

# Social Media & Text Analysis

## lecture 5 - Paraphrase Data Sources

**CSE 5539-0010 Ohio State University**  
**Instructor: Alan Ritter**  
**Website: [socialmedia-class.org](http://socialmedia-class.org)**

# Natural Language Processing

Dan Jurafsky



## Language Technology

making good progress

mostly solved

### Spam detection

Let's go to Agra!



Buy V1AGRA ...



### Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

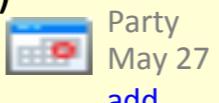
Colorless green ideas sleep furiously.

### Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

You're invited to our dinner party, Friday May 27 at 8:30



### Sentiment analysis

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.



### Coreference resolution

Carter told Mubarak he shouldn't run again.

### Word sense disambiguation (WSD)

I need new batteries for my **mouse**.



### Parsing



I can see Alcatraz from the window!

### Machine translation (MT)

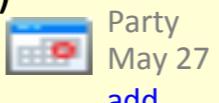
第13届上海国际电影节开幕...



The 13<sup>th</sup> Shanghai International Film Festival...

### Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



Party  
May 27  
add

still really hard

### Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

### Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

### Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose



Economy is good

### Dialog

Where is Citizen Kane playing in SF?



Castro Theatre at 7:30. Do you want a ticket?



# Natural Language Processing

Dan Jurafsky



## Language Technology

making good progress

mostly solved

### Spam detection

Let's go to Agra!



Buy V1AGRA ...



### Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

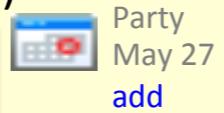
### Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

### Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



Party  
May 27  
add

### Sentiment analysis

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.



### Coreference resolution

Carter told Mubarak he shouldn't run again.

### Word sense disambiguation (WSD)

I need new batteries for my **mouse**.



### Parsing



I can see Alcatraz from the window!

### Machine translation (MT)

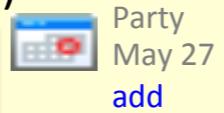
第13届上海国际电影节开幕...



The 13<sup>th</sup> Shanghai International Film Festival...

### Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



Party  
May 27  
add

still really hard

### Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

### Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

### Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose



Economy is good

### Dialog

Where is Citizen Kane playing in SF?



Castro Theatre at 7:30. Do you want a ticket?



# what is Paraphrase?

“sentences or phrases that convey approximately the same meaning using different words” — (Bhagat & Hovy, 2013)

# what is Paraphrase?

“sentences or phrases that convey approximately the same meaning using different words” — (Bhagat & Hovy, 2013)

*wealthy*

**word**

*rich*

# what is Paraphrase?

“sentences or phrases that convey approximately the same meaning using different words” — (Bhagat & Hovy, 2013)

*wealthy*

**word**

*rich*

*the king's speech*

**phrase**

*His Majesty's address*

# what is Paraphrase?

“sentences or phrases that convey approximately the same meaning using different words” — (Bhagat & Hovy, 2013)

*wealthy*

**word**

*rich*

*the king's speech*

**phrase**

*His Majesty's address*

*... the forced resignation  
of the CEO of Boeing,  
Harry Stonecipher, for ...*

**sentence**

*... after Boeing Co. Chief  
Executive Harry Stonecipher  
was ousted from ...*

# What's good about Paraphrases ?

*... the forced resignation  
of the CEO of Boeing,  
Harry Stonecipher, for ...*

*... after Boeing Co. Chief  
Executive Harry Stonecipher  
was ousted from ...*

# What's good about Paraphrases ?

**fundamentally useful for a wide range of applications**

*... the forced resignation  
of the CEO of Boeing,  
Harry Stonecipher, for ...*

*... after Boeing Co. Chief  
Executive Harry Stonecipher  
was ousted from ...*

# What's good about Paraphrases ?

**fundamentally useful for a wide range of applications**

## e.g. Question Answering

Who is the CEO stepping down from Boeing?

*... the forced resignation  
of the CEO of Boeing,  
Harry Stonecipher, for ...*

*... after Boeing Co. Chief  
Executive Harry Stonecipher  
was ousted from ...*

# What's good about Paraphrases ?

**fundamentally useful for a wide range of applications**

## e.g. Question Answering

Who is the CEO stepping down from Boeing?

**match**

*... the forced resignation of the CEO of Boeing, Harry Stonecipher, for ...*

*... after Boeing Co. Chief Executive Harry Stonecipher was ousted from ...*



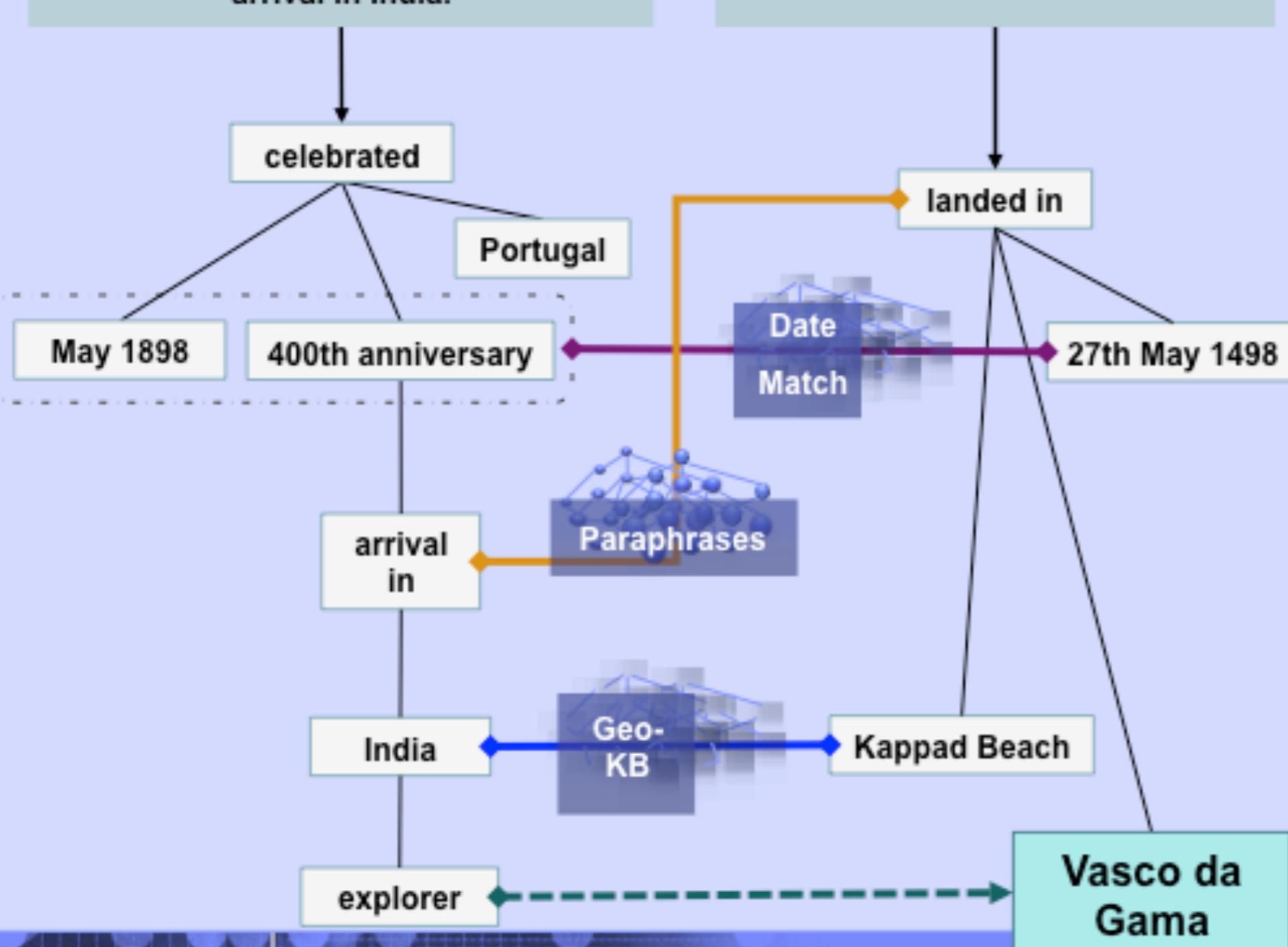
# Watson leverages multiple algorithms to perform deeper analysis

## [Question]

In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.

