

From distributional semantics to formal grammar and back

Chung-chieh Shan
Indiana University
Formal Grammar
11 August 2013



Approaches to semantics

“In order to say what a meaning *is*,
we may first ask what a meaning *does*,
and then find something that does that.” —David Lewis

Approaches to semantics

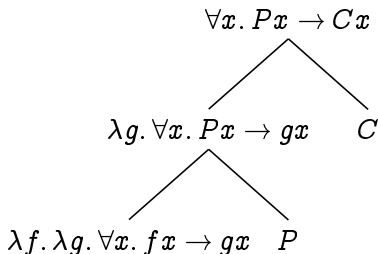
“In order to say what a meaning *is*,
we may first ask what a meaning *does*,
and then find something that does that.” —David Lewis

Truth, entailment

Every person cried. \models Every professor cried.

A person cried. $\not\models$ A professor cried.

Formal semantics



Approaches to semantics

“In order to say what a meaning *is*,
we may first ask what a meaning *does*,
and then find something that does that.” —David Lewis

Concepts, similarity

ambulance \sim battleship

ambulance $\not\sim$ bookstore

Distributional semantics

	abandon	abdominal	ability	academic	accept	...
ambulance	27	10	50	17	130	...
battleship	35	0	32	1	25	...
bookstore	5	0	6	33	13	...
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\ddots

Approaches to semantics

“In order to say what a meaning *is*,
we may first ask what a meaning *does*,
and then find something that does that.” —David Lewis

What is meaning?

- ▶ Over-constrained problem?

Meanings do lots of things.

- ▶ Under-constrained problem?

Passive observations alone can't distinguish what meanings are from how meanings are used.

- ▶ Seek unity in diversity. . .

At least, meanings should support both entailment and similarity judgments, possibly with the help of world knowledge.

Distributional semantics for entailment among words

For each word w , rank contexts c by descending $\frac{\Pr(c | w)}{\Pr(c)} > 1$.

“pointwise mutual information”

Distributional semantics for entailment among words

For each word w , rank contexts c by descending $\frac{\Pr(c | w)}{\Pr(c)} > 1$.

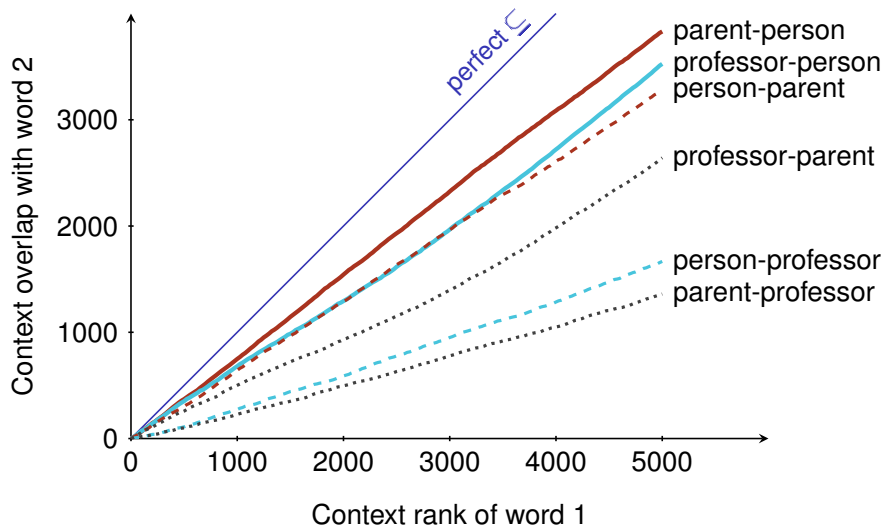
“pointwise mutual information”

parent argcount_n arglist_n arglist_j phane_n specity_n qdisc_n carthy_n
parents-to-be_n non-resident_j step-parent_n tc_n ballons_n
eliza_n symptons_n adoptive_j stepparent_n nonresident_j
home-school_n scabrid_n petiolule_n ...

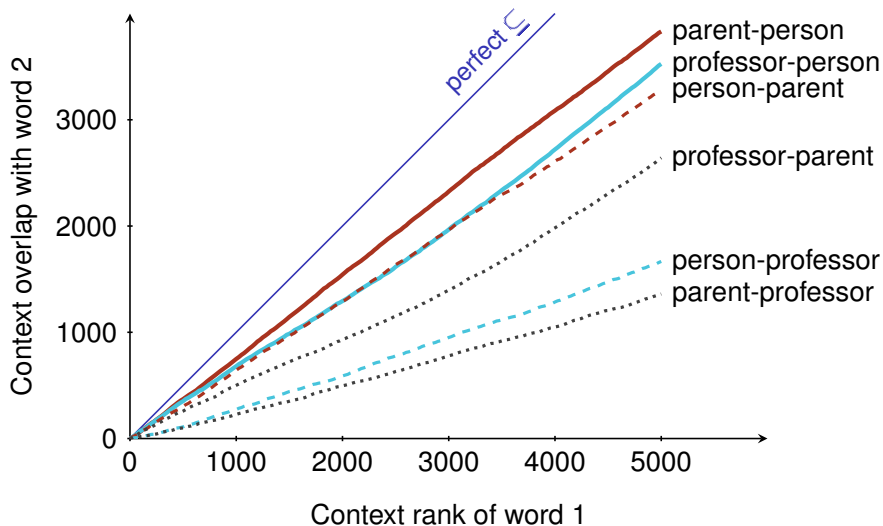
person anglia_n first-mentioned_j unascertained_j enure_v
deposit-taking_j bonis_n iconclass_j cotswolds_n aforesaid_n
haver_v foresaid_j gha_n sub-paragraphs_n enacted_j geest_j
non-medicinal_j sub-paragraph_n intimation_n arrestment_n
incumbrance_n ...

professor william_n extraordinarius_n ordinarius_n francis_n reid_n
emeritus_n emeritus_j derwent_n regius_n laurence_n edward_n
carisoprodol_n adjunct_j winston_n privatdozent_j edward_j
xanax_n tenure_v cialis_n florence_n ...

Distributional semantics for entailment among words



Distributional semantics for entailment among words



More sophisticated: *Kullback-Leibler divergence*,
skew divergence (Lee), *balAPinc* (Kotlerman et al.), ...

Sparse data strikes back

Successes for words and short phrases:

- ▶ similarity
- ▶ entailment
- ▶ sentiment

'common sense' from noisy large corpora

For *long, rare, episodic* phrases and sentences, need




- ▶ syntactic structure
- ▶ semantic reference
- ▶ pragmatic context
- ▶ grounding in other information sources

'linguistic generalization' from poor stimulus

This need goes way back—

From documents \times terms to words \times contexts

Information retrieval started with bag of terms in each document.
Stopwords, stemming, tagging; TF-IDF.

	abandon	abdominal	ability	academic	...
	27	10	50	17	...
	35	0	32	1	...
	5	0	6	33	...
⋮	⋮	⋮	⋮	⋮	⋮

From documents \times terms to words \times contexts

Information retrieval started with bag of terms in each document.
Stopwords, stemming, tagging; TF-IDF. Dimensionality reduction reveals topics.

$$\begin{array}{c} \text{document 1} \\ \text{document 2} \\ \text{document 3} \\ \vdots \end{array} \begin{pmatrix} \text{abandon} & \text{abdominal} & \text{ability} & \text{academic} & \dots \\ 27 & 10 & 50 & 17 & \dots \\ 35 & 0 & 32 & 1 & \dots \\ 5 & 0 & 6 & 33 & \dots \\ \vdots & \vdots & \vdots & \ddots & \end{pmatrix}$$
$$= \begin{array}{c} \text{document 1} \\ \text{document 2} \\ \text{document 3} \\ \vdots \end{array} \begin{pmatrix} \phantom{\text{abandon}} \\ \phantom{\text{abdominal}} \\ \phantom{\text{ability}} \\ \phantom{\text{academic}} \\ \end{pmatrix} \times \begin{pmatrix} \text{abandon} & \text{abdominal} & \text{ability} & \text{academic} & \dots \\ \phantom{\text{abandon}} & \phantom{\text{abdominal}} & \phantom{\text{ability}} & \phantom{\text{academic}} & \\ \phantom{\text{abandon}} & \phantom{\text{abdominal}} & \phantom{\text{ability}} & \phantom{\text{academic}} & \\ \phantom{\text{abandon}} & \phantom{\text{abdominal}} & \phantom{\text{ability}} & \phantom{\text{academic}} & \\ \phantom{\text{abandon}} & \phantom{\text{abdominal}} & \phantom{\text{ability}} & \phantom{\text{academic}} & \end{pmatrix}$$

From documents \times terms to words \times contexts

Information retrieval started with bag of terms in each document. Stopwords, stemming, tagging; TF-IDF. Dimensionality reduction reveals topics. Now rows are phrases and columns are contexts.

$$\begin{array}{l} \text{ambulance} \\ \text{battleship} \\ \text{bookstore} \\ \vdots \end{array} \begin{pmatrix} \text{abandon} & \text{abdominal} & \text{ability} & \text{academic} & \dots \\ 27 & 10 & 50 & 17 & \dots \\ 35 & 0 & 32 & 1 & \dots \\ 5 & 0 & 6 & 33 & \dots \\ \vdots & \vdots & \vdots & \ddots & \end{pmatrix}$$

$$= \begin{array}{l} \text{ambulance} \\ \text{battleship} \\ \text{bookstore} \\ \vdots \end{array} \begin{pmatrix} \phantom{\text{abandon}} \\ \phantom{\text{abdominal}} \\ \phantom{\text{ability}} \\ \phantom{\text{academic}} \\ \end{pmatrix} \times \begin{pmatrix} \text{abandon} & \text{abdominal} & \text{ability} & \text{academic} & \dots \\ \phantom{\text{abandon}} & \phantom{\text{abdominal}} & \phantom{\text{ability}} & \phantom{\text{academic}} & \\ \phantom{\text{abandon}} & \phantom{\text{abdominal}} & \phantom{\text{ability}} & \phantom{\text{academic}} & \\ \phantom{\text{abandon}} & \phantom{\text{abdominal}} & \phantom{\text{ability}} & \phantom{\text{academic}} & \\ \phantom{\text{abandon}} & \phantom{\text{abdominal}} & \phantom{\text{ability}} & \phantom{\text{academic}} & \end{pmatrix}$$

Composite phrases?

