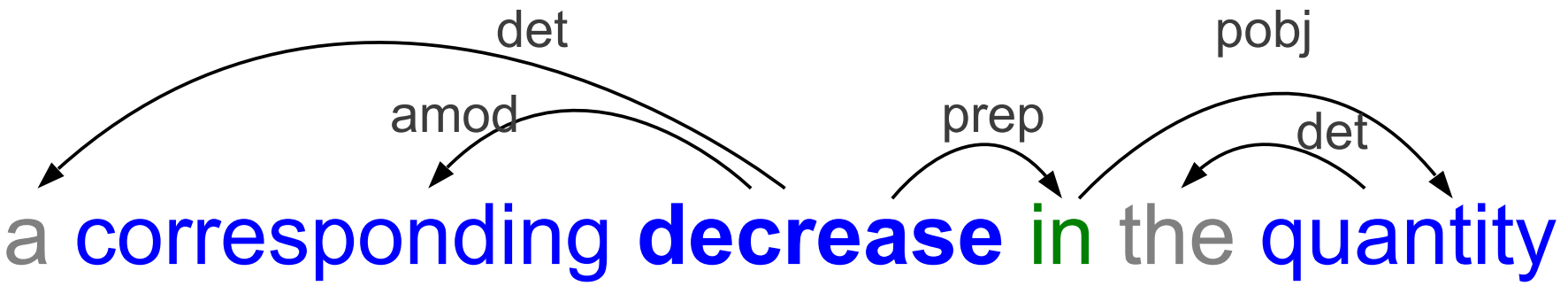
nounargs: noun and all its modifiers
(+ all functional markers)



nounargs: noun and all its modifiers
(+ all functional markers)



nounargs: noun and all its modifiers
(+ all functional markers)



nounargs: noun and all its modifiers
(+ all functional markers)

can be used for estimating
dependency language models

nounargs: noun and all its modifiers (+ all functional markers)

other interesting questions:

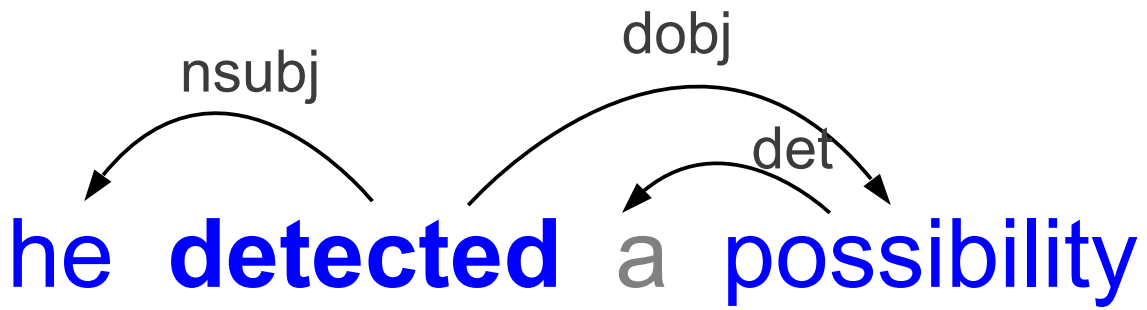
PP co-occurrence patterns?

adjectivial co-occurrence?