## [Supporting Evidence]

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach



## Legend

- Temporal Reasoning
- Statistical Paraphrasing
- GeoSpatial Reasoning
- Reference Text
- Answer

*Stronger evidence can be much harder to find and score...*

- Search far and wide
- Explore many hypotheses
- Find judge evidence
- Many inference algorithms



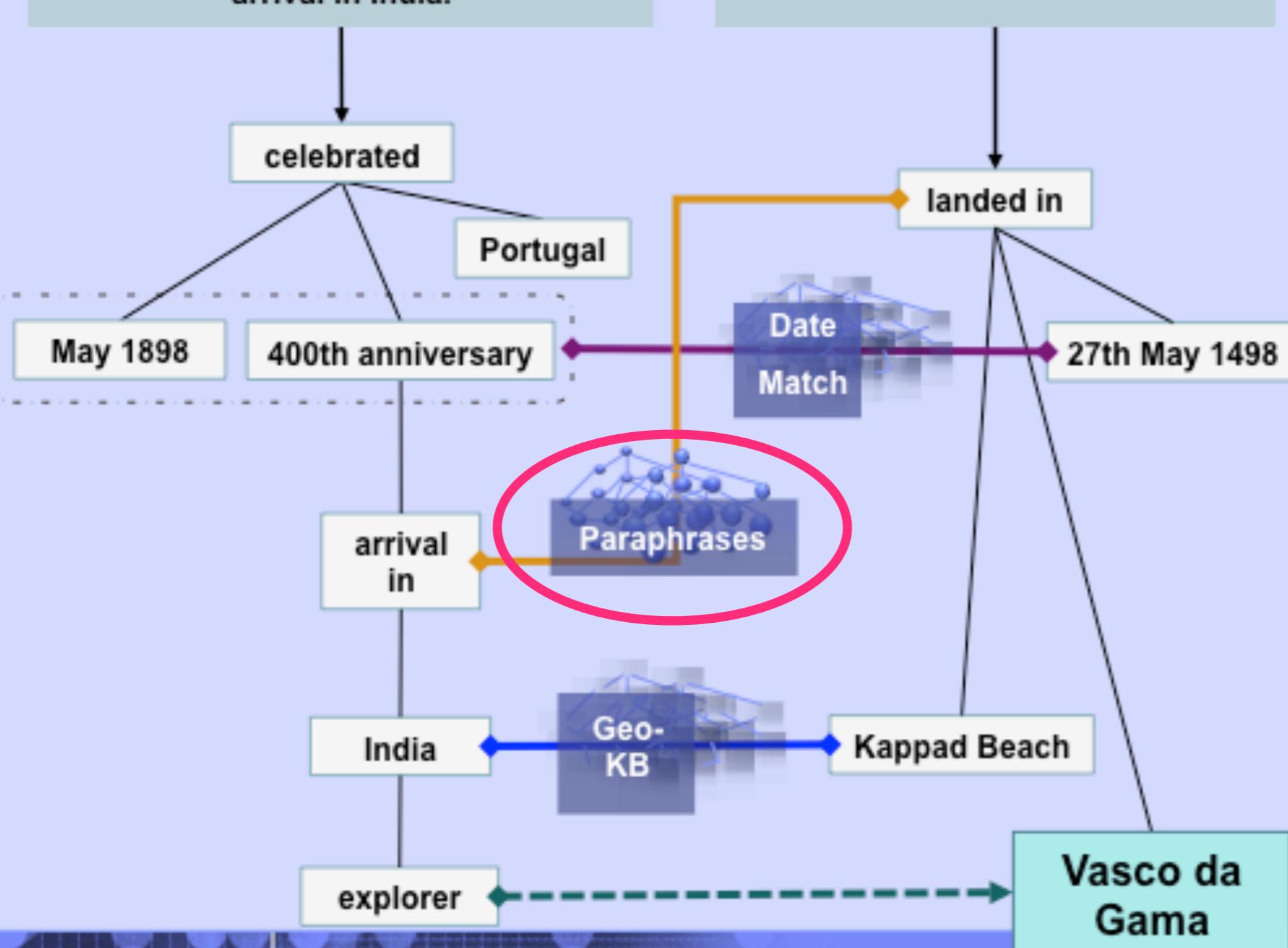
# Watson leverages multiple algorithms to perform deeper analysis

## [Question]

In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.

## [Supporting Evidence]

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach



## Legend

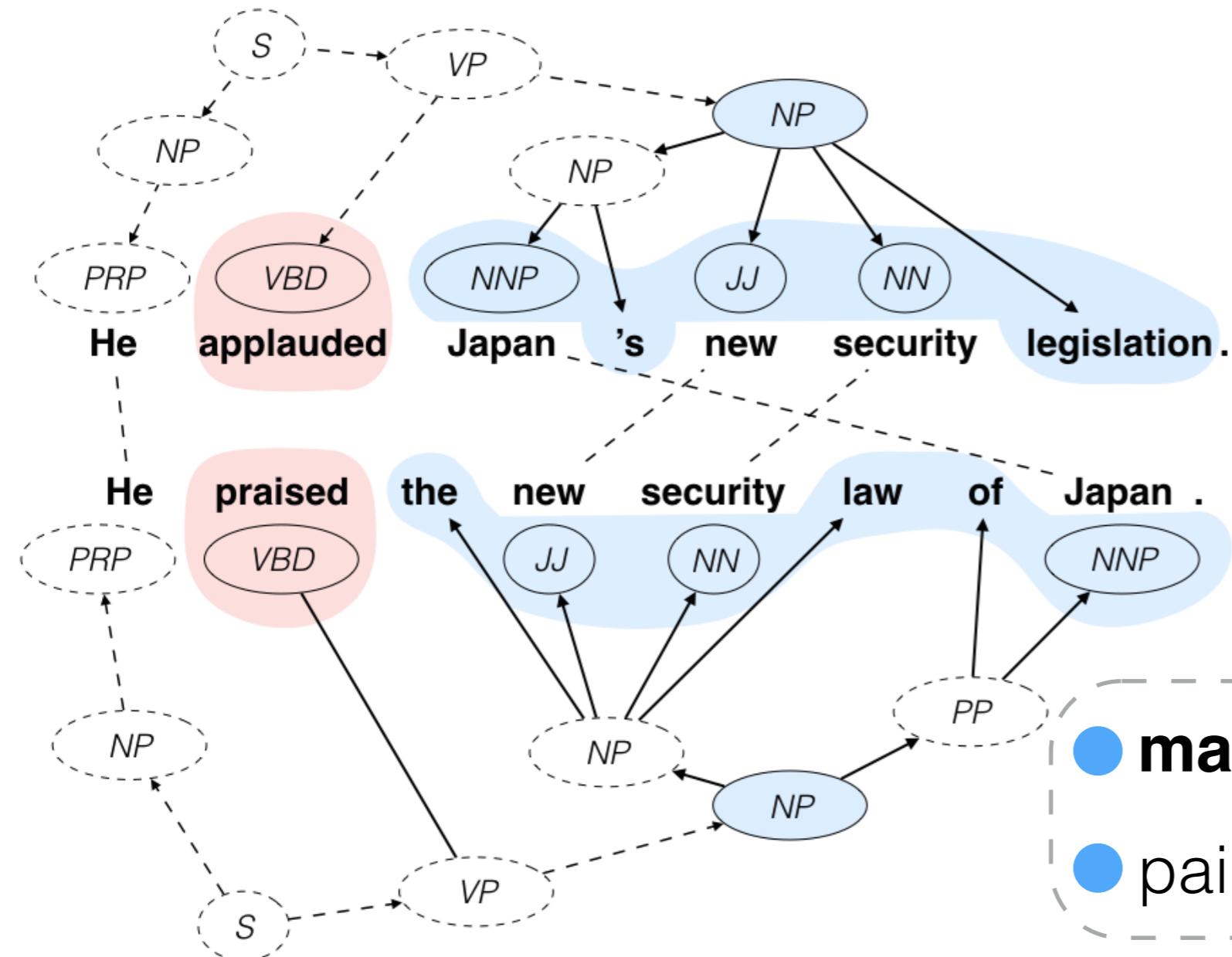
- Temporal Reasoning
- Statistical Paraphrasing
- GeoSpatial Reasoning
- Reference Text
- Answer

*Stronger evidence can be much harder to find and score...*

- Search far and wide
- Explore many hypotheses
- Find judge evidence
- Many inference algorithms

# Natural Language Generation

e.g. Text Simplification



## Techniques

- machine translation

- pairwise ranking optimization

# Digital Humanities



e.g. Stylistic Rewriting / Poetry Generation



Palpatine:  
*If you will not be turned, you will be destroyed!*

↓

*If you will not be turn'd, you will be undone!*

Luke:  
*Father, please! Help me!*

↓

*Father, I pray you! Help me!*



Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. "Paraphrasing for Style" In COLING (2012)

Quanze Chen, Chenyang Lei, Wei Xu, Ellie Pavlick, Chris Callison-Burch.