Need structure: substitution? locality? types? compositionality?

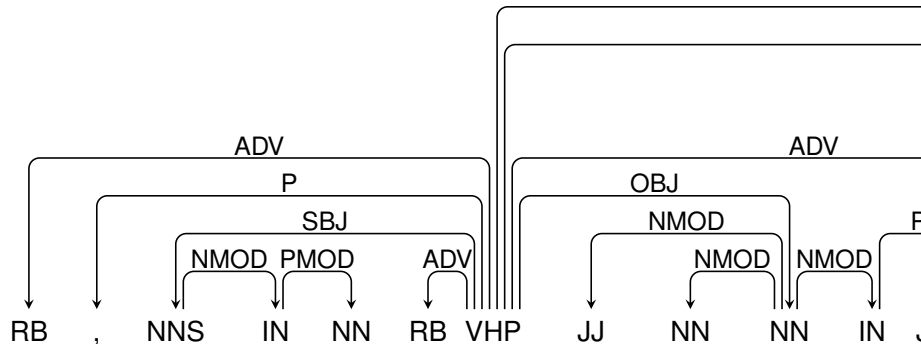
RB , NNS IN NN RB VHP JJ NN NN IN ,

however , individual with autism also have abnormal brain activation in m

However , individuals with autism also have abnormal brain activation in m

Composite phrases?

Need structure: substitution? locality? types? compositionality?

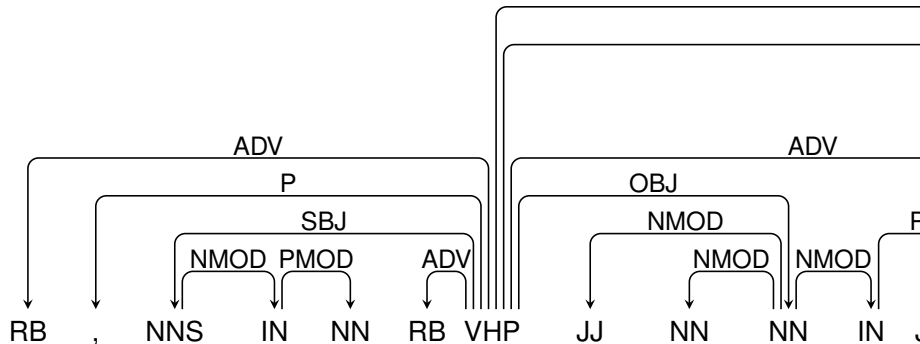


however, individual with autism also have abnormal brain activation in m...

However, individuals with autism also have abnormal brain activation in m...

Composite phrases?

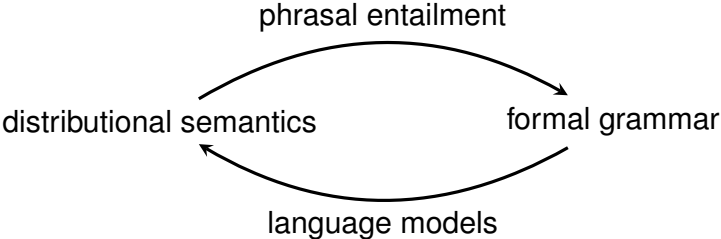
Need structure: substitution? locality? types? compositionality?

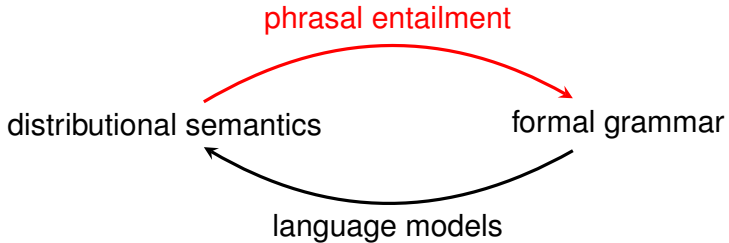


however, individual with autism also have abnormal brain activation in m...

However, individuals with autism also have abnormal brain activation in m...

To cope with sparse data, NLP (parsing, translation, compression) applies linguistic insight (typing, factoring, smoothing).





Above the word level

Phrases have corpus distributions too!

N	cat
AN	white cat
QN	every cat

Above the word level

Phrases have corpus distributions too! But **N** \approx **AN** $\not\approx$ **QN**

		Syntactic category
N	cat	N
AN	white cat	N
QN	every cat	QP

Above the word level

Phrases have corpus distributions too! But **N** \approx **AN** $\not\approx$ **QN**

		Syntactic category	Semantic type
N	cat	N	$e \rightarrow t$
AN	white cat	N	$e \rightarrow t$
QN	every cat	QP	$(e \rightarrow t) \rightarrow t$

Above the word level

Phrases have corpus distributions too! But $N \approx AN \not\approx QN$

		Syntactic category	Semantic type
N	cat	N	$e \rightarrow t$
AN	white cat	N	$e \rightarrow t$
AAN	big white cat	N	$e \rightarrow t$
QN	every cat	QP	$(e \rightarrow t) \rightarrow t$
QAN	every big cat	QP	$(e \rightarrow t) \rightarrow t$
* AQN	big every cat		
* QQN	some every cat		

Our questions

Entailment among **composite phrases** rather than nouns?

Entailment among logical words rather than content words?

Different entailment relations at different semantic types?

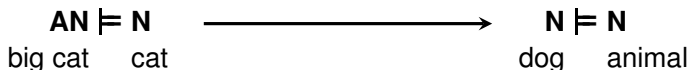


Our questions

Entailment among **composite phrases** rather than nouns?

Entailment among **logical words** rather than content words?

Different entailment relations at different semantic types?

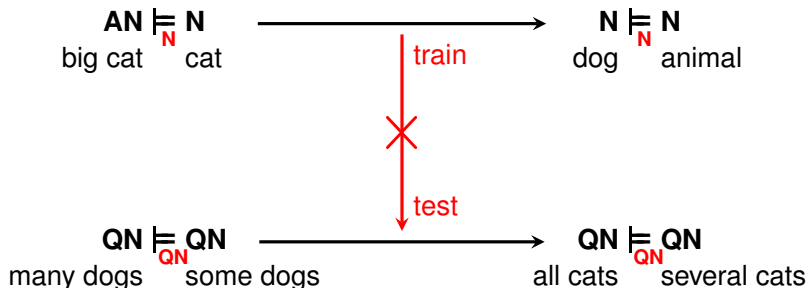


Our questions

Entailment among **composite phrases** rather than nouns?

Entailment among **logical words** rather than content words?

Different entailment relations at different **semantic types**?



Our semantic space

BNC, WackyPedia, ukWaC

↓ TreeTagger (Schmid)

lemmatized, POS-tagged tokens (2.8G)

↓ words and phrases in the same sentence

most frequent
A, N, V (27K)

AN
QN
A
Q
N
(48K)

$$\left(\begin{array}{c} \#(c, w) \end{array} \right)$$

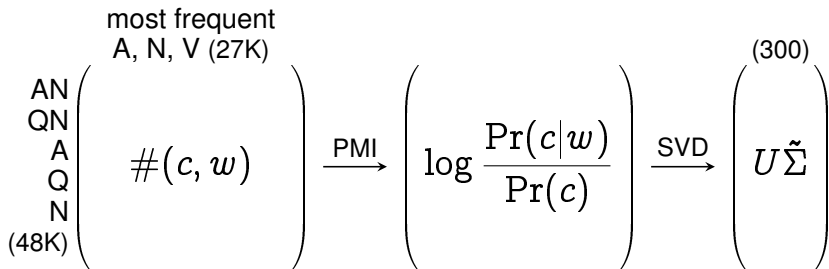
Our semantic space

BNC, WackyPedia, ukWaC

↓ TreeTagger (Schmid)

lemmatized, POS-tagged tokens (2.8G)

↓ words and phrases in the same sentence



Our semantic space

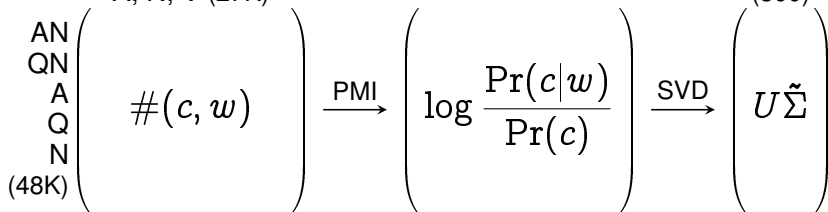
BNC, WackyPedia, ukWaC

↓ TreeTagger (Schmid)

lemmatized, POS-tagged tokens (2.8G)

