A Little Bit of NLP Goes A Long Way: Adding Phrases to the Term-Document Matrix using Finite-State Shallow Parsing

https://github.com/slanglab/phrasemachine
http://slanglab.cs.umass.edu/phrasemachine/

Abram Handler (UMass Amherst), Matthew J. Denny (Penn State),
Hanna Wallach (MSR), and Brendan O'Connor (UMass Amherst)

New Directions in Text As Data, October 15, 2016
Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

103d CONGRESS
1st Session

H. R. 3

[Report No. 103-375, Part I]

To amend the Federal Election Campaign Act of 1971 to provide for a voluntary system of spending limits and benefits for congressional election campaigns, and for other purposes.

Raw Text

103d CONGRESS
1st Session

H. R. 3

[Report No. 103-375, Part I]

To amend the Federal Election Campaign Act of 1971 to provide for a voluntary system of spending limits and benefits for congressional election campaigns, and for other purposes.

Raw Text

Term extraction

Unigrams

Word	Count
candidate	215
section	158
Federal	154
election	140
committee	120
under	115
that	114
...	...

```
103d CONGRESS
  1st Session

                                H. R. 3

                    [Report No. 103-375, Part I]

  To amend the Federal Election Campaign Act of 1971 to provide for a
  voluntary system of spending limits and benefits for congressional
               election campaigns, and for other purposes.
```

Raw Text

Term extraction

Unigrams

Word	Count
candidate	215
section	158
Federal	154
election	140
committee	120
under	115
that	114
...	...

Downstream analysis

Topic models!

Term-covariate ranking!

Supervised learning!

103d CONGRESS
1st Session

H. R. 3

[Report No. 103-375, Part I]

To amend the Federal Election Campaign Act of 1971 to provide for a voluntary system of spending limits and benefits for congressional election campaigns, and for other purposes.

Raw Text

Term extraction

Unigrams

Word	Count
candidate	215
section	158
Federal	154
election	140
committee	120
under	115
that	114
...	...

Phrase	Count
Federal Election	89
political party	34
ballot initiative	24
eligible candidate	14
political committees	14
voter communication vouchers	12
...	...

Phrases

Downstream analysis

Topic models!

Term-covariate ranking!

Supervised learning!

Why phrases / multiword expressions?

Nobel Prize example: http://languagelog.ldc.upenn.edu/nll/?p=28833

- Advantages
  - More specific concepts *clinton* vs. *hillary clinton*
  - Different word senses *social* vs. *social security*

- Disadvantages
  - Sparsity
  - Overlaps

N-grams need filtering

Unit of Analysis	Meaningful Text
Sentence	{Should a Federal agency seek to restrict photography of its installations or personnel, it shall obtain a court order that outlines the national security or other reasons for the restriction. }
Tokens	{Should}, {a}, {Federal}, {agency}, {seek}, {to}, {**restrict**}, {**photography**}, {of}, {its}, { installations}, {or}, {personnel}, {it}, {shall}, {obtain}, {a}, {**court**}, {order}, {that}, {outlines}, { the}, {national}, {**security**}, {or}, {other}, {reasons}, {for}, {the}, {restriction}
Bigrams	{Should a}, {a Federal}, {Federal agency}, {agency seek}, {seek to}, {to restrict}, {**restrict photography**}, {photography of}, {of its}, {its installations}, {installations or}, {or personnel}, {personnel it}, {it shall}, {shall obtain}, {obtain a}, {a court}, {**court order**}, {order that}, {that outlines}, {outlines the}, {the national}, {**national security**}, {security or}, {or other}, {other reasons}, {reasons for}, {for the}, {the restriction}
Trigrams	{Should a Federal}, {a Federal agency}, {**Federal agency seek**}, {agency seek to}, {seek to restrict}, {**to restrict photography**}, {restrict photography of}, {photography of its}, {of its installations}, {its installations or}, {**installations or personnel**}, {or personnel it}, {personnel it shall}, {it shall obtain}, {shall obtain a}, {obtain a court}, {**a court order**}, {court order that}, {order that outlines}, {that outlines the}, {outlines the national}, {**the national security**}, {national security or}, {security or other}, {or other reasons}, {other reasons for}, {reasons for the}, {for the restriction}

N-gram filtering

1. Statistical collocations
    - e.g. Dunning 1993
2. Grammatical information
    - e.g. Justeson and Katz 1995

Should a Federal agency seek to restrict photography of its installations or personnel, it shall obtain a court order that outlines the national security or other reasons for the restriction.

POS Tags

Should/M a/D Federal/N agency/N seek/V to/T restrict/V photography/N of/P its/PR installations/N or/C personnel/N, it/PR shall/M obtain/V a/D court/N order/N that/d outlines/V the/D national/A security/N or/C other/A reasons/N for/P the/D restriction/N.

Tag Definitions

A = adjective, N = noun, V = verb, M = modal, D = determiner,
C = conjunction, PR = pronoun, T = to

Part of speech tagging has relatively high accuracy (compared to other NLP tasks...)

Demo: Part-of-speech patterns

```
Sloths          NNS
are             VBP
mammals         NNS

[...]

They            PRP
are             VBP
named           VBN
after           IN
the             DT
capital         NN
sin             NN
of              IN
sloth           NN
```

Pattern: N
dr., sloths, half, america, fingers, birth, bodies, attackers, dwellers, germany

Pattern: NN
toed sloths, central america, sloth fur, b. pygmaeus, national park, toed sloths, toed sloths, toed sloth, costa rica, bradypus torquatus

Pattern: AN
stubby tails, most sloths, metabolic adaptations, tiny ears, specialized claws, most mammals, female moths, special problems, monophyletic group, curved claws

Pattern: (A|N)*N
phylogeny, sloths, toed sloth, sloths, birth, sloth, placental mammals, types, snouts, cecropia

Pattern: (A|N)*N
institute, strong body, trunk, ecuador, shasta ground, superorder, 13 ft, 60 million years, defenselessness, ground sloths

Pattern: (A|N){3}N
small ground sloths acratocnus, study testing sloth sleep, manuel antonio national park, testing sloth sleep patterns, many megafaunal ground sloths, many other rainforest animals, fur hosts two species, huge ground sloths megalonyx

Pattern: NPN
group of mammals, visits to ground, jungles of central, island in panama, speed during emergency, range of economy, lineage of ground, deaths in costa, relationships with moths, tribes in ecuador

Noun Phrase (NP) patterns

- Noun phrases: can stand alone in a list of terms
- Recognition: compositionality of language
  - BaseNP = (Adj | Noun)* Noun
  - PP = Prep Det* BaseNP
  - NP = BaseNP PP*

Noun Phrase (NP) patterns

- Noun phrases: can stand alone in a list of terms
- Recognition: compositionality of language
  - BaseNP = (Adj | Noun)* Noun
  - PP = Prep Det* BaseNP
  - NP = BaseNP PP*

- Justeson and Katz (1995): restricted to bigrams and trigrams

Tag Pattern	**Example**
AN	*equal employment*
NN	*research project*
AAN	*local educational agency*
ANN	*recreational land resource*
NAN	*health related service*
NNN	*health care provider*
NPN	*election by majority*

Grammar!