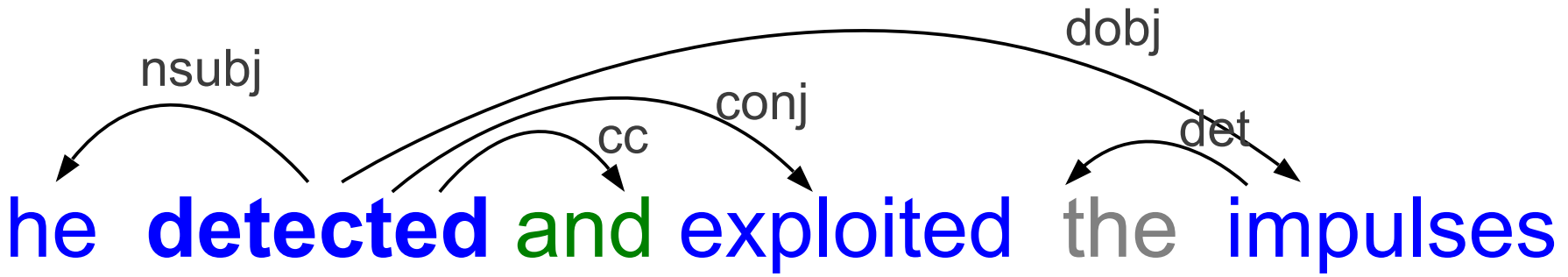
definiteness patterns?

verbargs: verb and all its modifiers
(+ all functional markers)

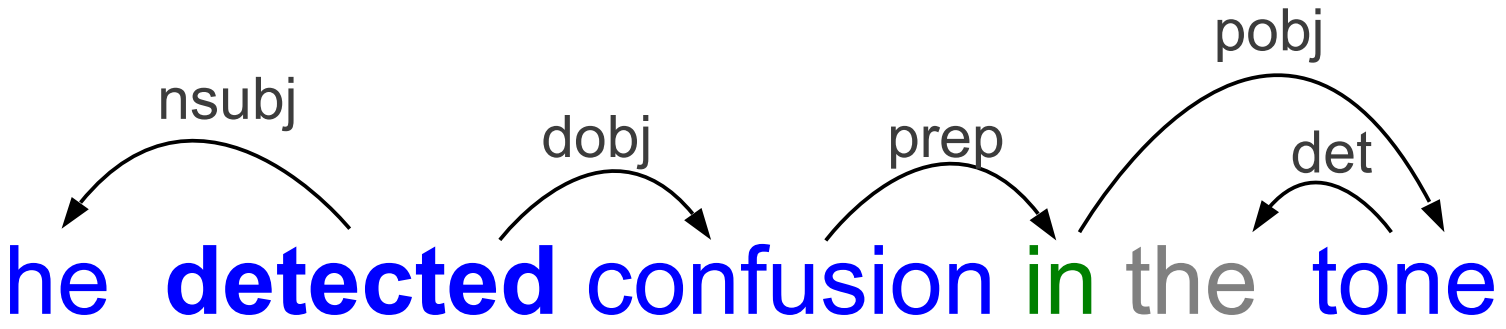
verbargs: verb and all its modifiers
(+ all functional markers)



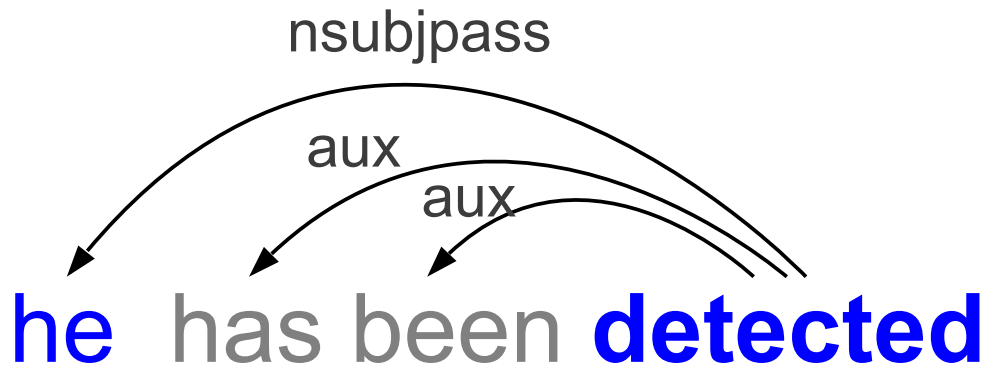
verbargs: verb and all its modifiers
(+ all functional markers)