↓ words and phrases in the same sentence

most frequent
A, N, V (27K)



frequency
baseline

cosine
baseline

baIAPinc

SVM

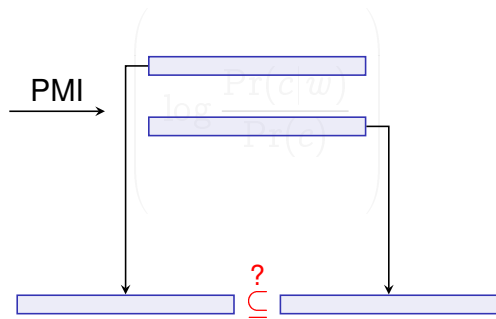
Our entailment classifiers

$$\xrightarrow{\text{PMI}} \left(\log \frac{\Pr(c|w)}{\Pr(c)} \right)$$

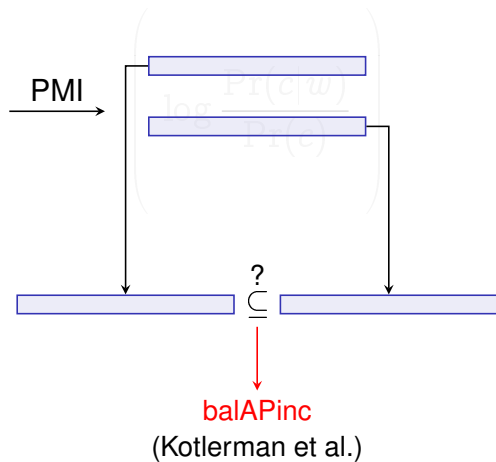
Our entailment classifiers

$$\xrightarrow{\text{PMI}} \left(\begin{array}{c} \text{[redacted]} \\ \log \frac{\text{Pr}(c|w)}{\text{Pr}(c)} \\ \text{[redacted]} \end{array} \right)$$

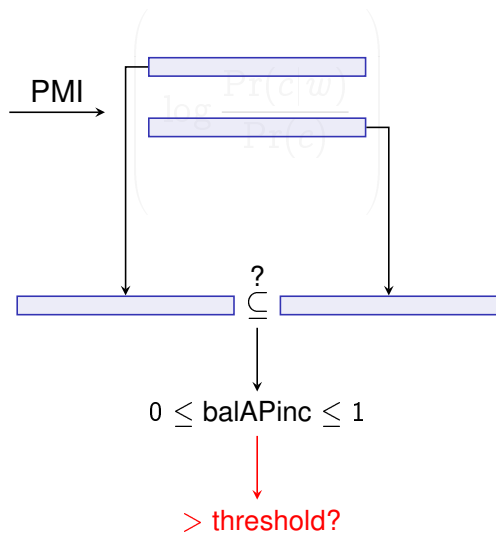
Our entailment classifiers



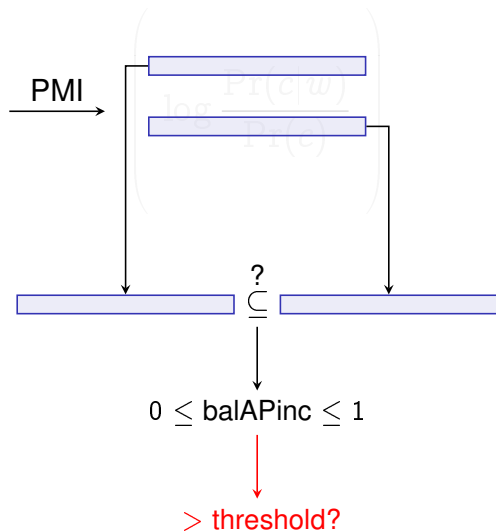
Our entailment classifiers



Our entailment classifiers

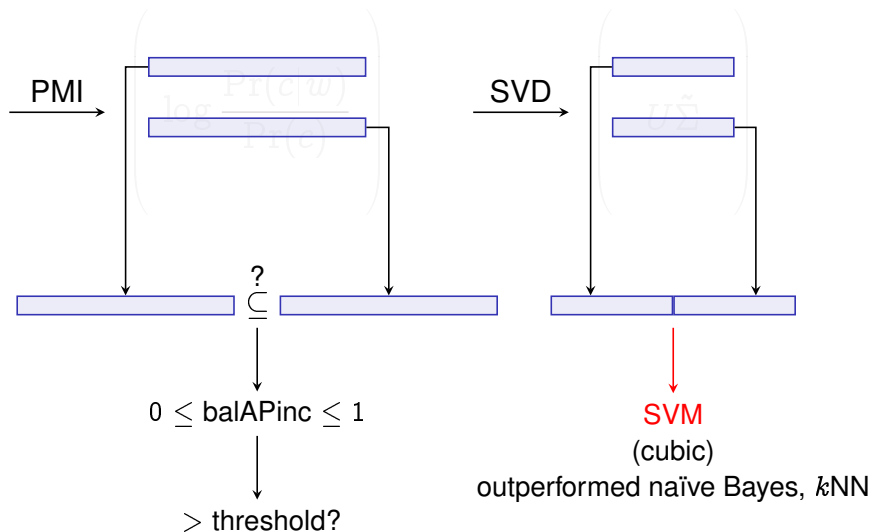


Our entailment classifiers



Train	Test
AN \models N	N \models N
QN \models QN	QN \models QN
AN \models N	QN \models QN

Our entailment classifiers



Our data sets

WordNet



pope \models spiritual_leader

spiritual_leader \models leader

cat \models feline

feline \models carnivore

\vdots

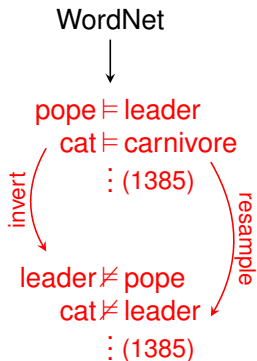
Our data sets

WordNet



pope \models leader
cat \models carnivore
 \vdots (1385)

Our data sets

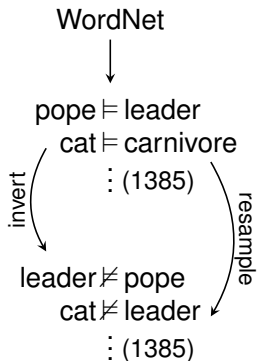


Our data sets

most frequent



big
former
⋮ (300)



Our data sets

most frequent



big

~~former~~

⋮ (256)

WordNet



pope \models leader

cat \models carnivore

⋮ (1385)

invert

leader $\not\models$ pope

cat $\not\models$ leader

⋮ (1385)

resample

Our data sets

most frequent

↓
big
~~former~~
: (256)

BLESS

↓
apple
shirt
: (200)

WordNet

↓
pope \models leader
cat \models carnivore
: (1385)
invert ↙
leader $\not\models$ pope
cat $\not\models$ leader
: (1385)
↘ resample

↙
big apple \models apple
big shirt \models shirt
: (1246)
↘ resample

↙ resample
big apple $\not\models$ shirt
big shirt $\not\models$ apple
: (1244)

Our data sets

most frequent



big
~~former~~
⋮ (256)

BLESS



apple
shirt
⋮ (200)

WordNet



pope \models leader
cat \models carnivore
⋮ (1385)

invert

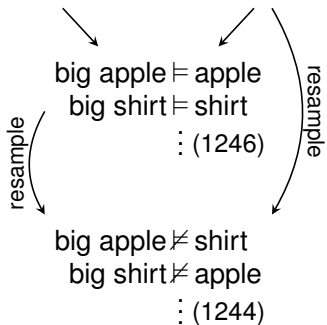
leader $\not\models$ pope
cat $\not\models$ leader
⋮ (1385)

resample

most frequent



all
both
each
either
every
few
many
most
much
no
several
some
⋮



Our data sets

most frequent



big
~~former~~
⋮ (256)

big apple \models apple
big shirt \models shirt
⋮ (1246)

resample

big apple $\not\models$ shirt
big shirt $\not\models$ apple
⋮ (1244)

BLESS



apple
shirt
⋮ (200)

resample

WordNet



pope \models leader
cat \models carnivore
⋮ (1385)

invert

leader $\not\models$ pope
cat $\not\models$ leader
⋮ (1385)

resample

most frequent



all \models some
many \models several
⋮ (13)

some $\not\models$ every
both $\not\models$ many
⋮ (17)

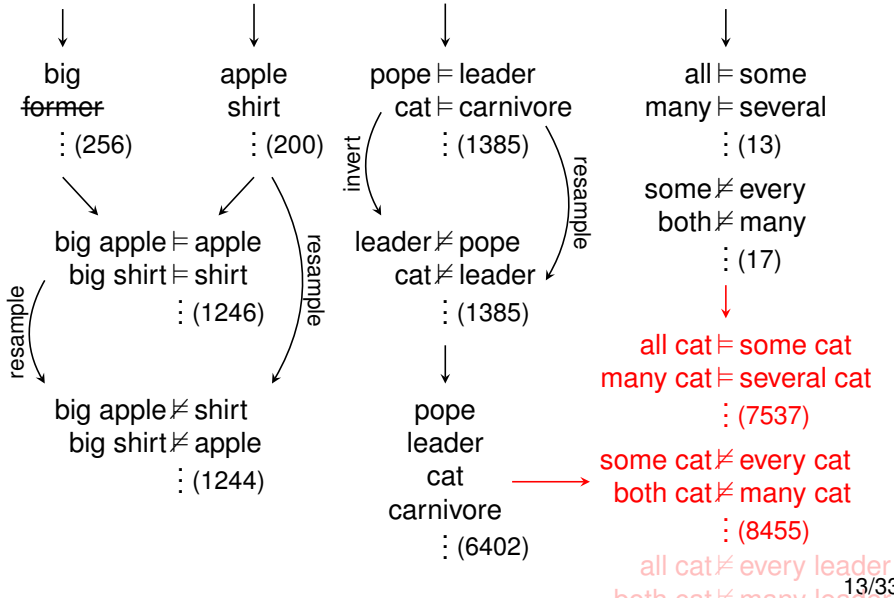
Our data sets

most frequent

BLESS

WordNet

most frequent



Our data sets

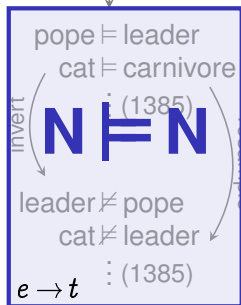
most frequent

big
former
⋮ (256)