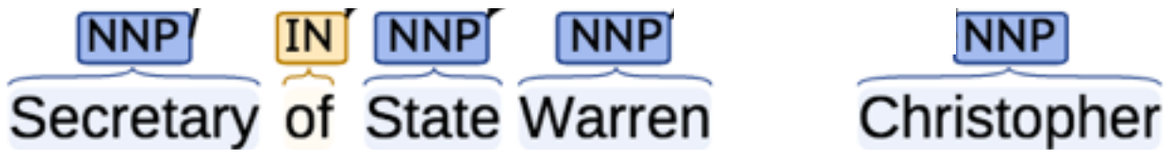
Appendix: FullNP Grammar

The following `foma` grammar defines the rewrite phrase transducer P:

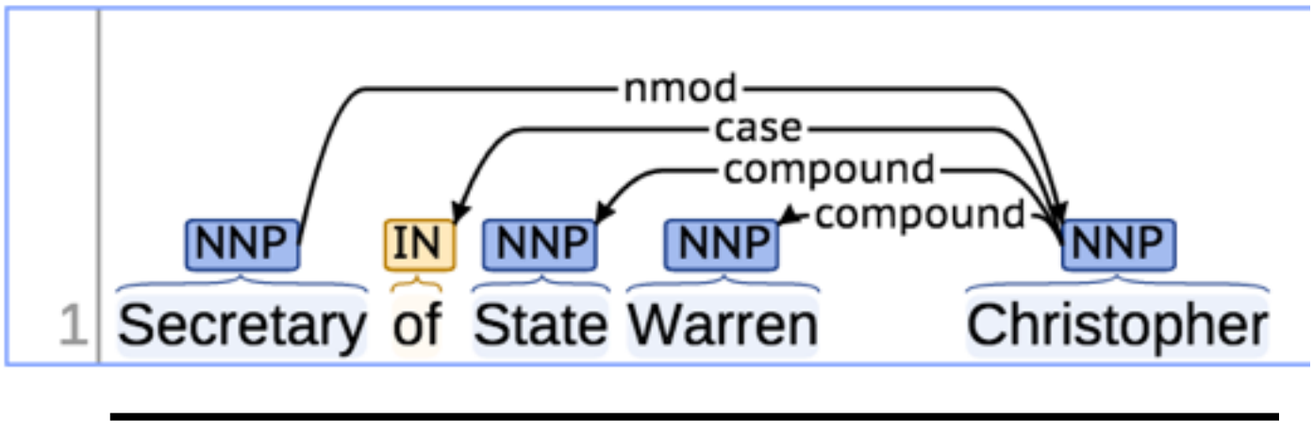
```
# POS tag categories. "Coarse" refer to the Petrov Univeral tag set.
# We directly use PTB tags, but for Twitter, we assume they've been
# preprocessed to coarse tags.
# CD is intentionally under both Adj and Noun.
define Adj1     [JJ | JJR | JJS | CD | CoarseADJ];
define Det1     [DT | CoarseDET];
define Prep
define Adv
# Note tha
define Ver
# PTB FW g
# Twitter
define Nou
define Ver
define Any
    Coarse
    Coarse
]
define Lpa
define Rpa
# Ideally,
```

```
# single-word coordinations
define Adj      Adj1 [CC Adj1]*;
define Det      Det1 [CC Det1]*;
define Adv      Adv1 [CC Adv1]*;
define Prep     Prep1 [CC Prep1]*;
define VerbMod  VerbMod1 [CC VerbMod1]*;

# NP (and thus BaseNP) have to be able to stand on their own.  They are not
# allowed to start with a determiner, since it's usually extraneous for our
# purposes. But when we want an NP right of something, we need to allow
# optional determiners since they're in between.
define BaseNP   [Adj|Noun]* Noun;
define PP       Prep+ [Det|Adj]* BaseNP;
define ParenP   Lparen AnyPOS^{1,50} Rparen;
define NP1      BaseNP [PP | ParenP]*;
define NP       NP1 [CC [Det|Adj]* NP1]*;

regex NP -> START ... END;
write att compiled_fsts/NP.attfoma
```

Full Syntax Parser?



Full Syntax Parsers Don't Work

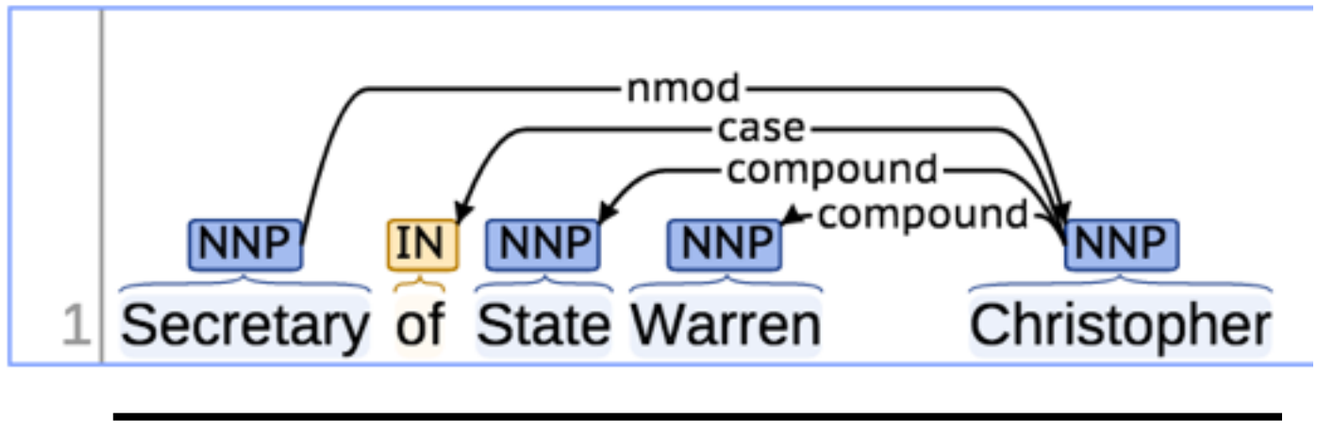


Curse from the 1990s: Penn Treebank's flat NPs

Supervised learning parsers can't escape annotation assumptions

Full Syntax Parsers Don't Work

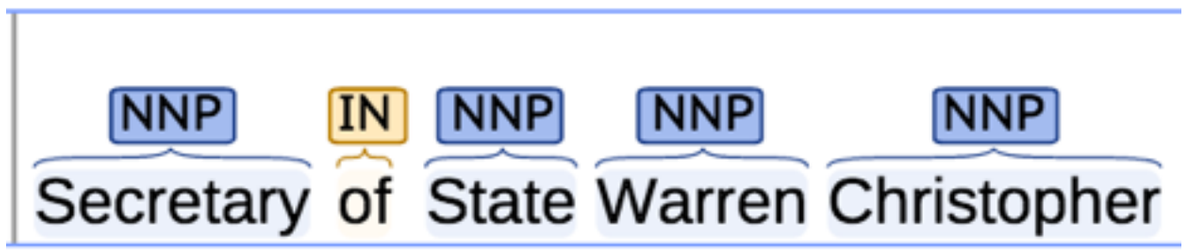
Basic Dependencies:



Curse from the 1990s:
Penn Treebank's flat NPs

Supervised learning parsers can't escape annotation assumptions

*Noun+ (Prep Noun+)**

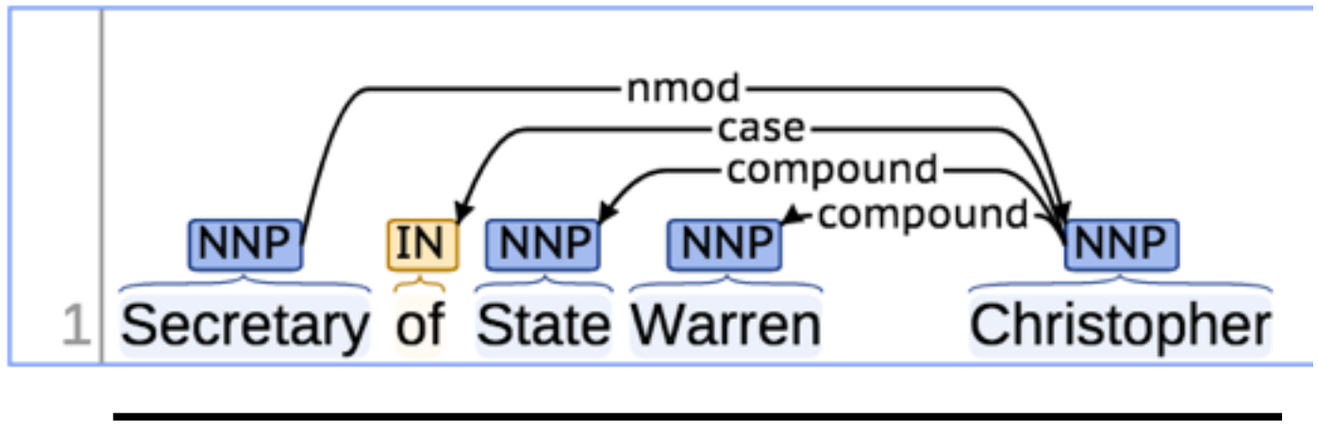


Our approach:
Specify linguistic ***rules*** over part-of-speech tags

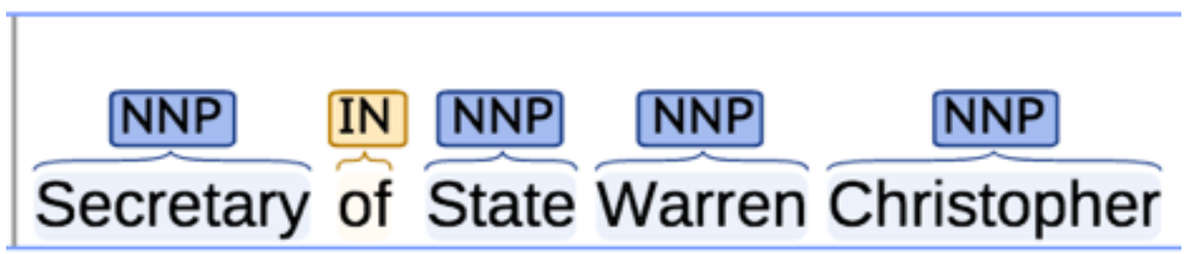
Overgenerate
(allow false positives)

Full Syntax Parsers Don't Work

Basic Dependencies:



*Noun+ (Prep Noun+)**



Curse from the 1990s:
Penn Treebank's flat NPs

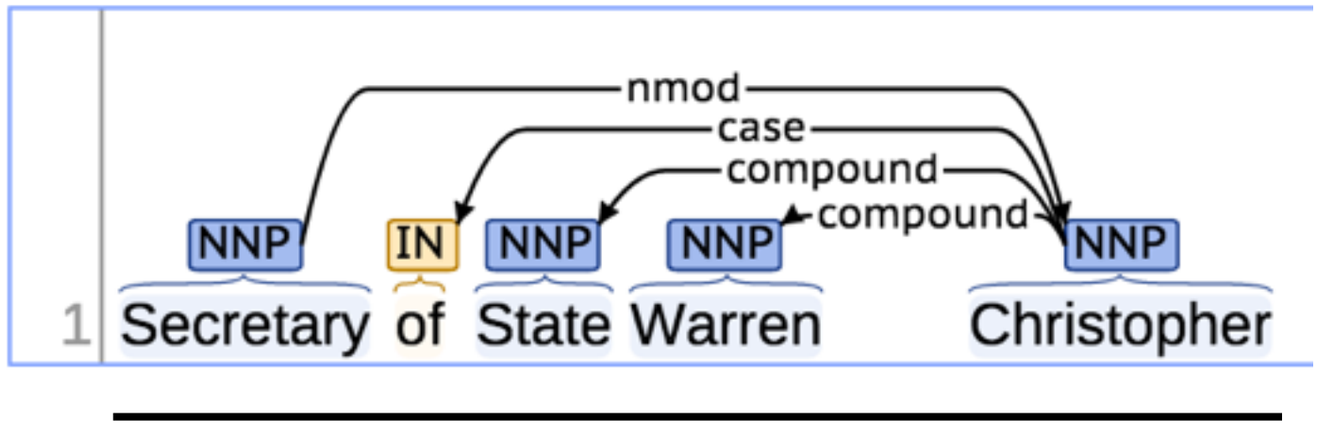
Supervised learning parsers can't escape annotation assumptions

Our approach:
Specify linguistic ***rules*** over part-of-speech tags

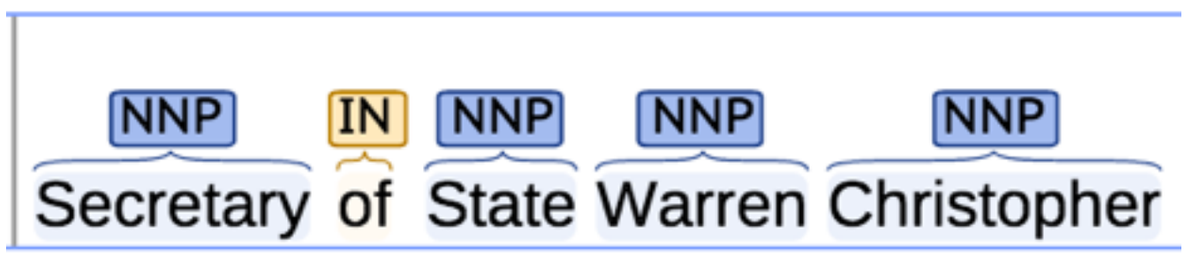
Overgenerate
(allow false positives)

Full Syntax Parsers Don't Work

Basic Dependencies:



*Noun+ (Prep Noun+)**



Curse from the 1990s:
Penn Treebank's flat NPs

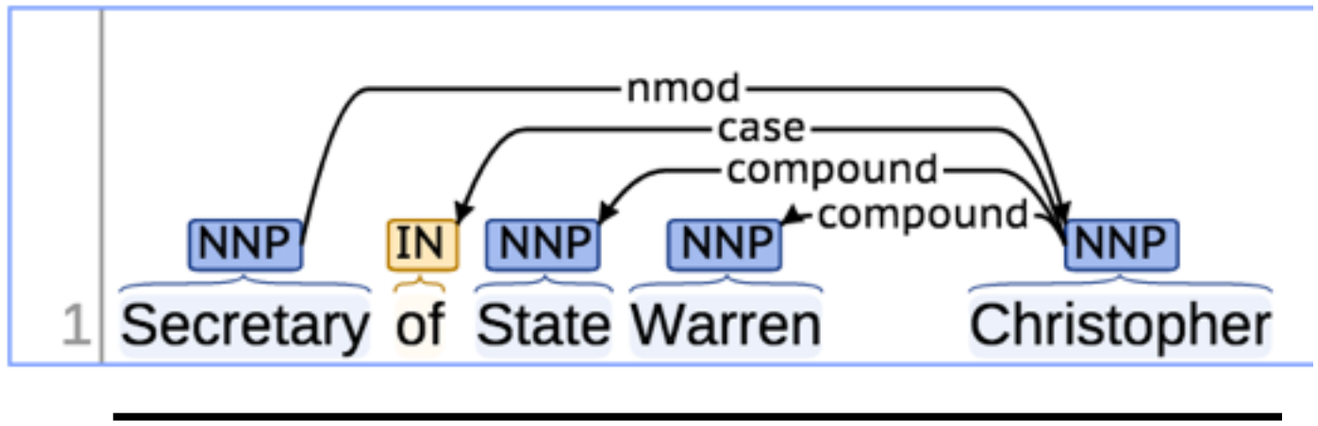
Supervised learning parsers can't escape annotation assumptions