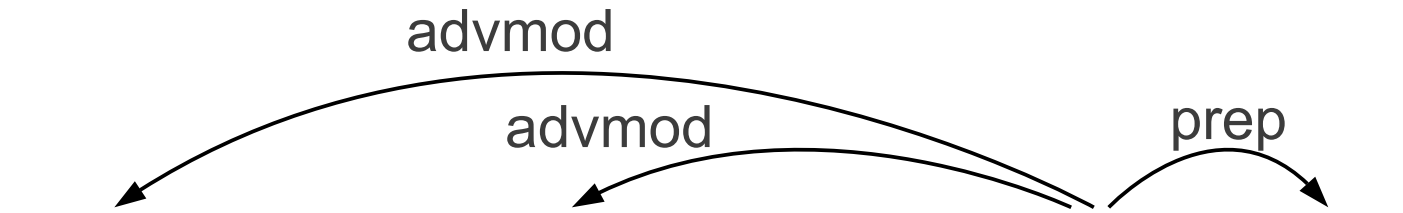
verbargs: verb and all its modifiers
(+ all functional markers)



verbargs: verb and all its modifiers
(+ all functional markers)



verbargs: verb and all its modifiers
(+ all functional markers)



often	*unfairly*	dealt	with
only	*briefly*	dealt	with
only	*marginally*	dealt	with
otherwise	*severely*	dealt	with
rather	*hardly*	dealt	with
so	*directly*	dealt	with
so	*easily*	dealt	with

verbargs: verb and all its modifiers
(+ all functional markers)

he conquered
he conquered egypt
he conquered for himself a duchy
he conquered himself with an effort
he conquered this obstacle by manufacturing
he conquered during his journey

verbargs: verb and all its modifiers (+ all functional markers)

frame induction?

better SRL?

verb groups?

reasons to use the syntactic ngrams corpus in your research:

freely available and easy to obtain

based on state-of-the-art parser

x100 bigger than previous efforts and ready-to-use

a standarized dataset means you can compare yourself to others

W. Schuler, PhD.
Cognitive Computational Linguist
Ohio State University

"I can use this to improve
my super cool brain-emulating parser!" *

* I my be paraphrasing a little

now for the really cool stuff!

Time-series information

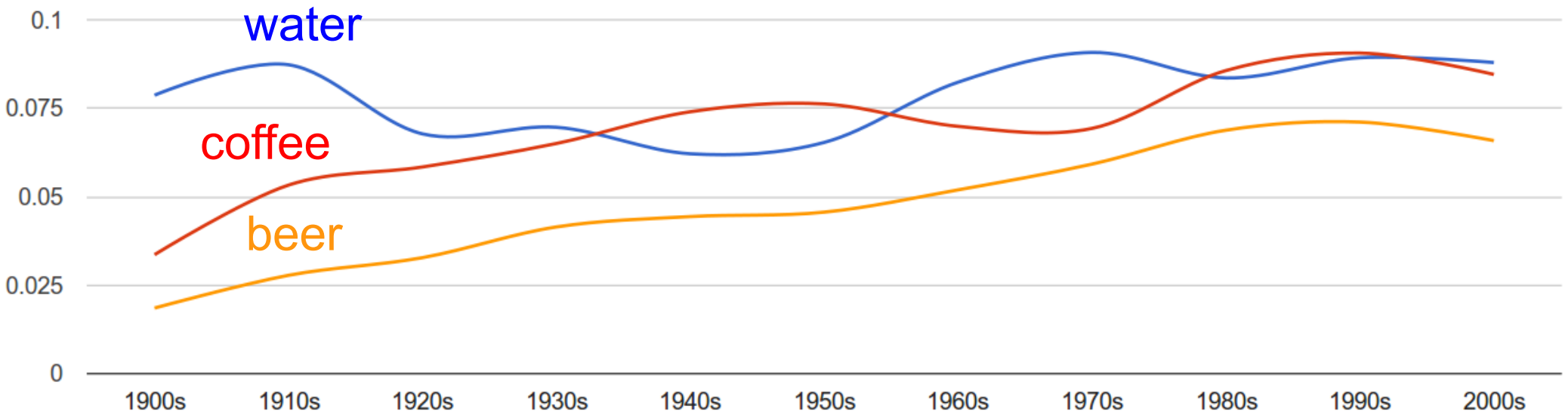
for each syntactic ngram,
we have counts **broken down by year**