BLESS

apple
shirt
⋮ (200)

WordNet

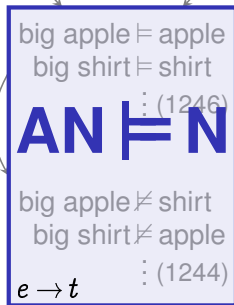


most frequent

all \models some
many \models several
⋮ (13)

some $\not\models$ every
both $\not\models$ many
⋮ (17)

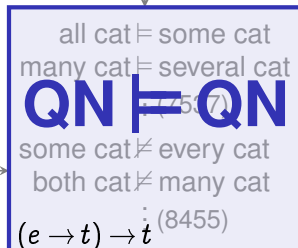
resample



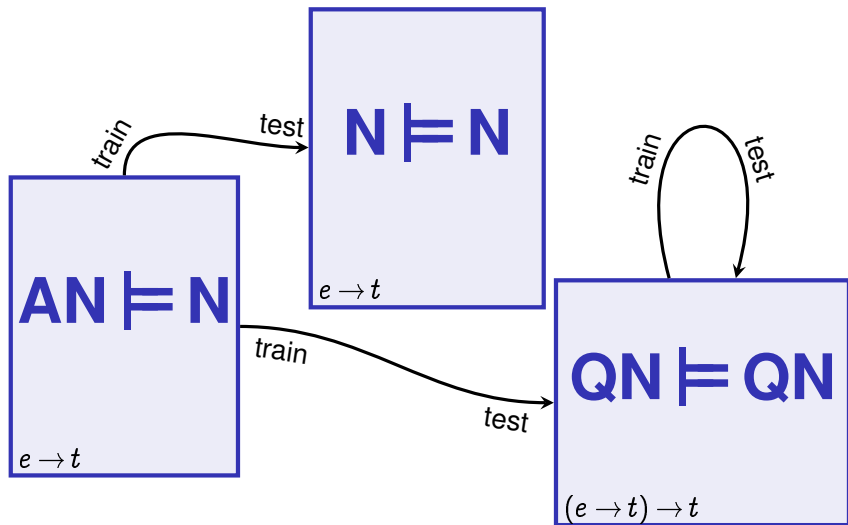
resample

pope
leader
cat
carnivore
⋮ (6402)

resample



Our data sets



Results at noun type

	P	R	F	Accuracy	(95% C.I.)
SVM _{upper}	88.6	88.6	88.5	88.6	(87.3–89.7)
balAPinc _{AN≠N}	65.2	87.5	74.7	70.4	(68.7–72.1)
balAPinc _{upper}	64.4	90.0	75.1	70.1	(68.4–71.8)
SVM _{AN≠N}	69.3	69.3	69.3	69.3	(67.6–71.0)
cos(N ₁ , N ₂)	57.7	57.6	57.5	57.6	(55.8–59.5)
fq(N ₁) < fq(N ₂)	52.1	52.1	51.8	53.3	(51.4–55.2)

Holding out QN data

⋈	all	both	each	either	every	few	many	most	much	no	several	some
all							+	+			+	+
both				+			-	-			-	+
each												+
either		-										
every							+					
few	-						-					
many	-				-			-		-	+	+
most							+					
much												+
no												
several	-				-	-						+
some	-	-			-		-					

Holding out QN data

⋈	all	both	each	either	every	few	many	most	much	no	several	some
all							+	+			+	+
both				+			-	-			-	+
each												+
either		-										
every							+					
few	-						-					
many	-				-			-		-	+	+
most							+					
much												+
no												
several	-				-	-						+
some	-		-		-		-					

pair-out

Holding out QN data

↙	all	both	each	either	every	few	many	most	much	no	several	some
all							+	+			+	+
both				+			-	-			-	+
each												+
either		-										
every							+					
few	-						-					
many	-							-		-	+	+
most							+					
much												+
no												
several	-				-	-						+
some	-	-	-	-	-	-	-					

Results at quantifier type

	P	R	F	Accuracy	(95% C.I.)
$SVM_{\text{pair-out}}$	76.7	77.0	76.8	78.1	(77.5–78.8)
$SVM_{\text{quantifier-out}}$	70.1	65.3	68.0	71.0	(70.3–71.7)
$SVM_{\text{pair-out}}^Q$	67.9	69.8	68.9	70.2	(69.5–70.9)
$SVM_{\text{quantifier-out}}^Q$	53.3	52.9	53.1	56.0	(55.2–56.8)
$\cos(QN_1, QN_2)$	52.9	52.3	52.3	53.1	(52.3–53.9)
$\text{balAPinc}_{AN \neq N}$	46.7	5.6	10.0	52.5	(51.7–53.3)
$SVM_{AN \neq N}$	2.8	42.9	5.2	52.4	(51.7–53.2)
$\text{fq}(QN_1) < \text{fq}(QN_2)$	51.0	47.4	49.1	50.2	(49.4–51.0)
$\text{balAPinc}_{\text{upper}}$	47.1	100	64.1	47.2	(46.4–47.9)

Holding out each quantifier

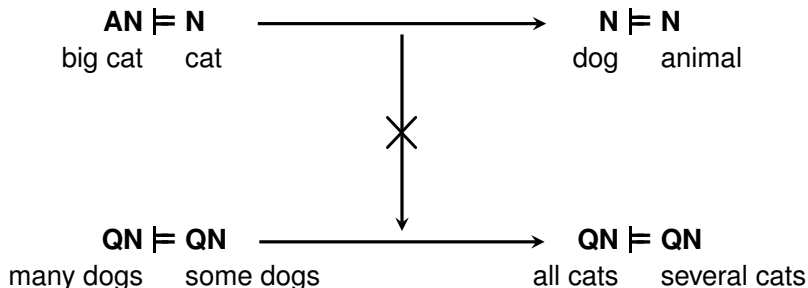
Quantifier	Instances		Correct		
	⊨	⊭	⊨	⊭	
each	656	656	649	637	(98%)
every	460	1322	402	1293	(95%)
much	248	0	216	0	(87%)
all	2949	2641	2011	2494	(81%)
several	1731	1509	1302	1267	(79%)
many	3341	4163	2349	3443	(77%)
few	0	461	0	311	(67%)
most	928	832	549	511	(60%)
some	4062	3145	1780	2190	(55%)
no	0	714	0	380	(53%)
both	636	1404	589	303	(44%)
either	63	63	2	41	(34%)
<i>Total</i>	<i>15074</i>	<i>16910</i>	<i>9849</i>	<i>12870</i>	<i>(71%)</i>

Interim summary

Entailment among composite phrases rather than nouns.
(Cheap training data!)

Entailment among logical words rather than content words.
(Part of Recognizing Textual Entailment?)

Different entailment relations at different semantic types.
(Prediction from formal semantics.)

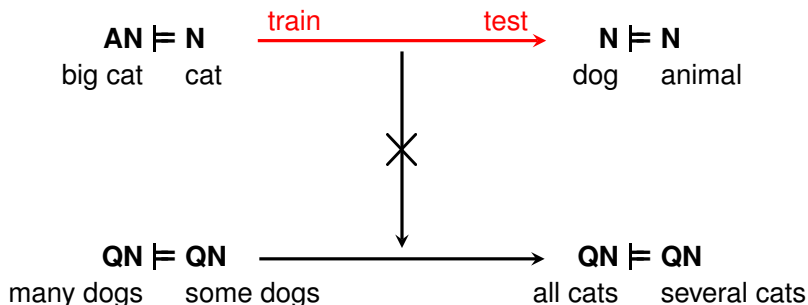


Interim summary

Entailment among **composite phrases** rather than nouns.
(Cheap training data!)

Entailment among logical words rather than content words.
(Part of Recognizing Textual Entailment?)

Different entailment relations at different semantic types.
(Prediction from formal semantics.)

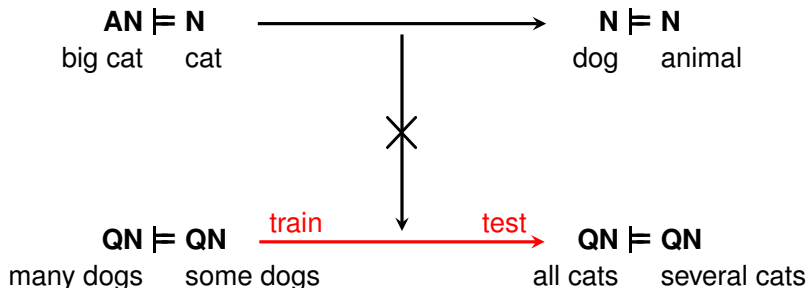


Interim summary

Entailment among **composite phrases** rather than nouns.
(Cheap training data!)