"Poetry of the Crowd: A Human Computation Algorithm to Convert Prose into Rhyming Verse" In HCOMP (2014)

# Language, Vision, Robotics, VR



Pick up a black table leg off of the floor.  
Pick up the black table leg.  
Walk over to the white table.  
Place black leg on white table bottom.  
Locate the black table leg on the floor by the white table.  
Find the black table leg and attach it to the white table.

# Language, Vision, Robotics, VR



Pick up a black table leg off of the floor.  
Pick up the black table leg.  
Walk over to the white table.  
Place black leg on white table bottom.  
Locate the black table leg on the floor by the white table.  
Find the black table leg and attach it to the white table.

Paraphrases!!

Paraphrases!!

# Other Applications

**fundamentally useful for a wide range of applications**

# Other Applications

**fundamentally useful for a wide range of applications**

semantic similarity \*

\* my research

# Other Applications

**fundamentally useful for a wide range of applications**

semantic similarity \*  
machine translation \*

\* my research

# Other Applications

**fundamentally useful for a wide range of applications**

- semantic similarity \*
- machine translation \*
- summarization \*

\* my research

# Other Applications

**fundamentally useful for a wide range of applications**

- semantic similarity \*
- machine translation \*
- summarization \*
- social science \*

\* my research

# Other Applications

**fundamentally useful for a wide range of applications**

- semantic similarity \*
- machine translation \*
- summarization \*
- social science \*
- information extraction \*

\* my research

# Other Applications

**fundamentally useful for a wide range of applications**

- semantic similarity \*
- machine translation \*
- summarization \*
- social science \*
- information extraction \*
- information retrieval \*

\* my research

# Other Applications

**fundamentally useful for a wide range of applications**

- semantic similarity \*
- machine translation \*
- summarization \*
- social science \*
- information extraction \*
- information retrieval \*
- semantic parsing

\* my research

# Other Applications

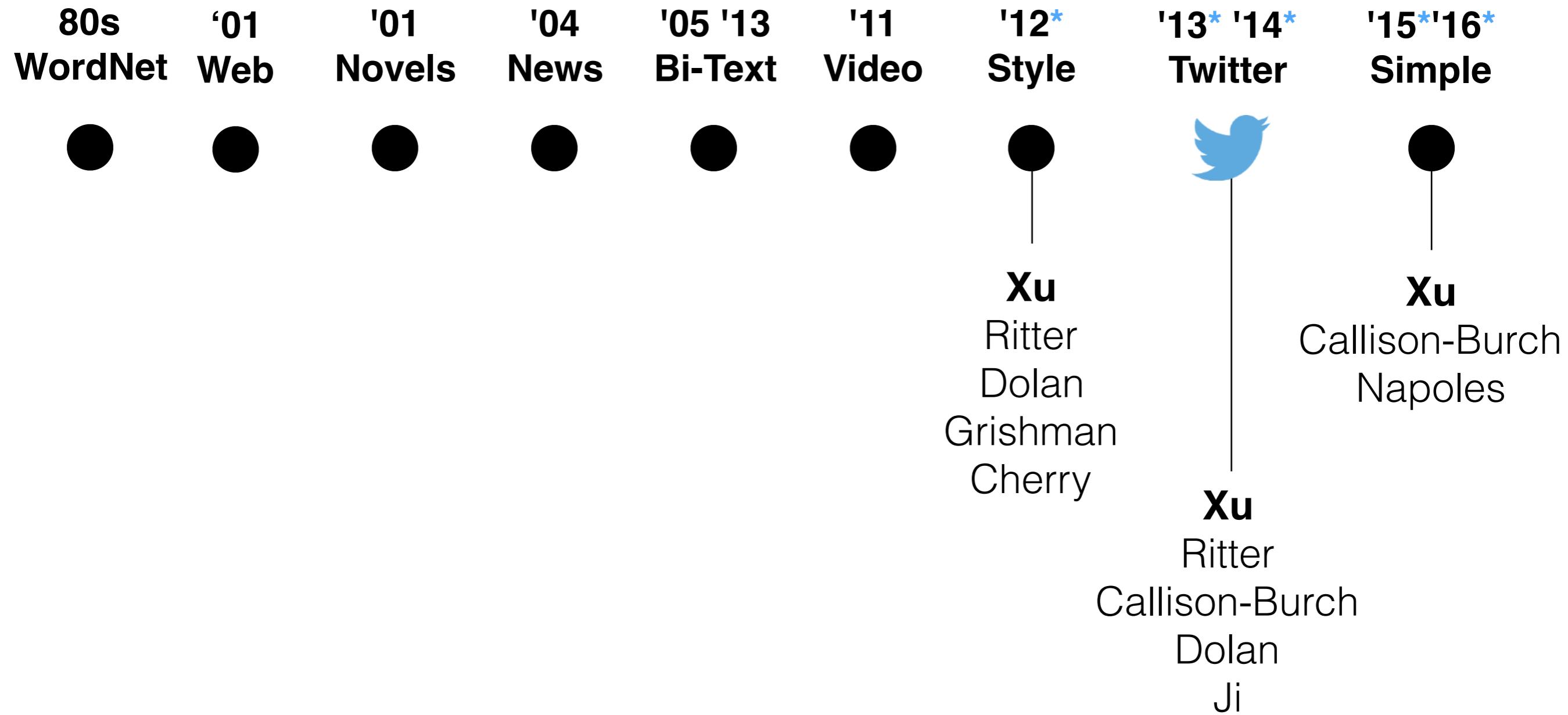
**fundamentally useful for a wide range of applications**

- semantic similarity \*
- machine translation \*
- summarization \*
- social science \*
- information extraction \*
- information retrieval \*
- semantic parsing

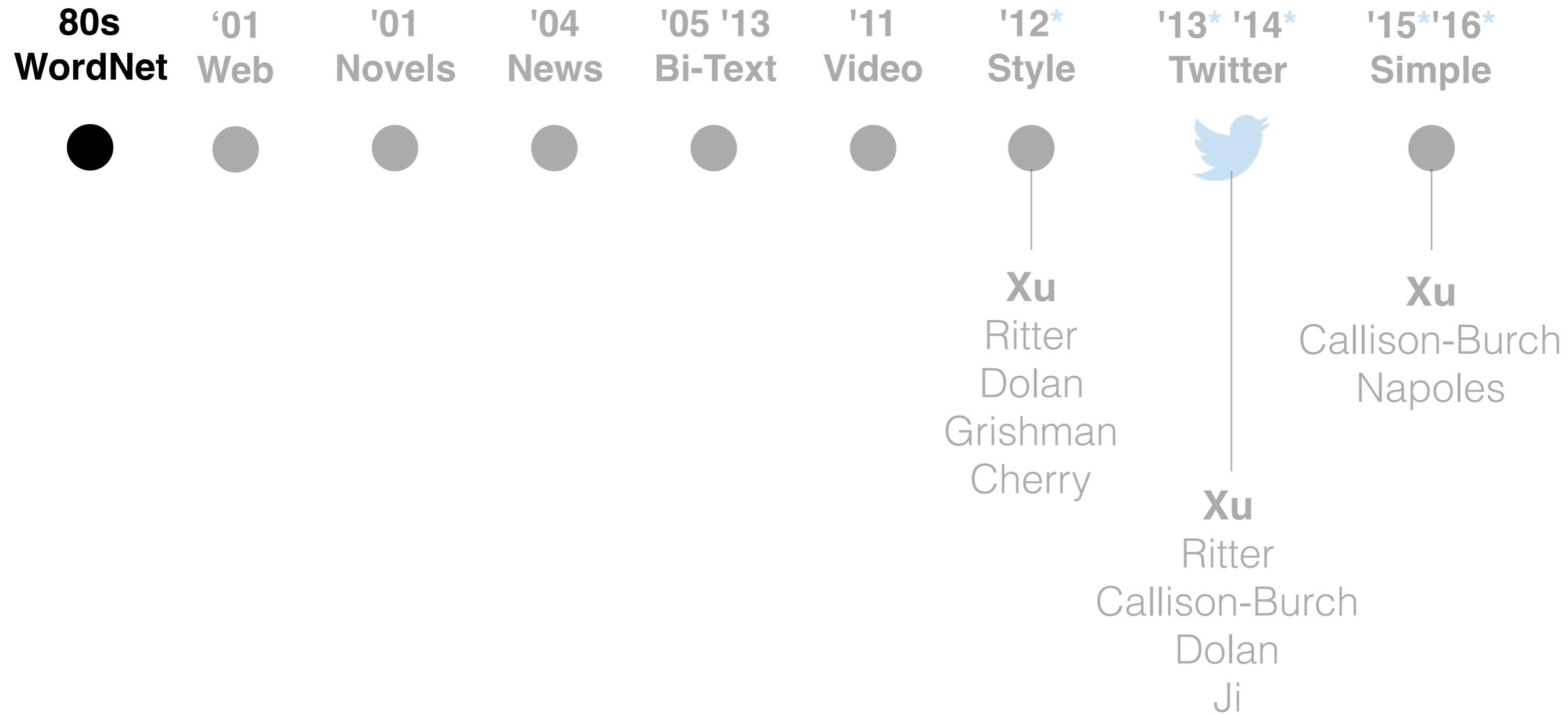
...