Our approach:
Specify linguistic ***rules*** over part-of-speech tags

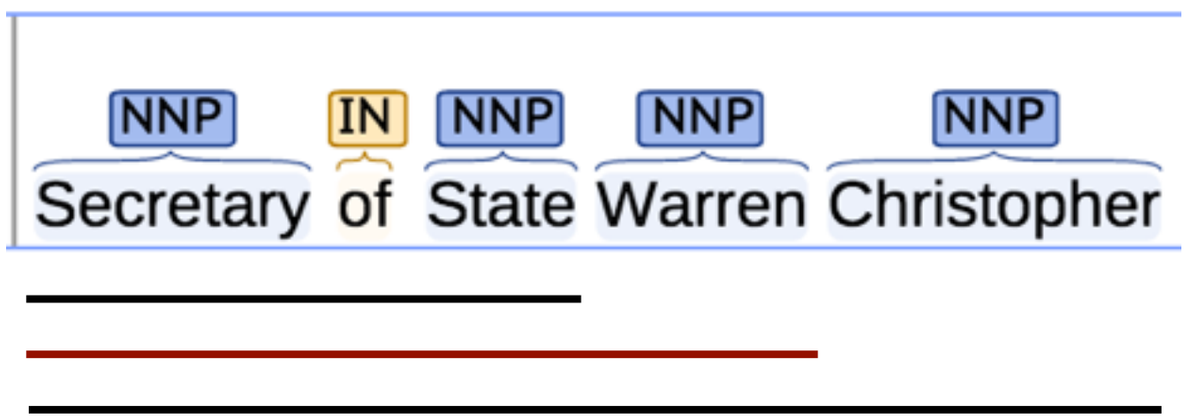
Overgenerate
(allow false positives)

Full Syntax Parsers Don't Work

Basic Dependencies:



*Noun+ (Prep Noun+)**



Curse from the 1990s:
Penn Treebank's flat NPs

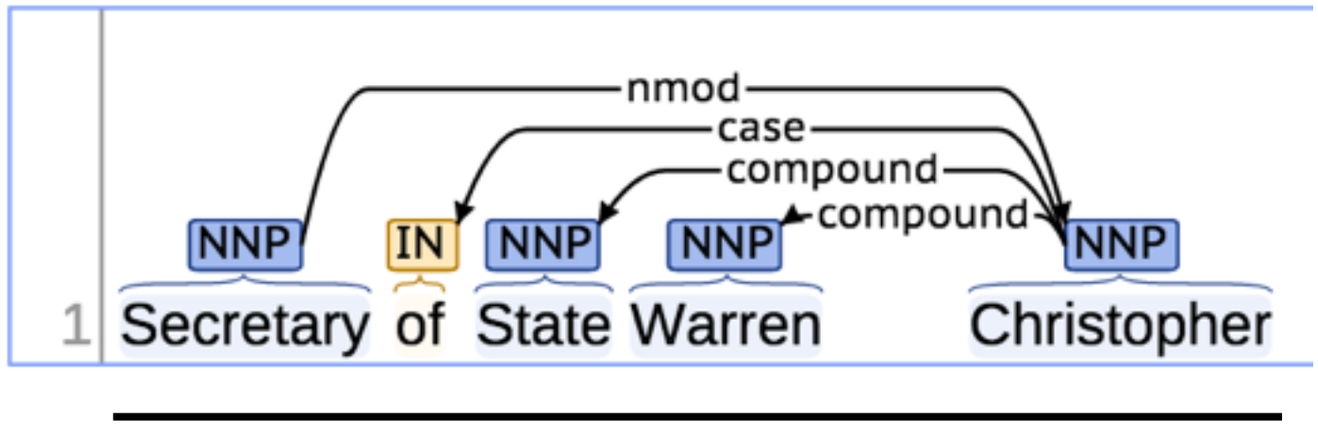
Supervised learning parsers can't escape annotation assumptions

Our approach:
Specify linguistic ***rules*** over part-of-speech tags

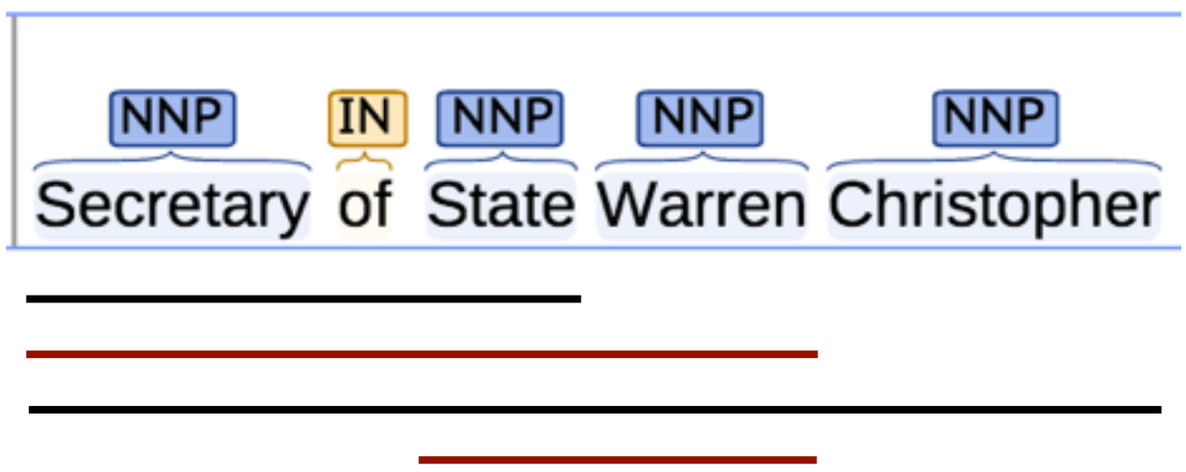
Overgenerate
(allow false positives)

Full Syntax Parsers Don't Work

Basic Dependencies:



*Noun+ (Prep Noun+)**



Curse from the 1990s:
Penn Treebank's flat NPs

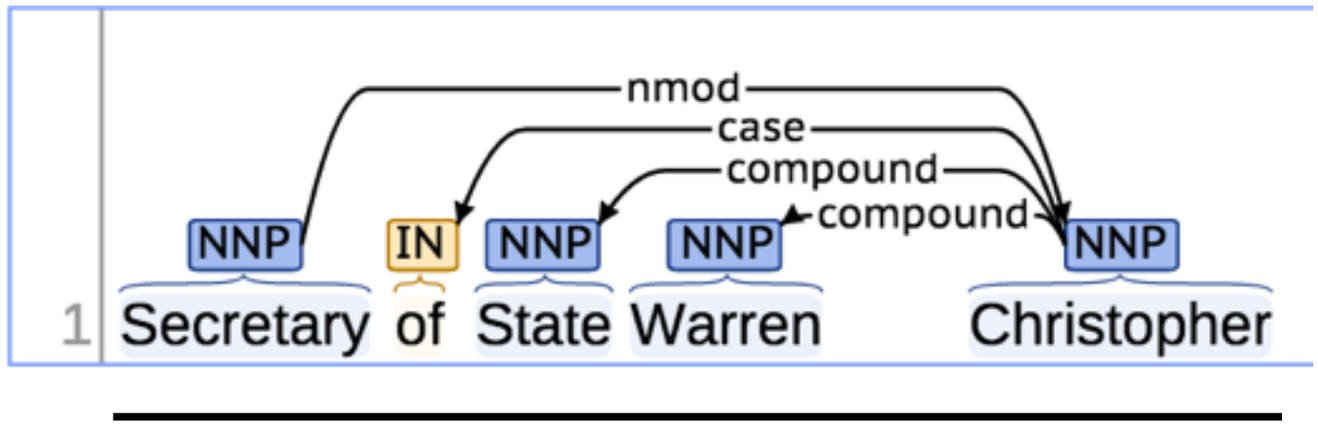
Supervised learning parsers can't escape annotation assumptions

Our approach:
Specify linguistic ***rules*** over part-of-speech tags

Overgenerate
(allow false positives)