Entailment among **logical words** rather than content words.
(Part of Recognizing Textual Entailment?)

Different entailment relations at different semantic types.
(Prediction from formal semantics.)

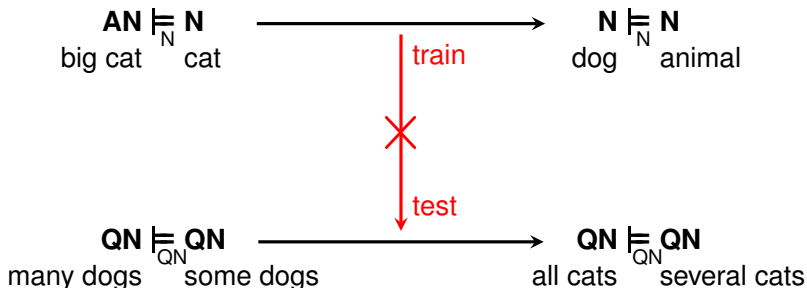


Interim summary

Entailment among **composite phrases** rather than nouns.
(Cheap training data!)

Entailment among **logical words** rather than content words.
(Part of Recognizing Textual Entailment?)

Different entailment relations at different **semantic types**.
(Prediction from formal semantics.)



Interim summary

Entailment among **composite phrases** rather than nouns.

(Cheap training data!)

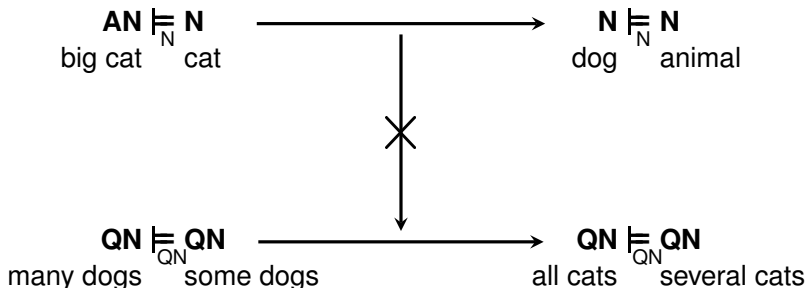
👉 Practical import

Entailment among **logical words** rather than content words.

(Part of Recognizing Textual Entailment?) 👉 Practical import

Different entailment relations at different **semantic types**.

(Prediction from formal semantics.)



Interim summary

Entailment among **composite phrases** rather than nouns.

(Cheap training data!)

👉 Practical import

Entailment among **logical words** rather than content words.

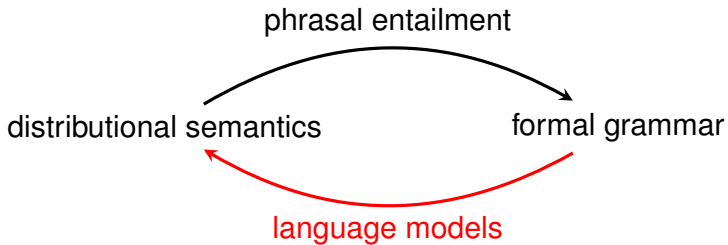
(Part of Recognizing Textual Entailment?) 👉 Practical import

Different entailment relations at different **semantic types**.

(Prediction from formal semantics.)

Ongoing work:

- ▶ How does the SVM work?
- ▶ Missing experiments?
- ▶ How to compose semantic vectors?



From language models to distributional semantics

A **language model** is a virtual infinite corpus:
not frequencies observed but probabilities estimated.

Let the distributional meaning of a phrase w be the probability distribution over its contexts c .

$$\llbracket w \rrbracket = \lambda c. \frac{\Pr(c[w])}{\sum_{c'} \Pr(c'[w])}$$

$$\llbracket \text{red army} \rrbracket = \lambda(l, r). \frac{\Pr(l \text{ red army } r)}{\sum_{(l', r')} \Pr(l' \text{ red army } r')}$$

$$\llbracket \text{red } w \rrbracket = \lambda(l, r). \frac{\llbracket w \rrbracket(l \text{ red}, r)}{\sum_{(l', r')} \llbracket w \rrbracket(l' \text{ red}, r')}$$

Probabilities from any model: bag of words, Markov, PCFG. . .
Pass the buck. Language models and corpora can (should?)
include world reference in utterance context.

From language models to distributional semantics

A language model is a virtual infinite corpus:
not frequencies observed but probabilities estimated.

Let the distributional meaning of a phrase w be the probability distribution over its contexts c .

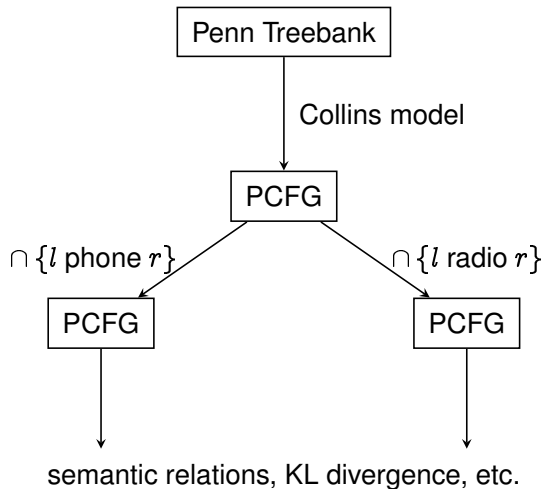
$$\llbracket w \rrbracket = \lambda c. \frac{\Pr(c[w])}{\sum_{c'} \Pr(c'[w])}$$

$$\llbracket \text{red army} \rrbracket = \lambda(l, r). \frac{\Pr(l \text{ red army } r)}{\sum_{(l', r')} \Pr(l' \text{ red army } r')}$$

$$\llbracket \text{red } w \rrbracket = \lambda(l, r). \frac{\llbracket w \rrbracket(l \text{ red}, r)}{\sum_{(l', r')} \llbracket w \rrbracket(l' \text{ red}, r')}$$

Probabilities from any model: **bag of words**, Markov, **PCFG**...
Pass the buck. Language models and corpora can (should?)
include world reference in utterance context.

From Penn Treebank to distributional semantics



Intersection grammars reveal meaning?

Random sentences

[[[[[/, he/PRP] -LCB-/-LRB-] [the/DT company/NN]] [[[[the/DT +unknown+/NNP] +unknown+/NNP] +unknown+/NNP] compared/VBN]] [says/VBZ [:/: :/:]]]

[have/VBP [lately/RB [[in/IN [July/NNP [[[[weighted/JJ large/JJ] exchange/NN] for/IN] clients/NNS]]] been/VBN]]]

was/VBD

[[/, [Mr./NNP Bush/NNP]] said/VBD]

[plunged/VBD [4/CD ,/,]]

[[start/NN of/IN] [was/VBD [me/PRP [1.9/CD pence/NN]]]]]

has/VBZ

[[/, they/PRP] [see/VBP [aided/VBN [[in/IN [[some/DT seconds/NNS] [[of/IN [executive/NN [[a/DT share/NN] [yesterday/NN [[the/DT last/JJ] market/NN]] acted/VBD]] [[the/DT advance/NN] revenue/NN]]]]] [a/DT share/NN]]]]] [by/IN [[the/DT pound/NN] and/CC] Exchange/NNP]]]]]]]

would/MD

[[Dec./NNP 28/CD] [announced/VBD that/IN]]]

[[[[/, ,/,] ,/,] [is/VBZ [[[[[the/DT only/JJ] third-quarter/JJ] net/JJ] emphasis/NN] [[very/RB executive/JJ] [[not/RB Soviet/JJ] to/TO]]]]]]] says/VBZ [[President/NNP Co./NNP] [Lee/NNP [chamber/NN supervisor/NN]]]]]

Intersection grammars reveal meaning?

Random sentences containing “car dealer”

[[[many/JJ [[the/DT car/NN] dealer/NN] Developers/NNS] ,/,] are/VBP]

[[the/DT characteristic/NN] [[that/WDT agreed/VBD] in/IN]] [is/VBZ [a/DT new/JJ federal/JJ car/NN] dealer/NN] of/IN [/: [in/IN [[their/PRP\$ modern/JJ] movie/NN] [likely/JJ ,/,]]]]]]

[says/VBZ [the/DT crude/NN] oil/NN] man/NN] plans/NNS] car/NN]] [dealer/NN [of/IN [in/IN [the/DT U.S./NNP] Cambodia/NNP] Europe/NNP]]]] But/CC] was/VBD] [anxiety/NN of/IN]] [[John/NNP +unknown+/NNP] +unknown+/NNP ,/,] ,/,] [a/DT +unknown+/NNP] In/IN] [said/VBD [bonds/NNS] Yet/RB] nonperforming/VBG] will/MD]] [the/DT role/NN]] [[the/DT executive/NN] 's/POS] son/NN] ,/,] [the/DT computer/NN]] ,/,] [engineering/NN [of/IN life/NN] of/IN]] that/DT] [+unknown+/NN +unknown+/NNP] [[The/DT head/NN] from/IN]] goes/VBZ]]

[resigned/VBD [to/TO [price/VB [the/DT car/NN] dealer/NN] [from/IN [Mr./NNP +unknown+/NNP] on/IN]]] [through/RP [such/PDT a/DT] offering/NN]]]] [showing/VBG [The/DT junk/NN] defense/NN] measure/NN] [bank/NN known/VBN]]]]

[[[still/RB [the/DT Gardens/NNPS] life/NN]] [initial/JJ transaction/NN]] [a/DT few/JJ] arrangement/NN]] administration/NN] [The/DT group/NN]] [[the/DT first/NN] price/NN] of/IN] ,/,] [is/VBZ [car/NN dealer/NN] [of/IN [\$/ \$ +unknown+/CD] the/DT futures/NNS]]]] [was/VBD [going/VBG [against/IN

Intersection grammars reveal meaning?