\* my research

# Paraphrase Research



# Paraphrase Research



# WordNet®

- What is it?
  - a large lexical database of English (155,287 words, latest version in 2005~6)
  - created (from mid-1980s) and maintained by Cognitive Science Lab of Princeton University
  - designed to establish the connections between words

# WordNet®

- What is it?
  - a combination of dictionary and thesaurus
  - try it out <http://wordnet.princeton.edu/>
  - In other languages: <http://globalwordnet.org/wordnets-in-the-world/>

Dictionary contains meaning, definition, pronunciation, orthography, and etymology of a word.

Thesaurus contains synonyms and antonyms of words.

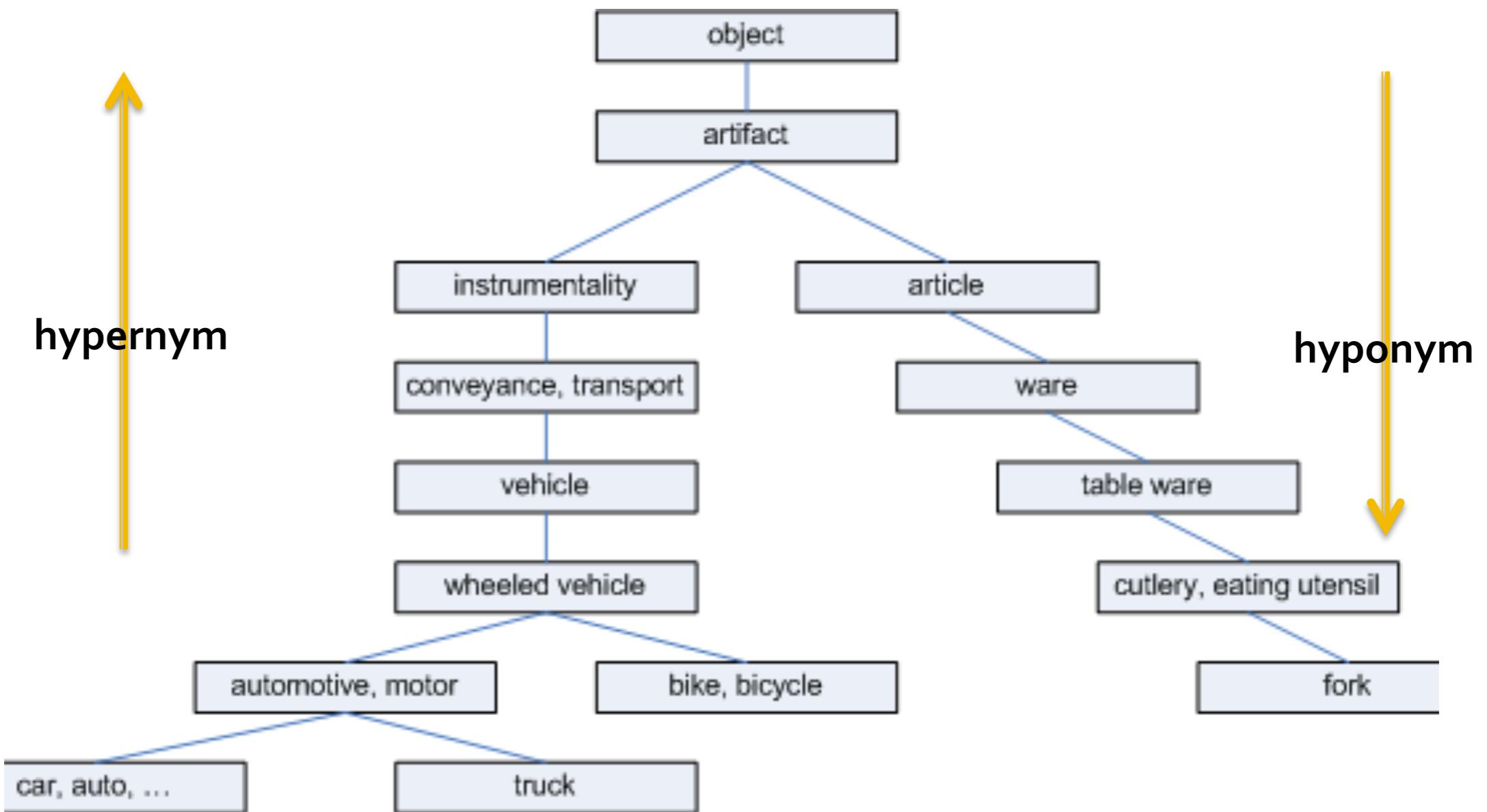
# WordNet®

- 4 types of Parts of Speech (POS)
  - Noun, Verb, Adjective, Adverb
- Synset (synonym set)
  - the smallest unit in WordNet
  - represents a specific meaning of a word
- S: (n) search (an investigation seeking answers) "*a thorough search of the ledgers revealed nothing*"; "*the outcome justified the search*"
- S: (v) search, seek, look for (try to locate or discover, or try to establish the existence of) "*The police are searching for clues*"; "*They are searching for the missing man in the entire county*"

# WordNet®

- Synsets are connected to one another through semantic and lexical relations
- Type of relations (based on POS)
  - hypernyms (kind-of): ‘vehicle’ is a hypernym of ‘car’
  - hyponyms (kind-of): ‘car’ is a hyponym of ‘vehicle’
  - holonym (part-of): ‘building’ is a holonym of ‘window’
  - meronym(part-of): ‘window’ is a meronym of ‘building’
  - similar to: ‘smart’ is similar to ‘intelligent’
  - antonyms: ‘smart’ is antonym of ‘unintelligent’

# WordNet®



# WordNet®

- Interfaces
  - Unix-style manual
  - Web Interfaces
  - Local Interfaces/APIs (Java, Python, Perl, C# ...)

<http://wordnet.princeton.edu/wordnet/related-projects/>

# WordNet®

Google Scholar  

Articles About 94,700 results (0.08 sec)

**Any time** [\[PDF\] semanticscholar.org](#)

Since 2017  
Since 2016  
Since 2013  
Custom range...