a very strong tool for **studying change over time**

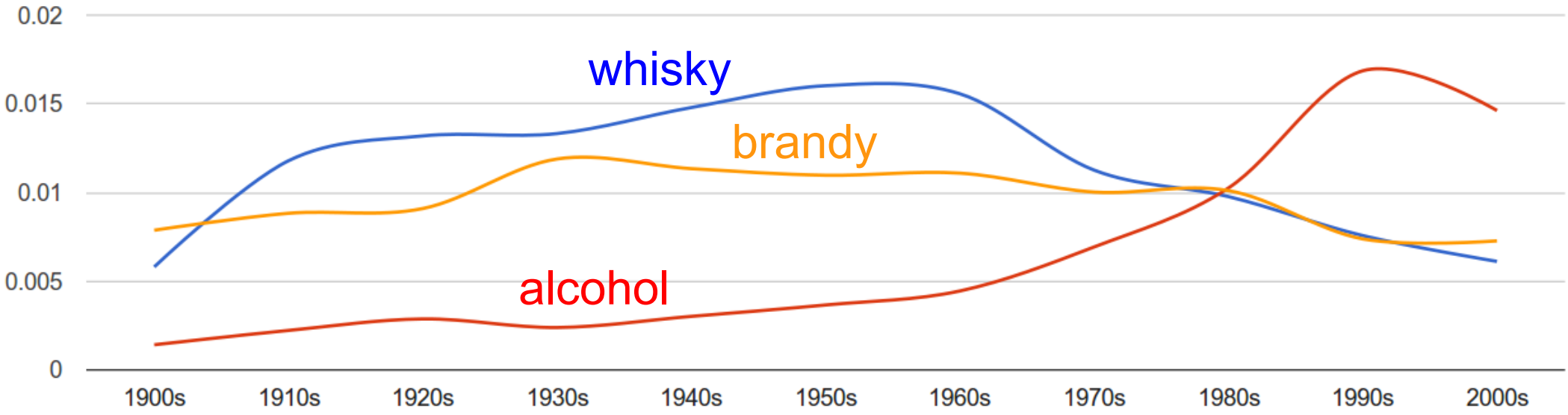
Simple stuff: what do people drink?

count noun object of drink/drank
when subject is proper-noun or pronoun,
graph by decade