Random sentences containing “drug dealer”

[[[[[[[[[[[In/IN ,/,] ,/,] [[[[[its/PRP\$ past/JJ] five/CD] structural/JJ] +unknown+/NN]
[of/IN [Allied/NNP stock/NN]]]]] [+unknown+/JJ farmer/NN]] +unknown+/NNS] ,/,]
[[the/DT most/JJS] [[who/WP [drag/VBP [require/VBP because/IN]]] [because/IN
of/IN [[the/DT company/NN] [[a/DT year/NN] [[+unknown+/JJ cooperative/JJ]
children/NNS]]]]]]] [[this/DT +unknown+/JJ] OTC/NNP] market/NN] for/IN]]
[The/DT company/NN]] [[The/DT drug/NN] dealer/NN]] [soared/VBD [/,
[reducing/VBG [/, [/, Monday/NNP]]]]]]

[[[[[[[[[[[/, +unknown+/NNP] [[rose/VBD [[from/IN [[[[[\$/\$] [+unknown+/CD
[million/CD [million/CD [+unknown+/CD [million/CD [billion/CD 15.6/CD]]]]]]]]]
[a/DT share/NN]] ,/, [today/NN tickets/NNS]]] [[A/DT deep/JJ] series/NN] [[of/IN
[fact/NN [[last/JJ car/NN] that/WDT]]] [with/IN looks/NNS]]] [[seven/CD
cents/NNS] [a/DT share/NN]]]]]] [on/IN [[[[The/DT +unknown+/JJ] ownership/NN]
or/CC] yesterday/NN]]]]] [[[[[another/DT year/NN] price/NN] [above/IN -/:]]
[was/VBD [[against/IN him/PRP] although/IN]]] [but/CC [[[[the/DT following/JJ]
week/NN] [was/VBD [[[[[[[[[a/DT specific/JJ] +unknown+/JJ] short-term/JJ]
Treasury/NNP] [economic/JJ trade/NN]]] [[the/DT FDA/NNP] ,/,] ,/,] [marks/NNS
[[[[the/DT coming/VBG] early/JJ] next/JJ] year/NN] little/RB]]] ,/,]]] [and/CC
[and/CC [/, and/CC]]]]]]]]] ,/,] [+unknown+/NNP +unknown+/NNS]] [has/VBZ
n't/RB]] [received/VBD [[[[[a/DT serious/JJ] drag/NN] on/IN] [now/RB [[when/WRB
[[[[the/DT drug/NN] dealer/NN] pay/VB] [[[[Sun/NNP Jeep/NNP] Stoll/NNP]
[entered/VBD [MCI/NNP] +unknown+/NNP] U.S./NNP]]] hurt/VB]]]]] [[the/DT

Intersection grammars reveal meaning?

Top sentences

2e-2 [[said/VBD]]
1e-2 [[is/VBZ]]
8e-3 [[was/VBD]]
7e-3 [[are/VBP]]
6e-3 [[has/VBZ]]
:
5e-4 [[[[[he/PRP]]] [said/VBD]]]
:
4e-4 [[[[[he/PRP]]] [says/VBZ]]]
:
9e-5 [[[[[the/DT] company/NN]]] [said/VBD]]
:
7e-5 [[[[[she/PRP]]] [says/VBZ]]]
:
2e-6 [[[[[Mr./NNP] Inc./NNP]]] [said/VBD]]
2e-6 [[is/VBZ [[first/JJ]]]]
2e-6 [more/RBR]
2e-6 [[has/VBZ [[failed/VBN]]]]

Intersection grammars reveal meaning?

Top sentences containing “car dealer”

2e-10 [car/NN] dealer/NN]] [said/VBD]]
8e-11 [a/DT] car/NN] dealer/NN]] [said/VBD]]
7e-11 [the/DT] car/NN] dealer/NN]] [said/VBD]]
5e-11 [the/DT] car/NN] dealer/NN]] [said/VBD]]
4e-11 [car/NN] dealer/NN]]
3e-11 [car/NN] dealer/NN]] [said/VBD]]
3e-11 [says/VBZ] [car/NN] dealer/NN]]
3e-11 [car/NN] dealer/NN]] ,/] [said/VBD]]
3e-11 [,/] [car/NN] dealer/NN]] [said/VBD]]
2e-11 [a/DT] car/NN] dealer/NN]] [said/VBD]]
2e-11 [car/NN] dealer/NN]] [was/VBD]]
2e-11 [a/DT] car/NN] dealer/NN]]
2e-11 [the/DT] car/NN] dealer/NN]]
2e-11 [a/DT] car/NN] dealer/NN]] [said/VBD]]
2e-11 [the/DT] car/NN] dealer/NN]] [said/VBD]]
2e-11 [says/VBZ] [a/DT] car/NN] dealer/NN]]
1e-11 [says/VBZ] [the/DT] car/NN] dealer/NN]]
1e-11 [car/NN] dealer/NN]] [says/VBZ]]
1e-11 [the/DT] car/NN] dealer/NN]]
1e-11 [a/DT] car/NN] dealer/NN]] ,/] [said/VBD]]
1e-11 [,/] [a/DT] car/NN] dealer/NN]] [said/VBD]]

Intersection grammars reveal meaning?

Top sentences containing “drug dealer”

6e-9 [[[[[drug/NN] dealer/NN]]] [said/VBD]]
6e-9 [[[[[the/DT] drug/NN] dealer/NN]]] [said/VBD]]
2e-9 [[[[[the/DT] drug/NN] dealer/NN]]] [said/VBD]]
2e-9 [[drug/NN] dealer/NN]]
2e-9 [[[[the/DT] drug/NN] dealer/NN]]
1e-9 [[[[[The/DT] drug/NN] dealer/NN]]] [said/VBD]]
1e-9 [[says/VBZ] [[[[drug/NN] dealer/NN]]]]
1e-9 [[says/VBZ] [[[[[the/DT] drug/NN] dealer/NN]]]]
1e-9 [[[[[a/DT] drug/NN] dealer/NN]]] [said/VBD]]
1e-9 [[[[[drug/NN] dealer/NN]]] ,/,] [said/VBD]]
1e-9 [[[,/] [[[[drug/NN] dealer/NN]]] [said/VBD]]
1e-9 [[[[[the/DT] drug/NN] dealer/NN]]] ,/,] [said/VBD]]
1e-9 [[[,/] [[[[[the/DT] drug/NN] dealer/NN]]] [said/VBD]]
8e-10 [[[[[drug/NN] dealer/NN]]] [was/VBD]]
8e-10 [[[[[the/DT] drug/NN] dealer/NN]]] [was/VBD]]
7e-10 [[[[[a/DT] drug/NN] dealer/NN]]] [said/VBD]]
5e-10 [[[[the/DT] drug/NN] dealer/NN]]
5e-10 [[[[[drug/NN] dealer/NN]]] [says/VBZ]]
5e-10 [[[[[the/DT] drug/NN] dealer/NN]]] [says/VBZ]]
4e-10 [[[[[drug/NN] dealer/NN]]] [said/VBD]]
4e-10 [[[[[drug/NN] dealer/NN]]] [noted/VBD]]

Intersection grammars reveal meaning?

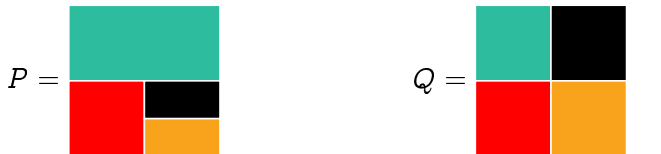
Top sentences containing “card dealer”

2e-11 [the/DT card/NN dealer/NN] [said/VBD]]
1e-11 [card/NN dealer/NN] [said/VBD]]
8e-12 [a/DT card/NN dealer/NN] [said/VBD]]
6e-12 [the/DT card/NN dealer/NN] [said/VBD]]
4e-12 [the/DT card/NN dealer/NN]]
4e-12 [a/DT card/NN dealer/NN] [said/VBD]]
3e-12 [card/NN dealer/NN]]
3e-12 [says/VBZ [the/DT card/NN dealer/NN]]]
3e-12 [The/DT card/NN dealer/NN] [said/VBD]]
3e-12 [the/DT card/NN dealer/NN] ,/ , [said/VBD]]
3e-12 [,/ , [the/DT card/NN dealer/NN] [said/VBD]]
3e-12 [says/VBZ [card/NN dealer/NN]]]
2e-12 [the/DT card/NN dealer/NN] [said/VBD]]
2e-12 [a/DT card/NN dealer/NN]]
2e-12 [The/DT card/NN dealer/NN] [said/VBD]]
2e-12 [card/NN dealer/NN] ,/ , [said/VBD]]
2e-12 [,/ , [card/NN dealer/NN] [said/VBD]]
2e-12 [the/DT card/NN dealer/NN] [was/VBD]]
2e-12 [card/NN dealer/NN] [said/VBD]]
2e-12 [the/DT card/NN dealer/NN]]
2e-12 [card/NN dealer/NN] [was/VBD]]