Full Syntax Parsers Don't Work

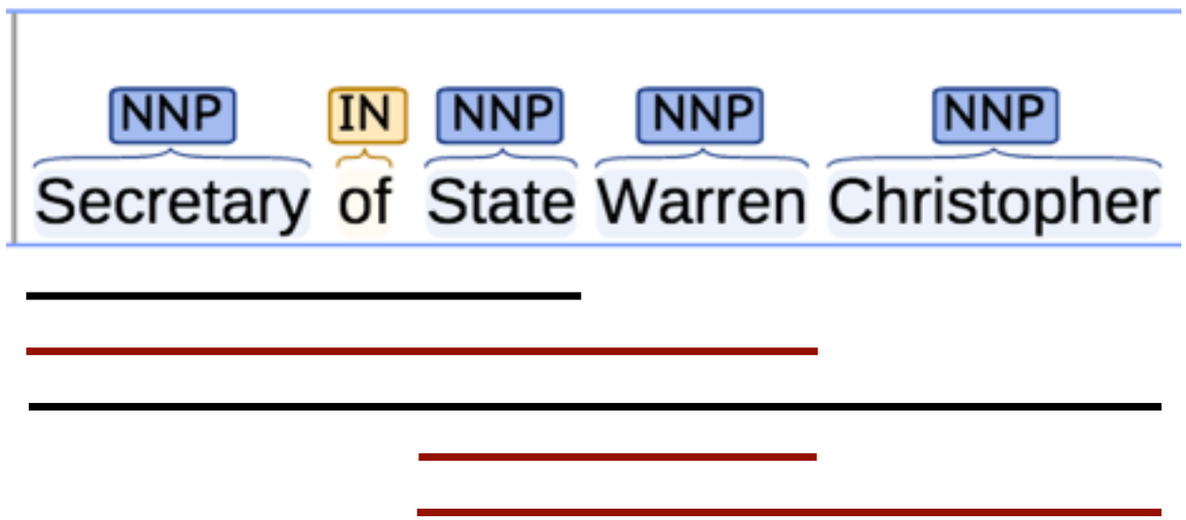
Basic Dependencies:



Curse from the 1990s:
Penn Treebank's flat NPs

Supervised learning parsers can't escape annotation assumptions

*Noun+ (Prep Noun+)**

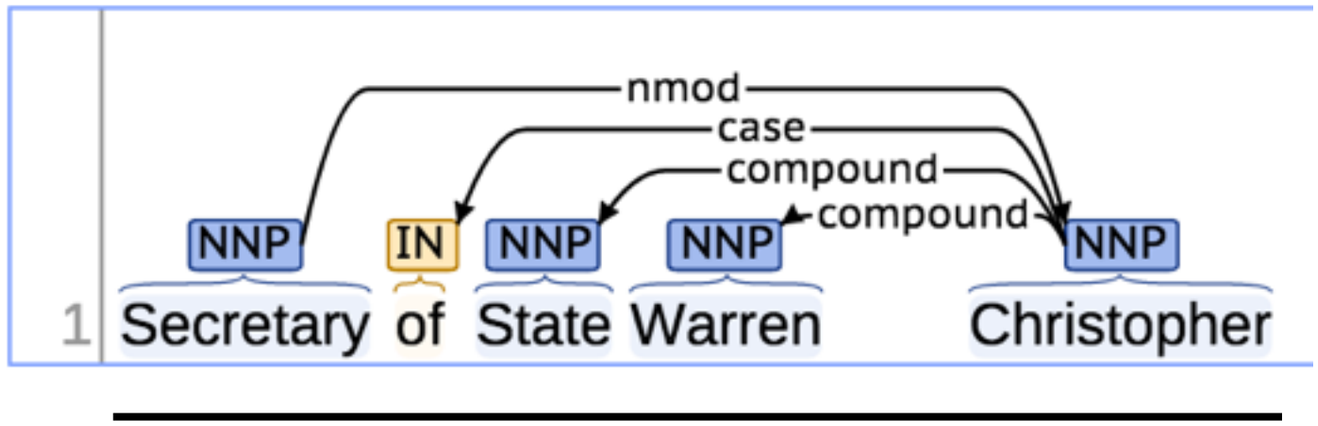


Our approach:
Specify linguistic ***rules*** over part-of-speech tags

Overgenerate
(allow false positives)

Full Syntax Parsers Don't Work

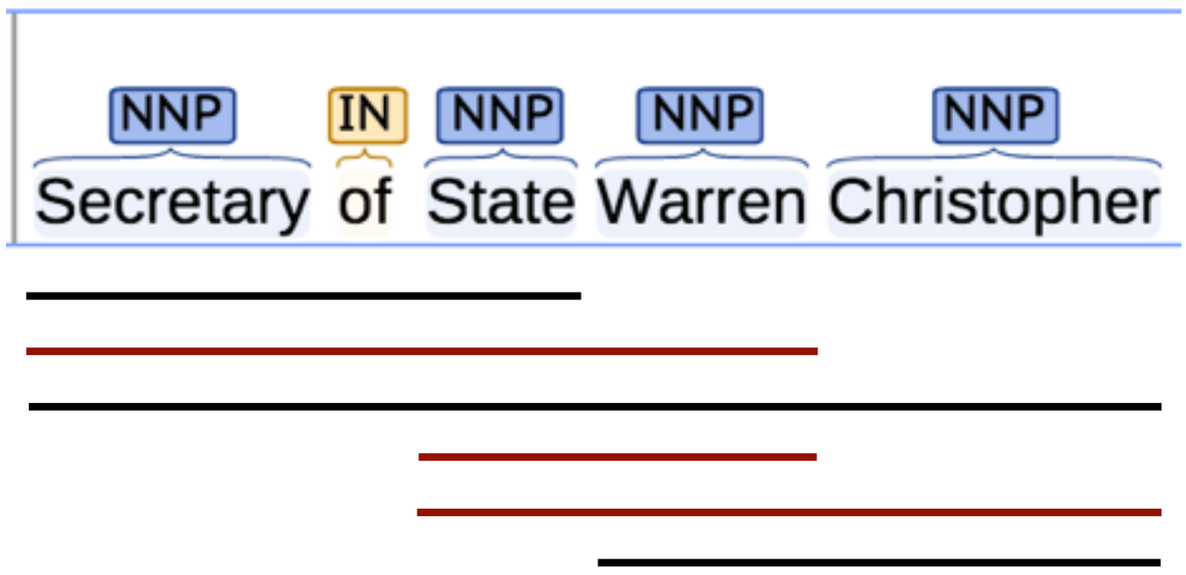
Basic Dependencies:



Curse from the 1990s:
Penn Treebank's flat NPs

Supervised learning parsers can't escape annotation assumptions

*Noun+ (Prep Noun+)**



Our approach:
Specify linguistic ***rules*** over part-of-speech tags

Overgenerate
(allow false positives)

- Procedure
  1. Run a POS tagger
  2. Extract phrases.
     Use both unigrams and phrases in doc-term matrix
  3. Run your favorite statistical text algorithm
     - Topic models! Scaling! TF-IDF! PMI! Monroe et al. z-scores!
  4. Term merging when viewing a ranked term list
     - Merging rules: string overlap or co-occurrence

- Evaluation: Traditional NER
- Interactive news exploration
- Congressional bills

Most frequent terms

Data Set	Method	Ranked List
Twitter	unigrams JK	snow, #tcot, al, dc, gore
	NPFST	
Old Bailey	unigrams ConsitParse JK	jacques, goodridge, rust, prisoner, sawtell
	NPFST	
NYT	unigrams ConstitParse JK	will, united, one, government, new
	NPFST	