---

**Sort by relevance** [\[BOOK\] WordNet](#)  
Sort by date

---

include patents  
 include citations

---

 Create alert

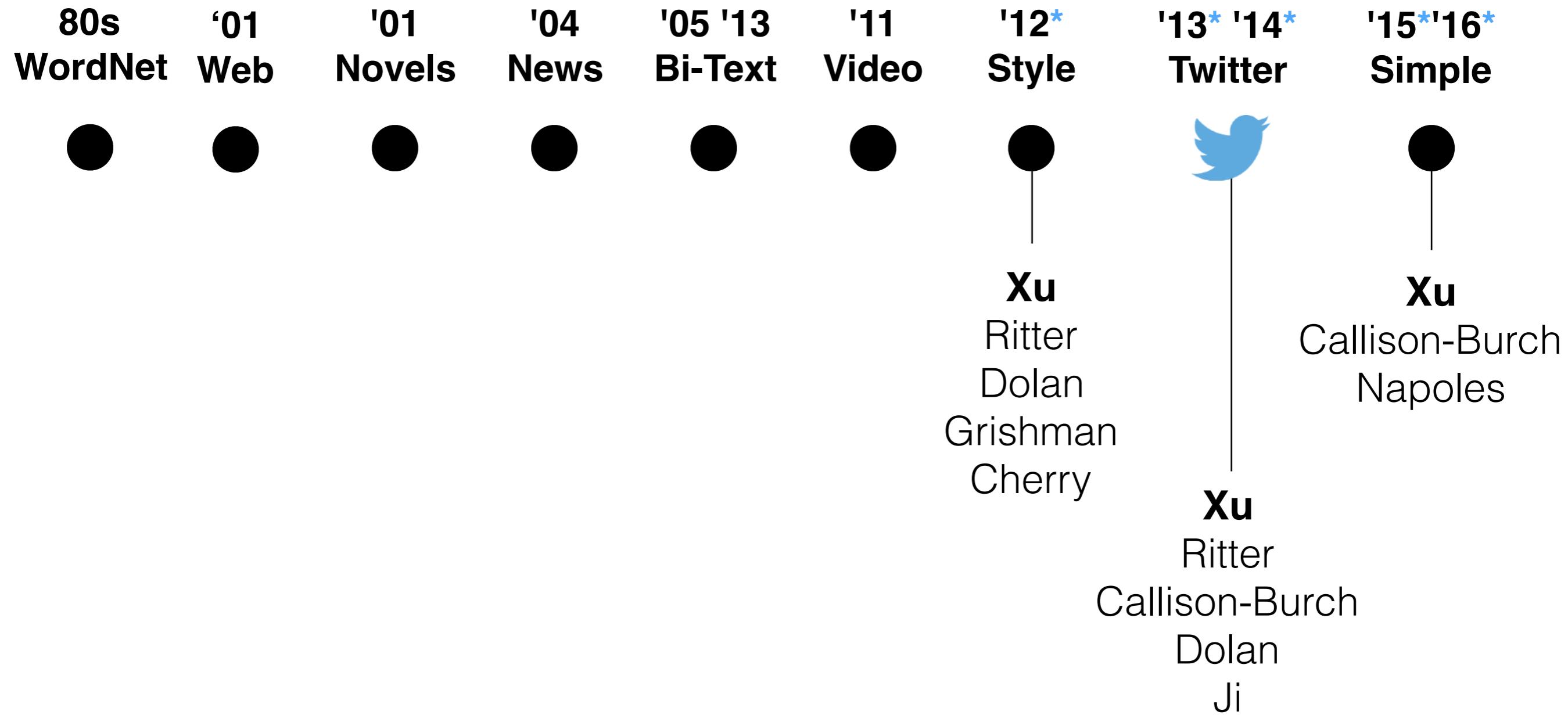
**WordNet: a lexical database for English** [\[PDF\] semanticscholar.org](#)  
GA Miller - Communications of the ACM, 1995 - dl.acm.org  
Abstract Because meaningful sentences are composed of meaningful words, any system that hopes to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, and  
☆ 99 Cited by 9594 Related articles All 34 versions Web of Science: 2440 »»

**[BOOK] WordNet** [\[PDF\] academia.edu](#)  
C Fellbaum - 1998 - Wiley Online Library  
Abstract **WordNet** (Miller, Beckwith, Fellbaum, Gross, & Miller 1990; Miller & Fellbaum, 1991; Miller, 1995; Fellbaum, 1998), a lexical database for English, can be thought of as a large electronic dictionary. It contains information about some 155,000 nouns, verbs, adjectives,  
☆ 99 Cited by 13461 Related articles All 12 versions »»

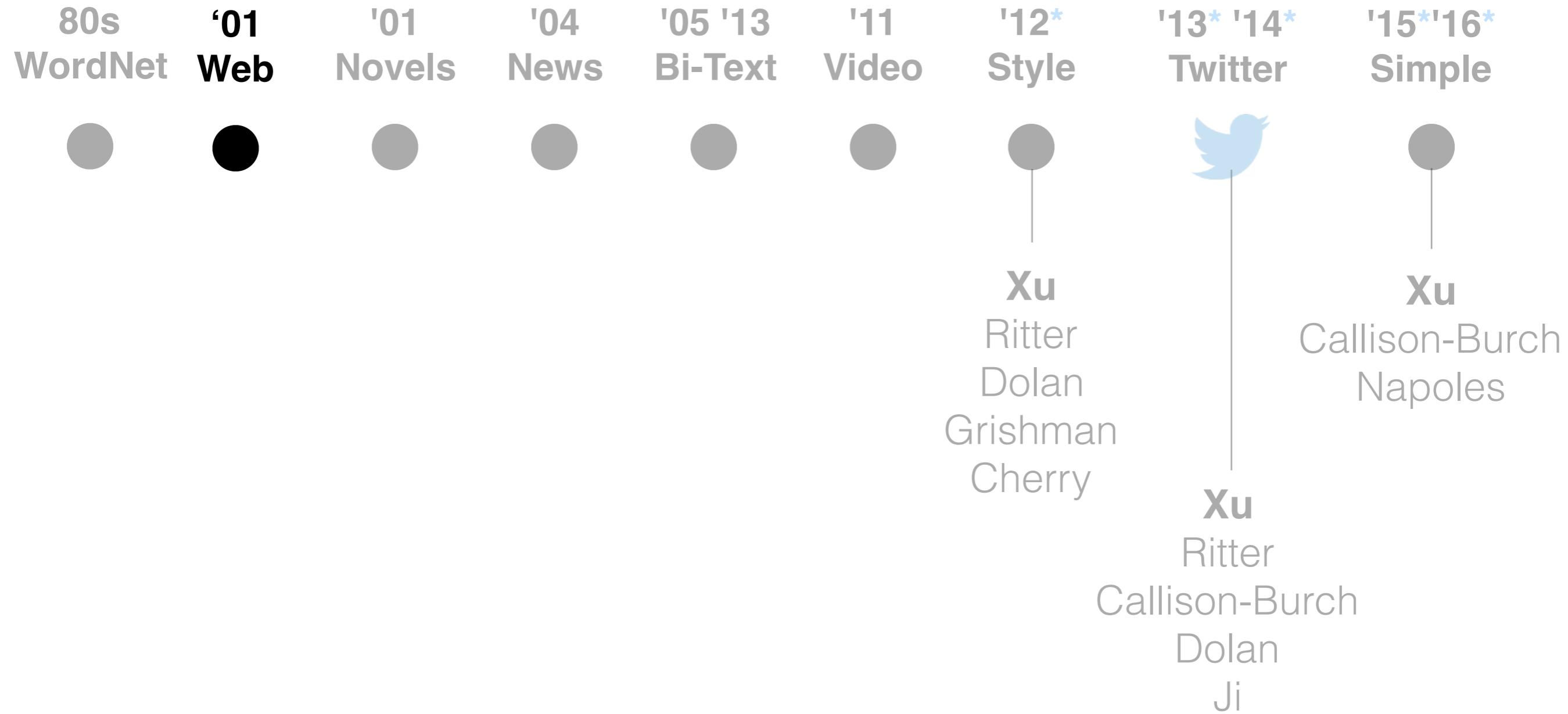
**Introduction to WordNet: An on-line lexical database** [\[PDF\] academia.edu](#)  
GA Miller, R Beckwith, C Fellbaum... - International journal ..., 1990 - academic.oup.com  
Abstract **WordNet** is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different  
☆ 99 Cited by 5707 Related articles All 80 versions »»

**WordNet:: Similarity: measuring the relatedness of concepts** [\[PDF\] aaai.org](#)  
T Pedersen, S Patwardhan, J Michelizzi - Demonstration papers at HLT- ..., 2004 - dl.acm.org  
Abstract **WordNet:: Similarity** is a freely available software package that makes it possible to measure the semantic similarity and relatedness between a pair of concepts (or synsets). It provides six measures of similarity, and three measures of relatedness, all of which are  
☆ 99 Cited by 1504 Related articles All 37 versions