Kullback-Leibler divergence

$$D_{\text{KL}}(P \parallel Q) = \overbrace{\sum_x P(x) \log \frac{1}{Q(x)}}^{\text{cross entropy}} - \overbrace{\sum_x P(x) \log \frac{1}{P(x)}}^{\text{entropy}}$$

Example



20 samples from P : ●●●●●●●●●●●●●●●●●●

Kullback-Leibler divergence

$$D_{\text{KL}}(P \parallel Q) = \overbrace{\sum_x P(x) \log \frac{1}{Q(x)}}^{\text{cross entropy}} - \overbrace{\sum_x P(x) \log \frac{1}{P(x)}}^{\text{entropy}}$$

Example



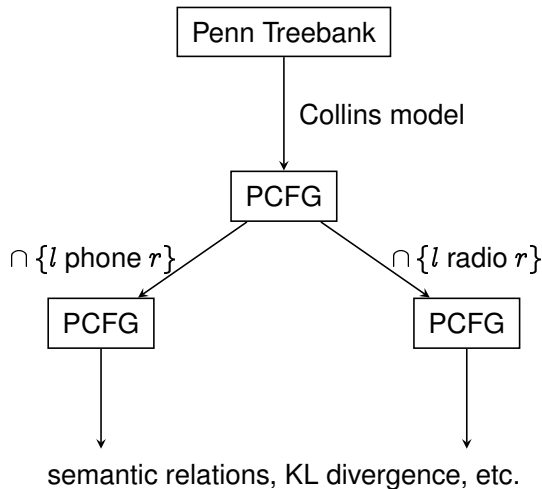
20 samples from P : ●●●●●●●●●●●●●●●●●●

encoded for P : 100010011010100110100101110110000111

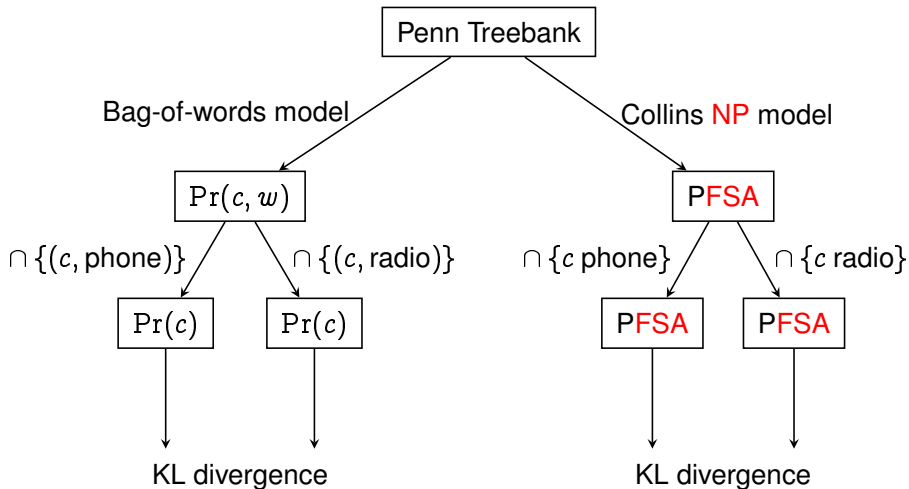
encoded for Q : 1000001000011010000110001011000100000011

KL divergence: 0.25 bits = 2.00 bits – 1.75bits

From Penn Treebank to distributional semantics



From Penn Treebank to distributional semantics

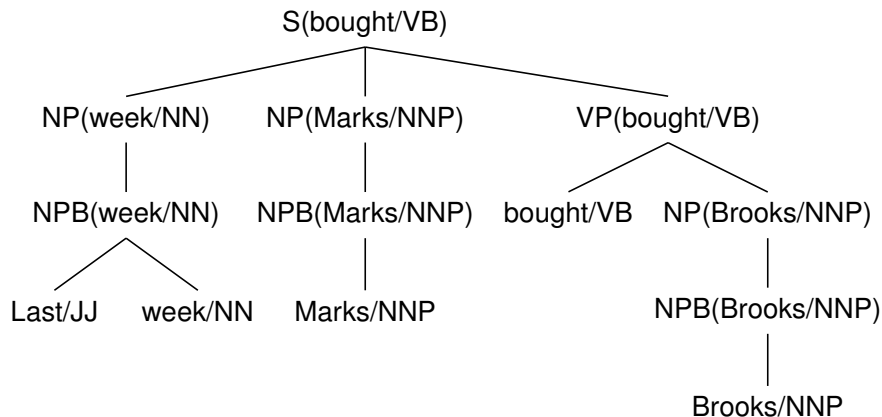


Collins model

Lexicalized PCFG for parsing (1997)

Not for generation (Post & Gildea 2008)

Bikel (2004) exegesis

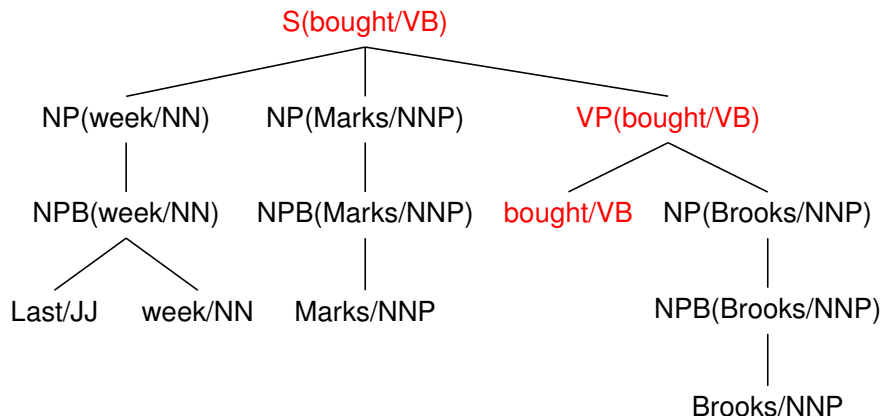


Collins model

Lexicalized PCFG for parsing (1997)

Not for generation (Post & Gildea 2008)

Bikel (2004) exegesis

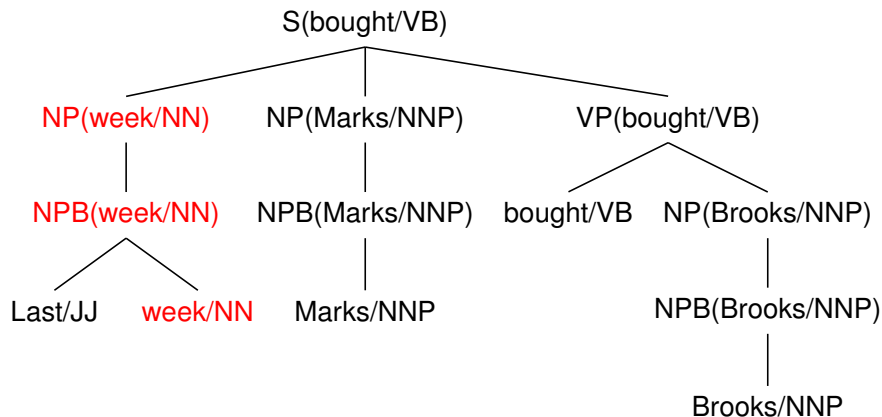


Collins model

Lexicalized PCFG for parsing (1997)

Not for generation (Post & Gildea 2008)

Bikel (2004) exegesis



Summary statistics

Standard English training set: Wall Street Journal §§02–21

- ▶ 39 832 sentences
- ▶ 950 028 word tokens
 - 44 113 unique words
 - 10 437 unique words that occur 6+ times
- ▶ 28 basic nonterminal labels
 - 42 parts of speech

Tiny for a corpus today.

Simplified Collins Model 1

- ▶ 575 936 nonterminals
 - 15 564 terminals
 - 12 611 676 rules

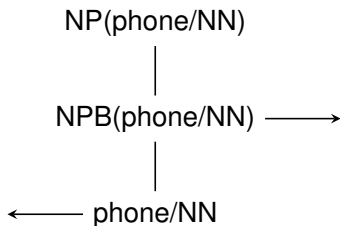
Big for a grammar today.

Pilot evaluation using BLESS data set

Concept	Relation	Relatum
phone	coord	computer
phone	coord	radio
phone	coord	stereo
phone	coord	television
phone	hyper	commodity
phone	hyper	device
phone	hyper	equipment
phone	hyper	good
phone	hyper	object
phone	hyper	system
phone	mero	cable
phone	mero	dial
phone	mero	number
phone	mero	plastic
phone	mero	wire
phone	random-n	choice
phone	random-n	clearance
phone	random-n	closing
phone	random-n	entrepreneur

Baroni and Lenci Evaluation
of Semantic Spaces (2011)

Only head nouns observed
in corpus:



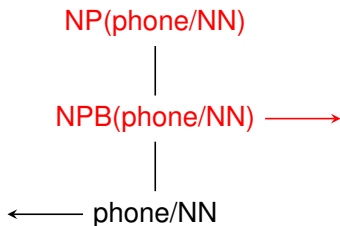
Compute KL divergences
among distributions over
modifier-nonterminal
sequences

Pilot evaluation using BLESS data set

Concept	Relation	Relatum
phone	coord	computer
phone	coord	radio
phone	coord	stereo
phone	coord	television
phone	hyper	commodity
phone	hyper	device
phone	hyper	equipment
phone	hyper	good
phone	hyper	object
phone	hyper	system
phone	mero	cable
phone	mero	dial
phone	mero	number
phone	mero	plastic
phone	mero	wire
phone	random-n	choice
phone	random-n	clearance
phone	random-n	closing
phone	random-n	entrepreneur