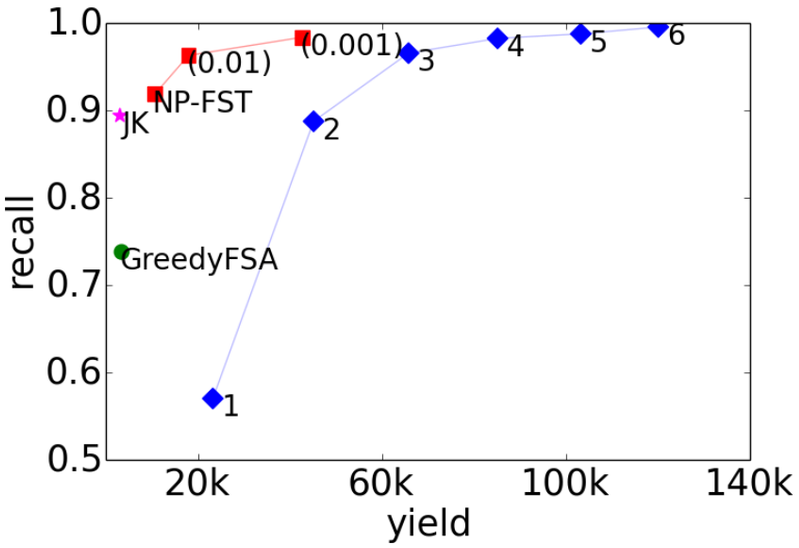
Most frequent terms

Data Set	Method	Ranked List
Twitter	unigrams	snow, #tcot, al, dc, gore
	JK	
	NPFST	
Old Bailey	unigrams	jacques, goodridge, rust, prisoner, sawtell
	ConsitParse	the prisoner, the warden, the draught, the fleet, the house
	JK	
	NPFST	
NYT	unigrams	will, united, one, government, new
	ConstitParse	the united states, the government, the agreement, the president, the white house
	JK	
	NPFST	

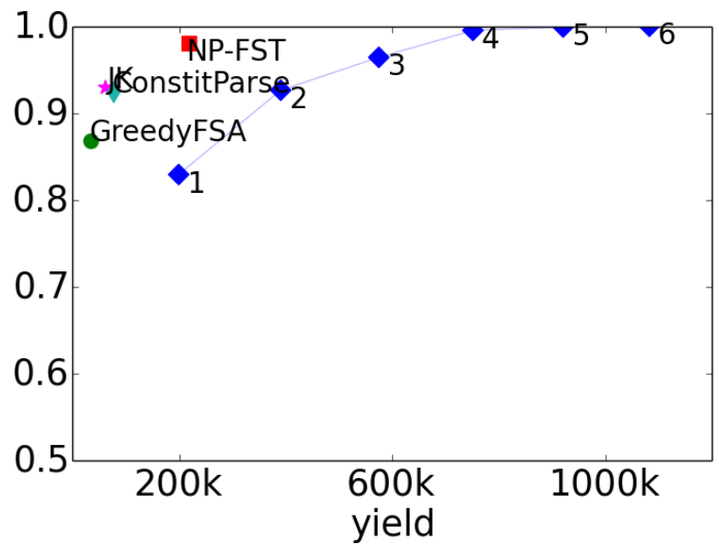
Most frequent terms

Data Set	Method	Ranked List
Twitter	unigrams JK	snow, #tcot, al, dc, gore
	NPFST	al gore's, snake oil science, 15 months, snow in dc, *bunch of snake oil science
Old Bailey	unigrams ConsitParse JK	jacques, goodridge, rust, prisoner, sawtell the prisoner, the warden, the draught, the fleet, the house
	NPFST	middlesex jury, public house, warrant of attorney, baron perryn, *middlesex jury before lord loughborough
NYT	unigrams ConstitParse JK	will, united, one, government, new the united states, the government, the agreement, the president, the white house
	NPFST	united states, united nations, white house, health care, *secretary of state warren christopher

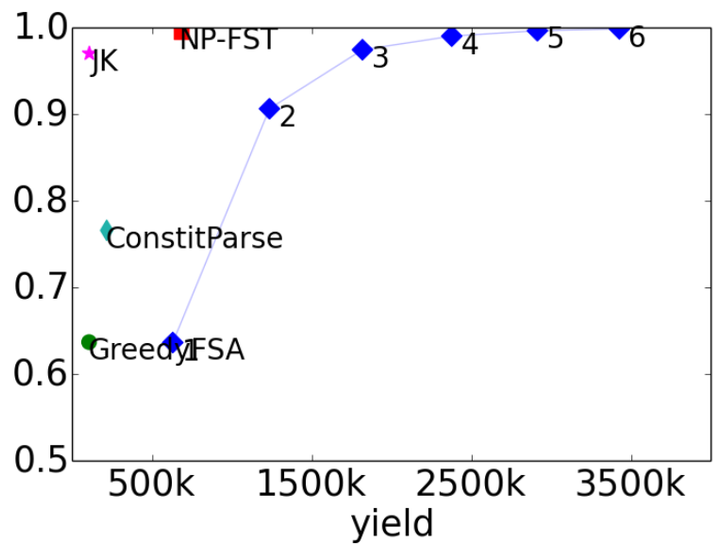
Evaluation: named entities



Twitter
(persons, brands...)



BioNLP
(proteins)



NYT
(persons, orgs...)

- Evaluation difficult: use previously defined named entity tasks as partial proxy
  - Compare: lexicon construction, keyphrase extraction
- Multitagging: allow any tag with posterior > 0.01
  - versus 1-best tagging

Rookie: Interactive news exploration

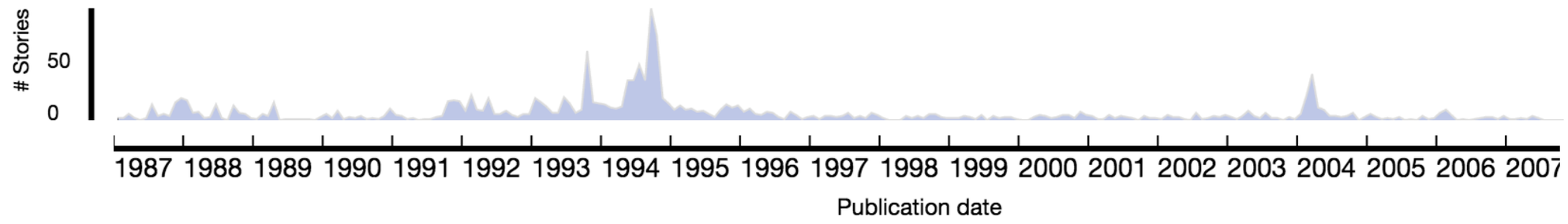
[Abram Handler, UMass]

United States

Submit

Found 1734 documents for **United States**

A

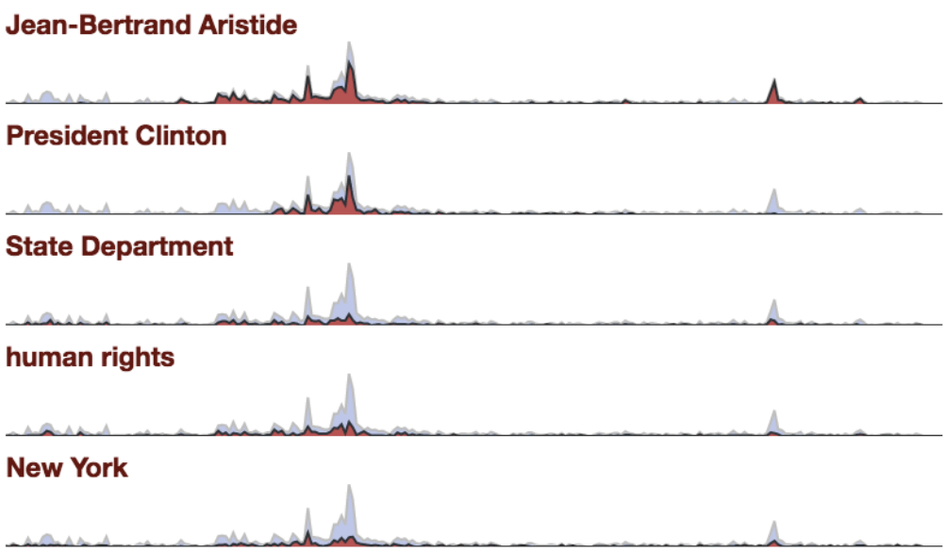


B

Subjects related to **United States**

more subjects

page 1 of 15



Summary of 1734 documents for **United States** from **Jan. 1987 — Dec. 2007**

More →

Oct. 21 1995 | The **United States**, which presided over the creation of the United Nations 50 years ago, is joining the big anniversary celebration next week as its most notorious deadbeat.