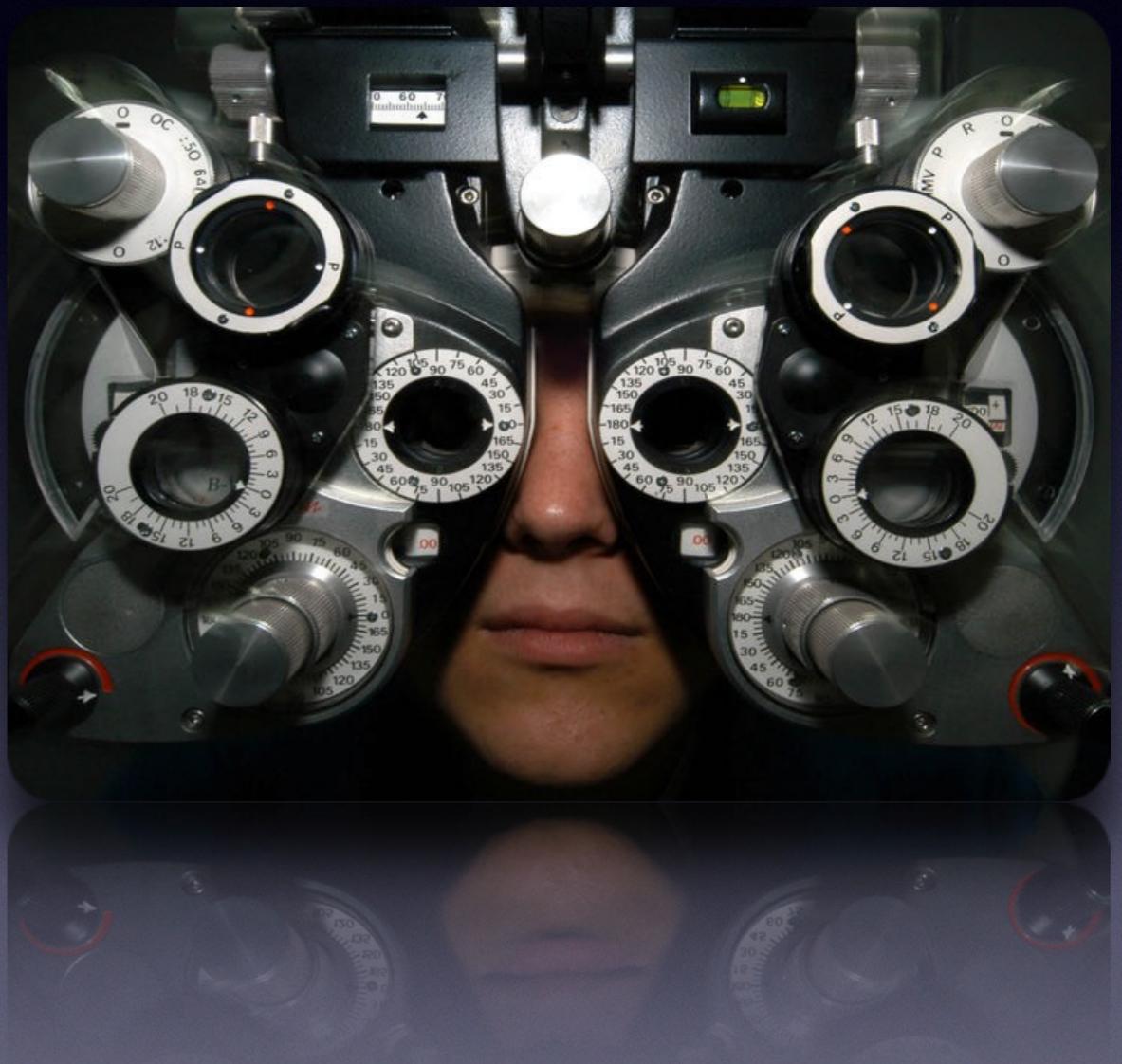
# Paraphrase Research



# Paraphrase Research



# Distributional Hypothesis



Source: Chris Callison-Burch

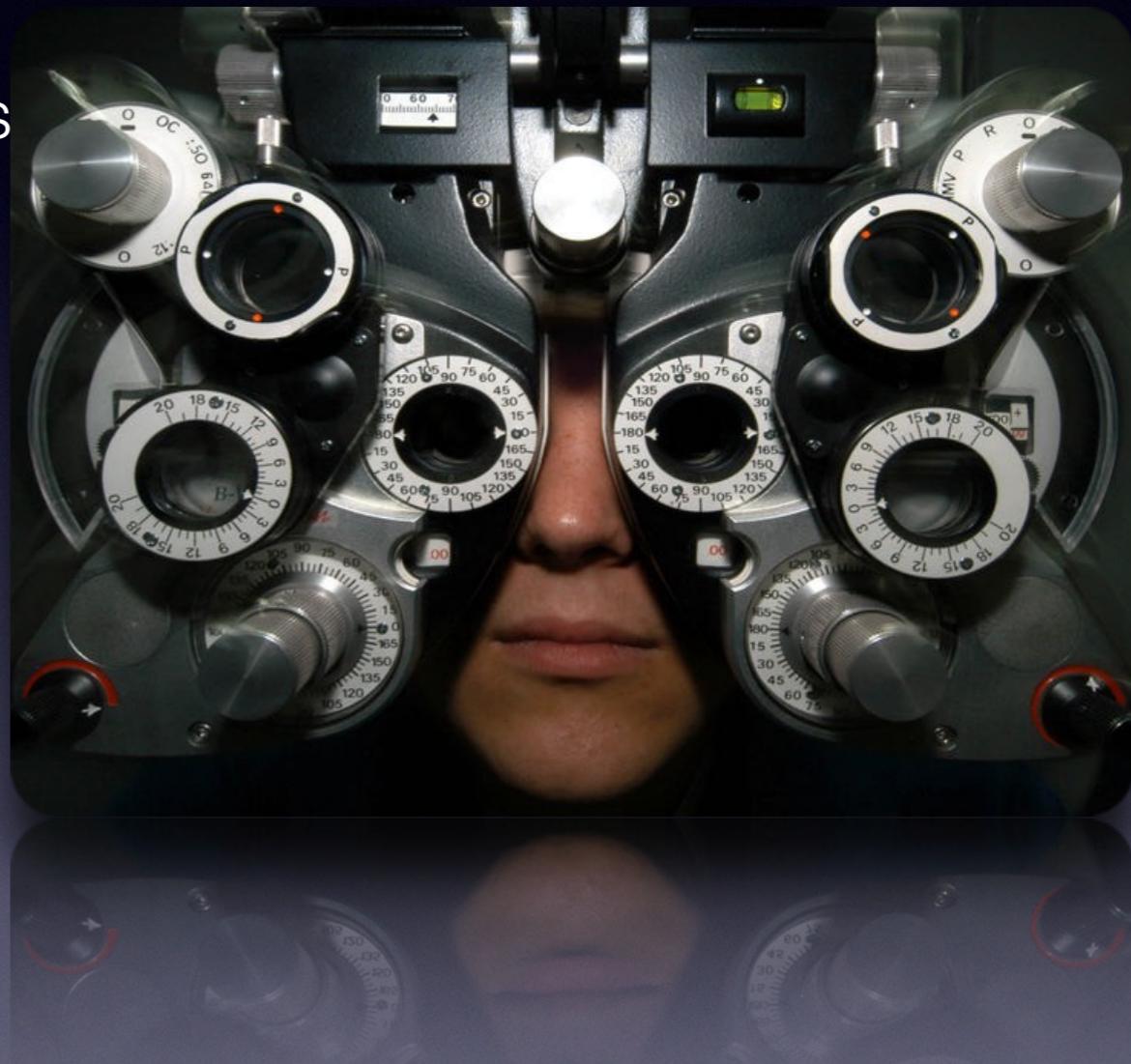
# Distributional Hypothesis

If we consider **oculist** and **eye-doctor** we find that, as our corpus of utterances grows, these two occur in almost the same environments. In contrast, there are many sentence environments in which **oculist** occurs but **lawyer** does not...

It is a question of the relative frequency of such environments, and of what we will obtain if we ask an informant to substitute any word he wishes for **oculist** (not asking what words have the same meaning).

These and similar tests all measure the probability of particular environments occurring with particular elements... If A and B have almost identical environments we say that they are synonyms.