drinking trends



drinking trends



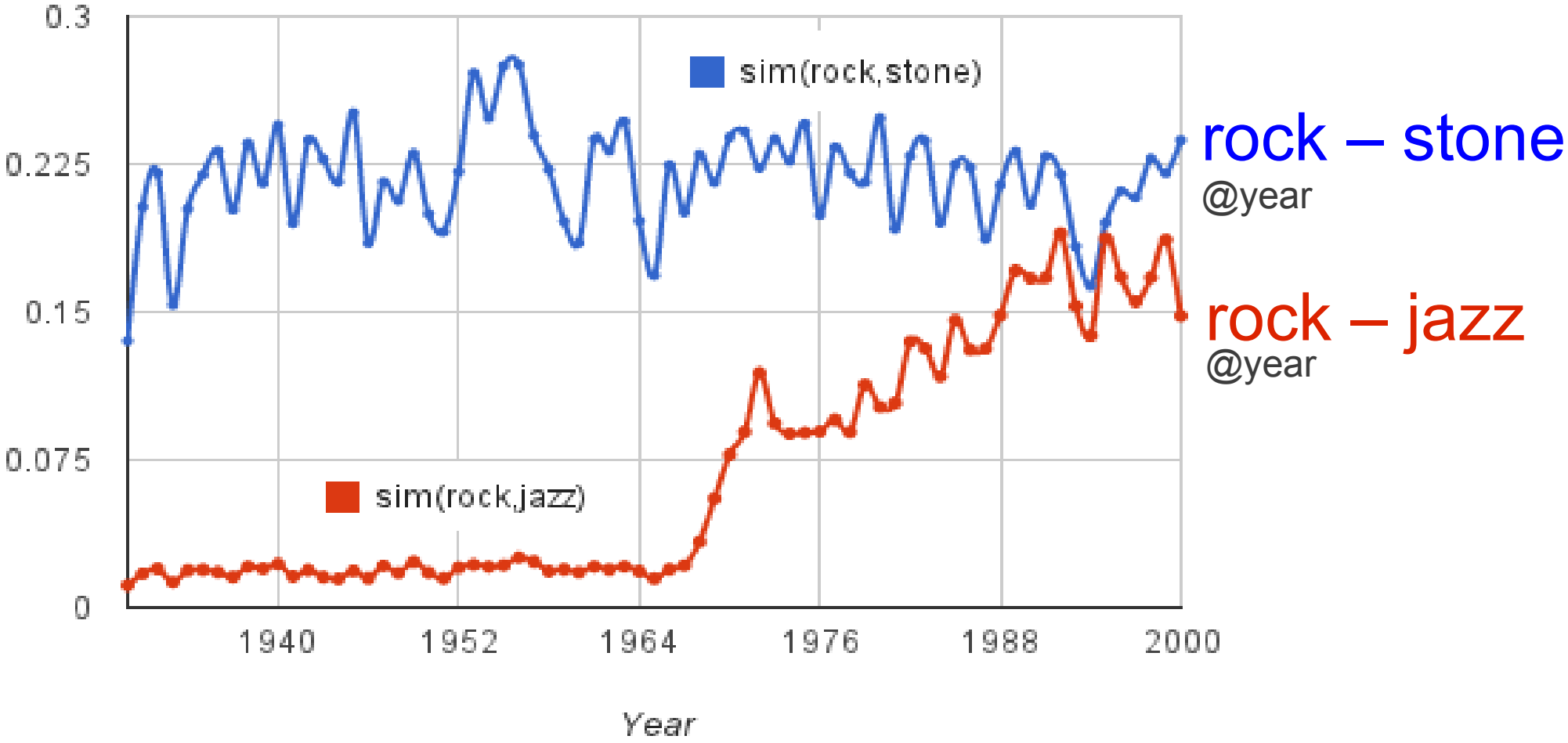
Somewhat less simple stuff:
but still fairly simple

change in word meaning over time
through distributional similarity
(hand picked example)

Compute distributional (cosine) similarity between:

rock – stone
@year

rock – jazz
@year


0.3
0.225
0.15
0.075
0
sim(rock,stone)
sim(rock,jazz)
1940
1952
1964
1976
1988
2000
Year
rock – stone
@year
rock – jazz
@year

many other fascinating questions:

language evolution!

number of senses over time

syntactic change over time

modification patterns over time

polarity change over time

…

the tip of an iceberg

What can YOU do
with ready-to-use, time-indexed
syntactic dependencies
from 350 billion words?

Sizes

- Eng / all: ~320GB compressed
  all + 1M + gb + us + fiction: ~680GB

• arcs	919M items	extended:1.08B
• biarcs	1.78B items	extended:1.62B
• triarcs	1.87B items	extended:1.71B
• quadarcs	187M items	extended:180M
• nounargs	275M items	unlex: 195M
• verbargs	130M items	unlex: 114M