May. 17 1992 | Officials said the **United States** had also stopped flights of Awacs and Orion radar planes over Peru's Upper Huallaga Valley, where 60 percent of the world's coca leaves -- the raw material for cocaine -- is grown.

May. 06 1994 | It defines strict conditions for setting up peacekeeping operations and for committing **United States** troops to take part in them.

C

Congressional bills

- 97,221 U.S. congressional bills 1993–2014
  - 269,601,458 total tokens
  - NP extraction in a few hours
- Analysis
  - Partisan ranking by Dirichlet-based z-scores (Monroe, Colaresi, Quinn 2008)
    488 bills: law and crime, 2013–2014
  - Topic model
    All bills in 2013

z-score ranked terms per party

Uni. Dem.	and, deleted, health, mental, domestic, inserting, grant, programs, prevention, violence, program striking, education, forensic, standards, juvenile, grants, partner, science, research
Uni. Rep.	any, offense, property, imprisoned, whoever, person, more, alien, knowingly, officer, not, united, intent, commerce, communication, forfeiture, immigration, official, interstate, subchapter
Phrases Dem.	
Phrases Rep.	

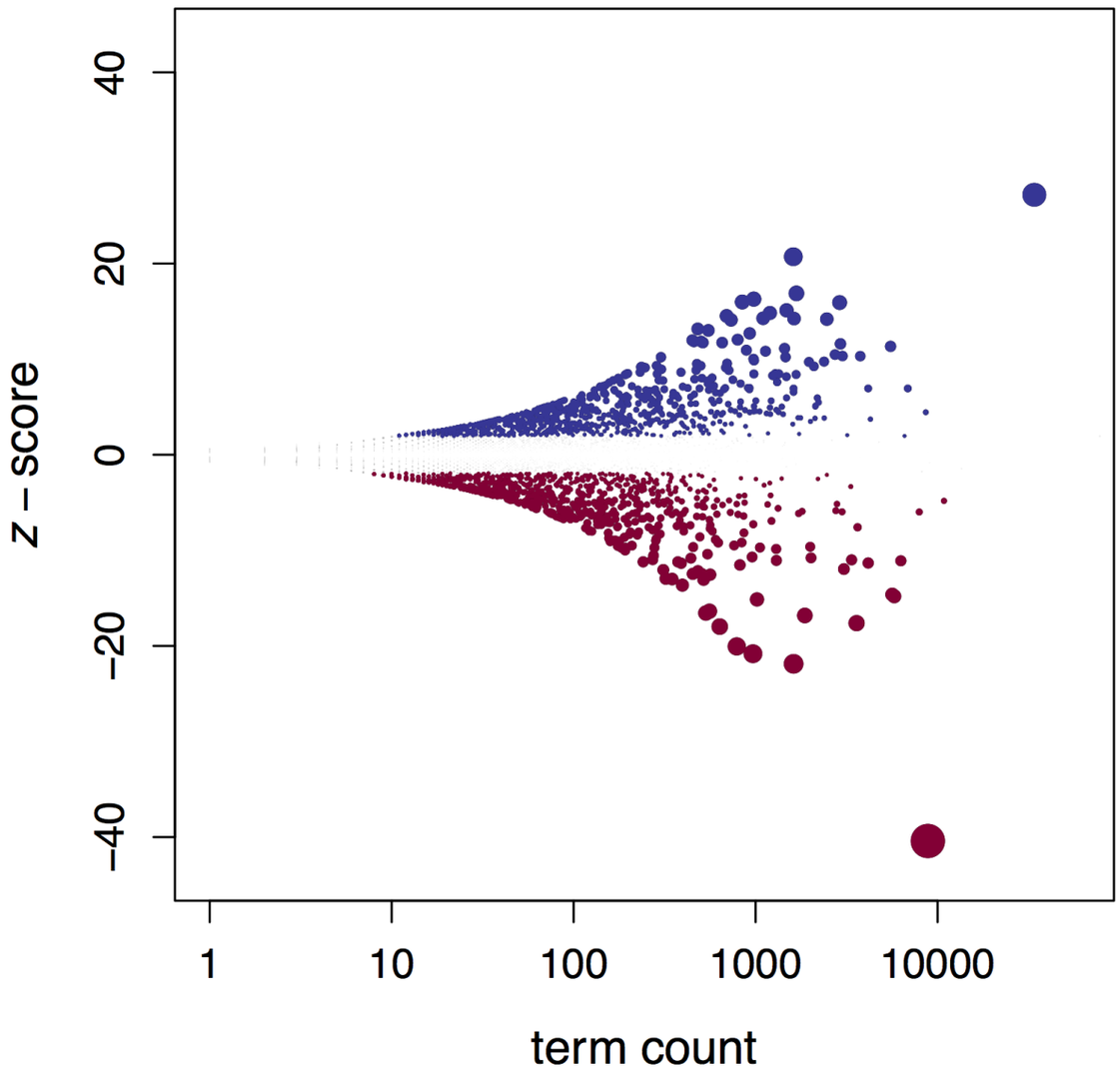
z-score ranked terms per party

Uni. Dem.	and, deleted, health, mental, domestic, inserting, grant, programs, prevention, violence, program striking, education, forensic, standards, juvenile, grants, partner, science, research
Uni. Rep.	any, offense, property, imprisoned, whoever, person, more, alien, knowingly, officer, not, united, intent, commerce, communication, forfeiture, immigration, official, interstate, subchapter
Phrases Dem.	mental health, juvenile justice and delinquency prevention act, victims of domestic violence, child support enforcement act of u.s.c., fiscal year, child abuse prevention and treatment act, omnibus crime control and safe streets act of u.s.c., date of enactment of this act, violence prevention, director of the national institute, former spouse, section of the foreign intelligence surveillance act of u.s.c., justice system, substance abuse criminal street gang, such youth, forensic science, authorization of appropriations, grant progran
Phrases Rep.	