—Zellig Harris (1954)



# DIRT

## (Discovery of Inference Rules from Text)

Lin and Pantel (2001) operationalize the Distributional Hypothesis using dependency relationships to define similar environments.

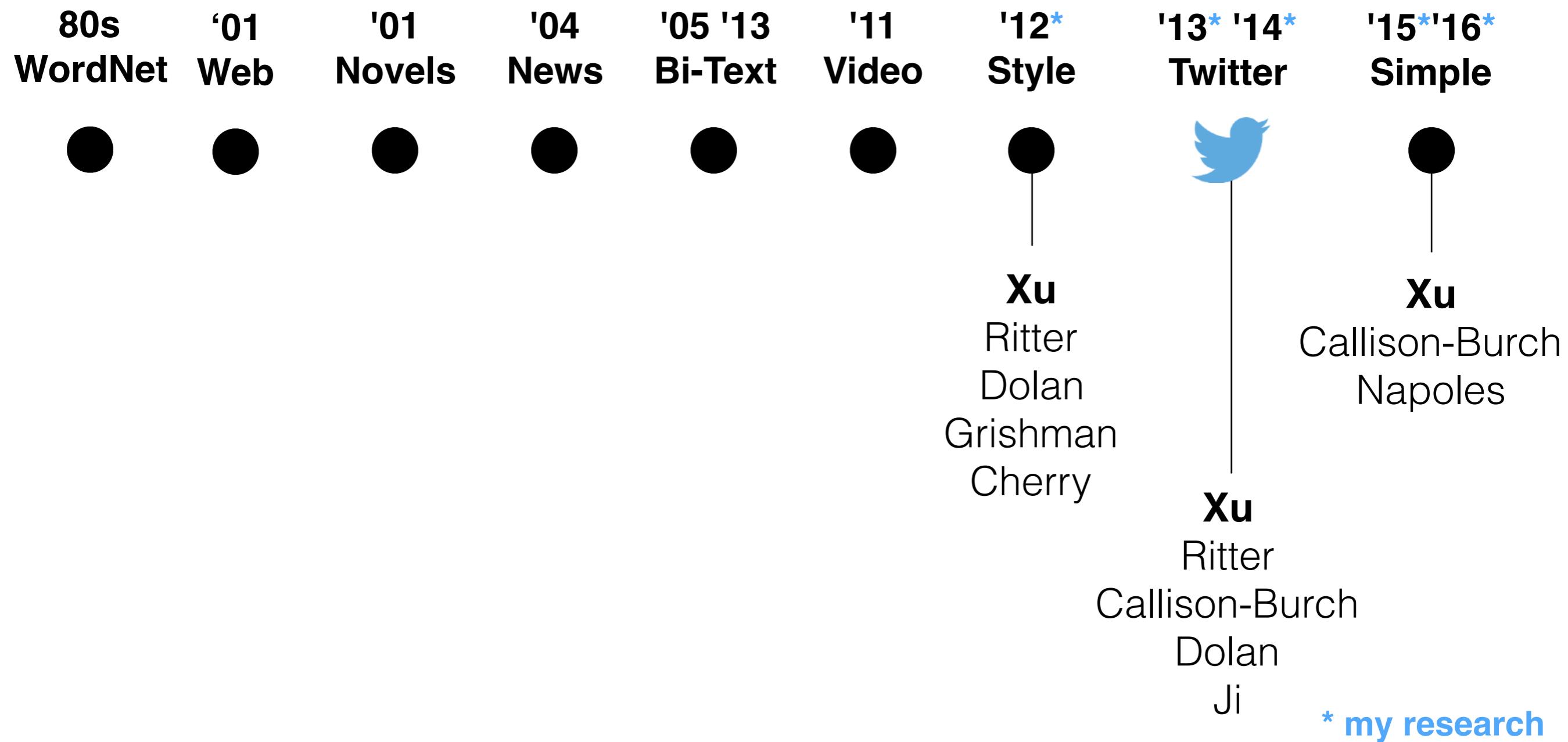
Duty and responsibility share a similar set of dependency contexts in large volumes of text:

modified by adjectives	objects of verbs
additional, administrative, assigned, assumed, collective, congressional, constitutional ...	assert, assign, assume, attend to, avoid, become, breach ...

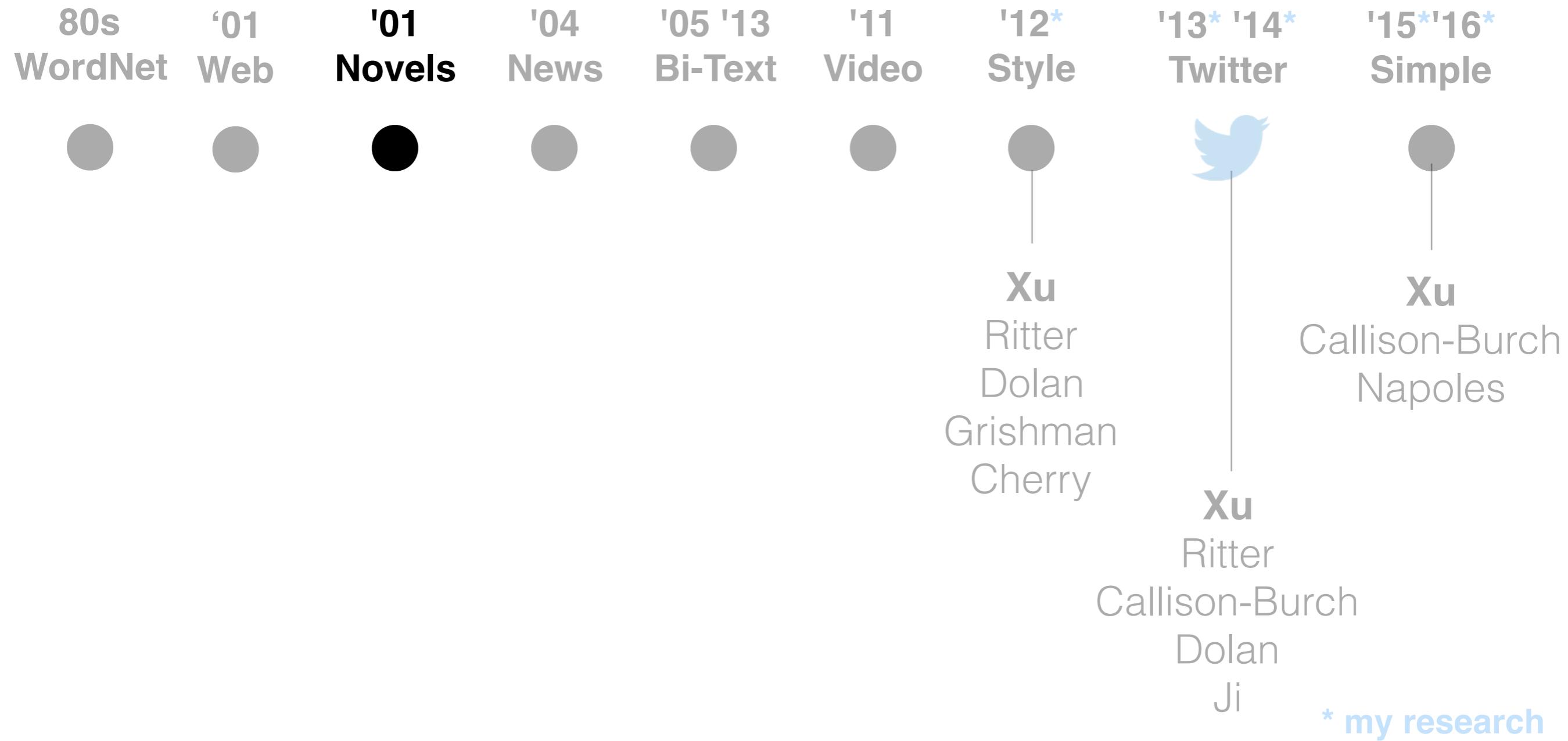
Source: Chris Callison-Burch

Deching Lin and Patrick Pantel. "DIRT - Discovery of Inference Rules from Text" In KDD (2001)