Baroni and Lenci Evaluation
of Semantic Spaces (2011)

Only head nouns observed
in corpus:



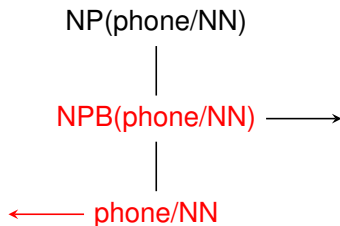
Compute KL divergences
among distributions over
modifier-nonterminal
sequences

Pilot evaluation using BLESS data set

Concept	Relation	Relatum
phone	coord	computer
phone	coord	radio
phone	coord	stereo
phone	coord	television
phone	hyper	commodity
phone	hyper	device
phone	hyper	equipment
phone	hyper	good
phone	hyper	object
phone	hyper	system
phone	mero	cable
phone	mero	dial
phone	mero	number
phone	mero	plastic
phone	mero	wire
phone	random-n	choice
phone	random-n	clearance
phone	random-n	closing
phone	random-n	entrepreneur

Baroni and Lenci Evaluation of Semantic Spaces (2011)

Only head nouns observed in corpus:



Compute KL divergences among distributions over *modifier-nonterminal sequences*

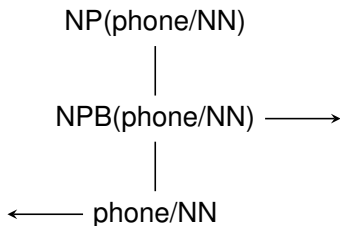
Pilot evaluation using BLESS data set

38 Concept Relation **687 Relatum**

phone	173 coord	computer
phone	coord	radio
phone	coord	stereo
phone	coord	television
phone	125 hyper	commodity
phone	hyper	device
phone	hyper	equipment
phone	hyper	good
phone	hyper	object
phone	hyper	system
phone	490 mero	cable
phone	mero	dial
phone	mero	number
phone	mero	plastic
phone	mero	wire
phone	561 random-n	choice
phone	random-n	clearance
phone	random-n	closing
phone	random-n	entrepreneur

Baroni and Lenci Evaluation of Semantic Spaces (2011)

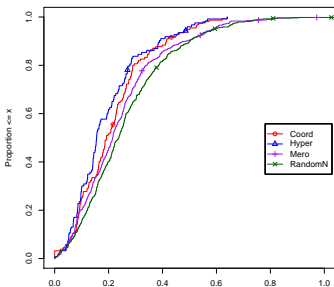
Only head nouns observed in corpus:



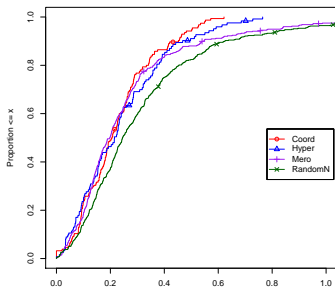
Compute KL divergences among distributions over *modifier-nonterminal sequences*

$D_{KL}(\text{Concept} \parallel \text{Relatum})$ $D_{KL}(\text{Relatum} \parallel \text{Concept})$

NP
 |
 NPB →

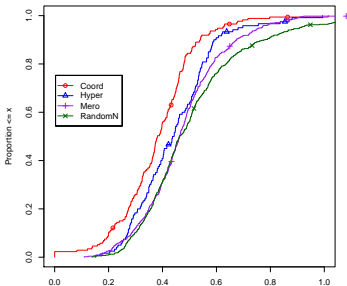


$r_{np_npb_KL}$ by Relation (Kruskal-Wallis rank sum test $p=1.73002e-05$)
 n:1288 m:61

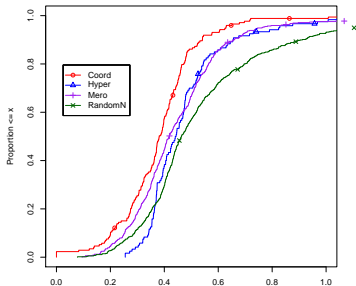


$r_{np_npb_LK}$ by Relation (Kruskal-Wallis rank sum test $p=5.88196e-06$)
 n:1288 m:61

NPB
 |
 ← NNS



l_{npb_KL} by Relation (Kruskal-Wallis rank sum test $p=2.54141e-13$)
 n:1340 m:9

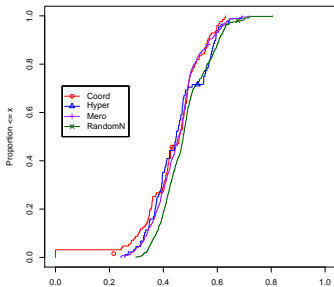


l_{npb_LK} by Relation (Kruskal-Wallis rank sum test $p=1.0453e-15$)
 n:1340 m:9

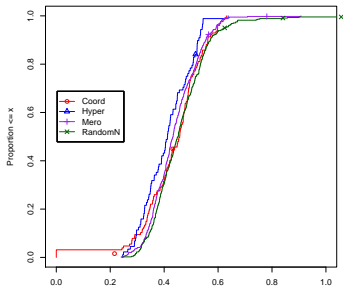
$D_{KL}(\text{Concept} \parallel \text{Relatum})$

$D_{KL}(\text{Relatum} \parallel \text{Concept})$

Bag of words



BagOfWords_KL by Relation (Kruskal-Wallis rank sum test $p=9.20412e-05$)
n:1048 m:0

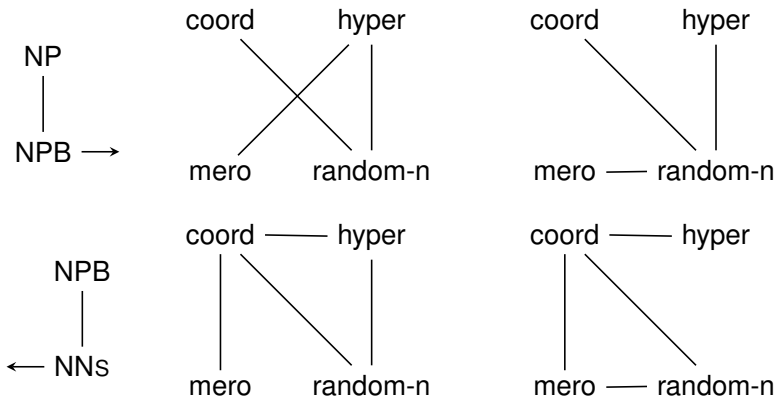


BagOfWords_LK by Relation (Kruskal-Wallis rank sum test $p=0.000968842$)
n:1048 m:0

Mann-Whitney-Wilcoxon rank sum test

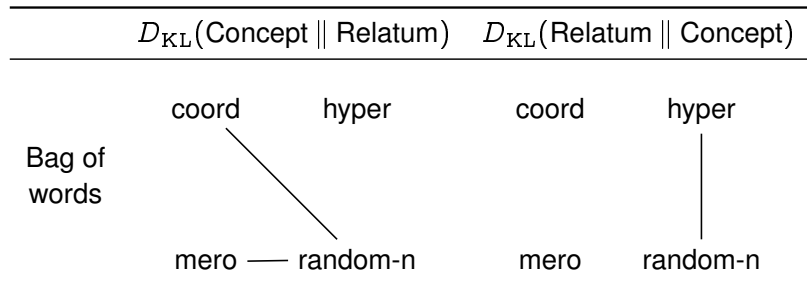
Edges indicate $p < .01$

$D_{KL}(\text{Concept} \parallel \text{Relatum})$ $D_{KL}(\text{Relatum} \parallel \text{Concept})$



Mann-Whitney-Wilcoxon rank sum test

Edges indicate $p < .01$



Summary

Distributional semantics from language models

- ▶ Estimate felicity *in context* from observed use
- ▶ Cope with sparse data using linguistic insight such as syntax
- ▶ Better distributional semantics from better language models?

Phrasal entailment from distributional semantics

- ▶ Logical words
- ▶ Semantic types

Thanks!

- ▶ Bolzano: Ngoc-Quynh Do, European Masters Program in Language and Communication Technologies
- ▶ Trento: Marco Baroni, Raffaella Bernardi, Roberto Zamparelli
- ▶ Rutgers: Jason Perry, Matthew Stone
- ▶ Cornell: John Hale, Mats Rooth

Holding out each quantifier pair

Quantifier pair	Instances	Correct	Quantifier pair	Instances	Correct
all = some	1054	1044 (99%)	some ≠ every	484	481 (99%)
all = several	557	550 (99%)	several ≠ all	557	553 (99%)
each = some	656	647 (99%)	several ≠ every	378	375 (99%)
all = many	873	772 (88%)	some ≠ all	1054	1043 (99%)
much = some	248	217 (88%)	many ≠ every	460	452 (98%)
every = many	460	400 (87%)	some ≠ each	656	640 (98%)
many = some	951	822 (86%)	few ≠ all	157	153 (97%)
all = most	465	393 (85%)	many ≠ all	873	843 (97%)
several = some	580	439 (76%)	both ≠ most	369	347 (94%)
both = some	573	322 (56%)	several ≠ few	143	134 (94%)
many = several	594	113 (19%)	both ≠ many	541	397 (73%)
most = many	463	84 (18%)	many ≠ most	463	300 (65%)
both = either	63	1 (2%)	either ≠ both	63	39 (62%)
			many ≠ no	714	369 (52%)
			some ≠ many	951	468 (49%)
			few ≠ many	161	33 (20%)
			both ≠ several	431	63 (15%)