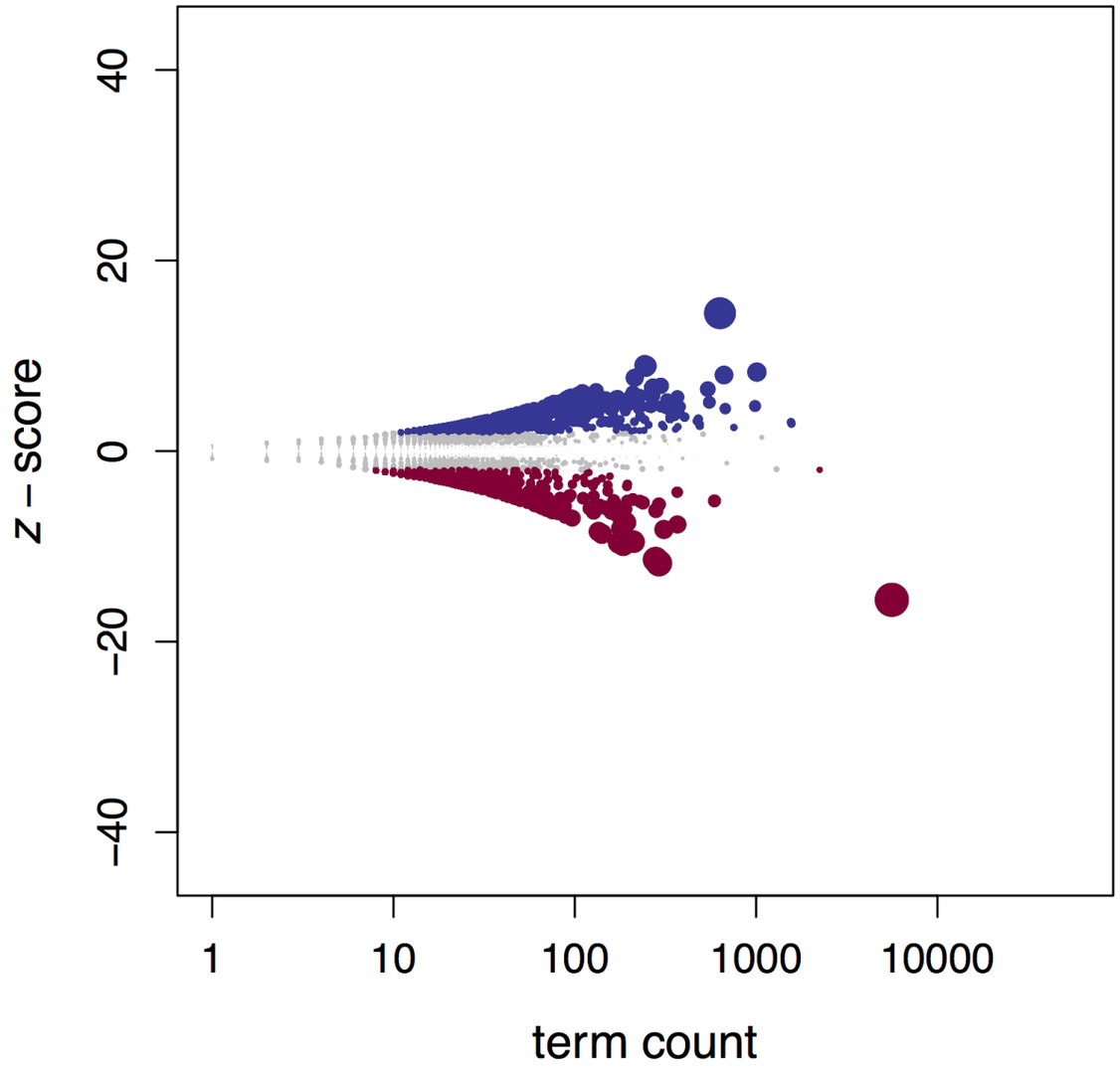
z-score ranked terms per party

Uni. Dem.	and, deleted, health, mental, domestic, inserting, grant, programs, prevention, violence, program striking, education, forensic, standards, juvenile, grants, partner, science, research
Uni. Rep.	any, offense, property, imprisoned, whoever, person, more, alien, knowingly, officer, not, united, intent, commerce, communication, forfeiture, immigration, official, interstate, subchapter
Phrases Dem.	mental health, juvenile justice and delinquency prevention act, victims of domestic violence, child support enforcement act of u.s.c., fiscal year, child abuse prevention and treatment act, omnibus crime control and safe streets act of u.s.c., date of enactment of this act, violence prevention, director of the national institute, former spouse, section of the foreign intelligence surveillance act of u.s.c., justice system, substance abuse criminal street gang, such youth, forensic science, authorization of appropriations, grant progran
Phrases Rep.	special maritime and territorial jurisdiction of the united states, interstate or foreign commerce, federal prison, section of the immigration and nationality act, electronic communication service provider, motor vehicles, such persons, serious bodily injury, controlled substances act, department or agency, one year, political subdivision of a state, civil action, section of the immigration and nationality act u.s.c., offense under this section, five years, bureau of prisons, foreign government, explosive materials, other person

Sparsity



(a) unigrams



(b) NPFST

Phrases in unsupervised learning

One LDA topic from 2013 congressional bills

	Terms (domain stop-terms removed)
unigrams	{**training**}, {**education**}, {**employment**}, {**workforce**}, {**service**}, {**job**}, {activity}, {state}, {individual}, {program}, {include}, {**labor**}, {**skill**}, {local}, {area}
phrases	{**educational agency**}, {**local educational agency**}, {**state educational agency**}, {**professional development**}, {**secondary school**}, {technical assistance}, {such agency}, {**higher education**}, {**elementary school**}, {**public school**}, {**student performance**}, {state plan}, {**education act**}, {such child}, {**national education**}, {**number of children**}

ShortVP: Extending the grammar for verb-argument patterns

Verb-Obj →

Tag Pattern	Example
VN	*reduce funding*
VAN	*encourage dissenting members*
VNN	*restrict federal agencies*
VDN	*establish a commission*

ShortVP: Extending the grammar for verb-argument patterns

	Tag Pattern	Example
Verb-Obj →	VN	*reduce funding*
	VAN	*encourage dissenting members*
	VNN	*restrict federal agencies*
	VDN	*establish a commission*
Verb-PP →	VPN	*prescribe in paragraph*

ShortVP: Extending the grammar for verb-argument patterns

	Tag Pattern	Example
Verb-Obj →	VN	*reduce funding*
	VAN	*encourage dissenting members*
	VNN	*restrict federal agencies*
	VDN	*establish a commission*
Verb-PP →	VPN	*prescribe in paragraph*
Subj-Verb →	ANV	*eligible employee means*
	NVV	*benefits are determined*

ShortVP: Extending the grammar for verb-argument patterns

	Tag Pattern	Example
Verb-Obj →	VN	*reduce funding*
	VAN	*encourage dissenting members*
	VNN	*restrict federal agencies*
	VDN	*establish a commission*
Verb-PP →	VPN	*prescribe in paragraph*
Subj-Verb →	ANV	*eligible employee means*
	NVV	*benefits are determined*

Table 7: Fifteen additional verb-phrases recovered from the complete text of H.R.5893 - Ansel Adams Act (a bill preventing government agencies from restricting photography at national parks), introduced in the 113th Congress.

{enacted regulations}, {**restrict photography**}, {**prohibit photography**}, {**threatened photographers**}, {**obtain permits**}, {buy insurance}, {are abridgments}, {are speech}, {bring home}, {make yosemite}, {allow restrictions}, {**require fees**}, {build public support}, {**restrict federal agencies**}, {buy insurance policies}

ShortVP: Extending the grammar for verb-argument patterns

	Tag Pattern	Example
Verb-Obj →	VN	*reduce funding*
	VAN	*encourage dissenting members*
	VNN	*restrict federal agencies*
	VDN	*establish a commission*
Verb-PP →	VPN	*prescribe in paragraph*
Subj-Verb →	ANV	*eligible employee means*
	NVV	*benefits are determined*

Table 7: Fifteen additional verb-phrases recovered from the complete text of H.R.5893 - Ansel Adams Act (a bill preventing government agencies from restricting photography at national parks), introduced in the 113th Congress.

{enacted regulations}, {**restrict photography**}, {**prohibit photography**}, {**threatened photographers**}, {**obtain permits**}, {buy insurance}, {are abridgments}, {are speech}, {bring home}, {make yosemite}, {allow restrictions}, {**require fees**}, {build public support}, {**restrict federal agencies**}, {buy insurance policies}

[At some point dependency/semantic parsing becomes preferable?]

Party Association of SHORTVP's Containing "increase"

Republican Sponsored Bills

Term	Rep.	Total
increase(s/ed/ing) ordinary loss	40	49
increase(s/ed/ing) project cost	26	32
increase(s/ed/ing) maximum penalty	16	21
increase(s/ed/ing) punishment	23	31
increase(s/ed/ing) contribution limit	14	19

Democrat Sponsored Bills

Term	Dem.	Total
increase(s/ed/ing) federal financial (aid)	29	35
increase(s/ed/ing) diversity	25	31
increase(s/ed/ing) sequestration	29	36
increase(s/ed/ing) wages	22	28
increase(s/ed/ing) accessibility	83	106

- Part-of-speech patterns allow fast and simple extraction of phrases for statistical text analysis
  - Work in many domains
    (politics is especially appropriate?)
  - Noun phrases are a useful default;
    or, write your own grammar
- Linguistic representations for future text-as-data research?

- ***phrasemachine***: open-source implementation in R and Python (basic NP grammar)
  - http://slanglab.cs.umass.edu/phrasemachine/
  - https://github.com/slanglab/phrasemachine