# Paraphrase Research



# Paraphrase Research



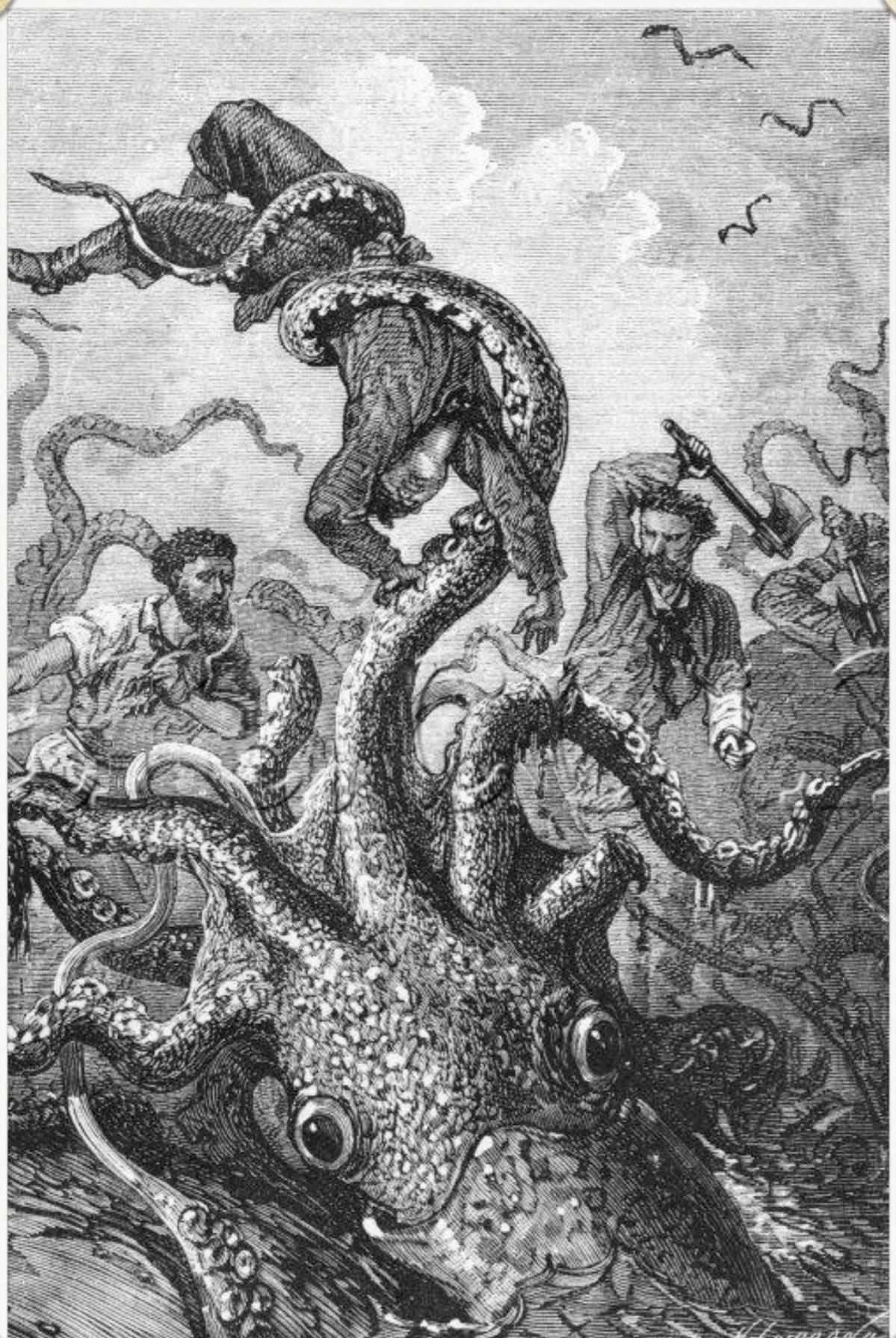


Source: Chris Callison-Burch



What a scene! Seized by the tentacle and **glued to** its suckers, the unfortunate man was **swinging in the air** at the **mercy** of this enormous appendage. He gasped, he choked, he yelled: "Help! Help!" I'll hear his **harrowing plea** the rest of my life!

The poor fellow was **done for.**



What a scene! Seized by the tentacle and **glued to** its suckers, the unfortunate man was **swinging in the air** at the **mercy** of this enormous appendage. He gasped, he choked, he yelled: "Help! Help!" I'll hear his **harrowing plea** the rest of my life!  
The **poor fellow** was **done for**.

What a scene! The unhappy man, seized by the tentacle and **fixed to** its suckers, was **balanced in the air** at the **caprice** of this enormous trunk. He rattled in his throat, he was stifled, he cried, "Help! help!" That **heart-rending cry**! I shall hear it all my life.  
The **unfortunate man** was **lost**.

# Novels (parallel monolingual data)

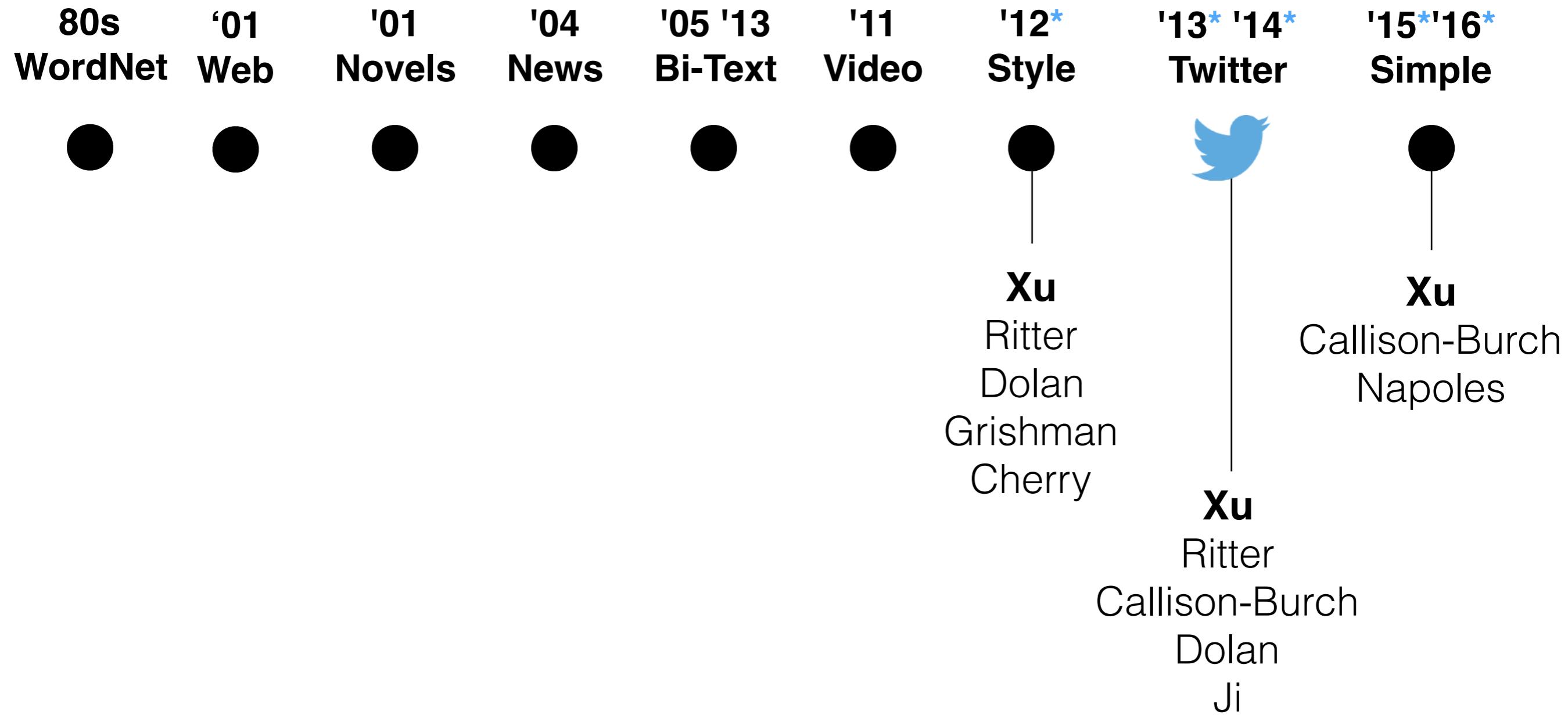
Barzilay and McKeown (2001) identify paraphrases using identical contexts in aligned sentences:

Emma burst into tears and he tried to comfort her,  
saying things to make her smile.

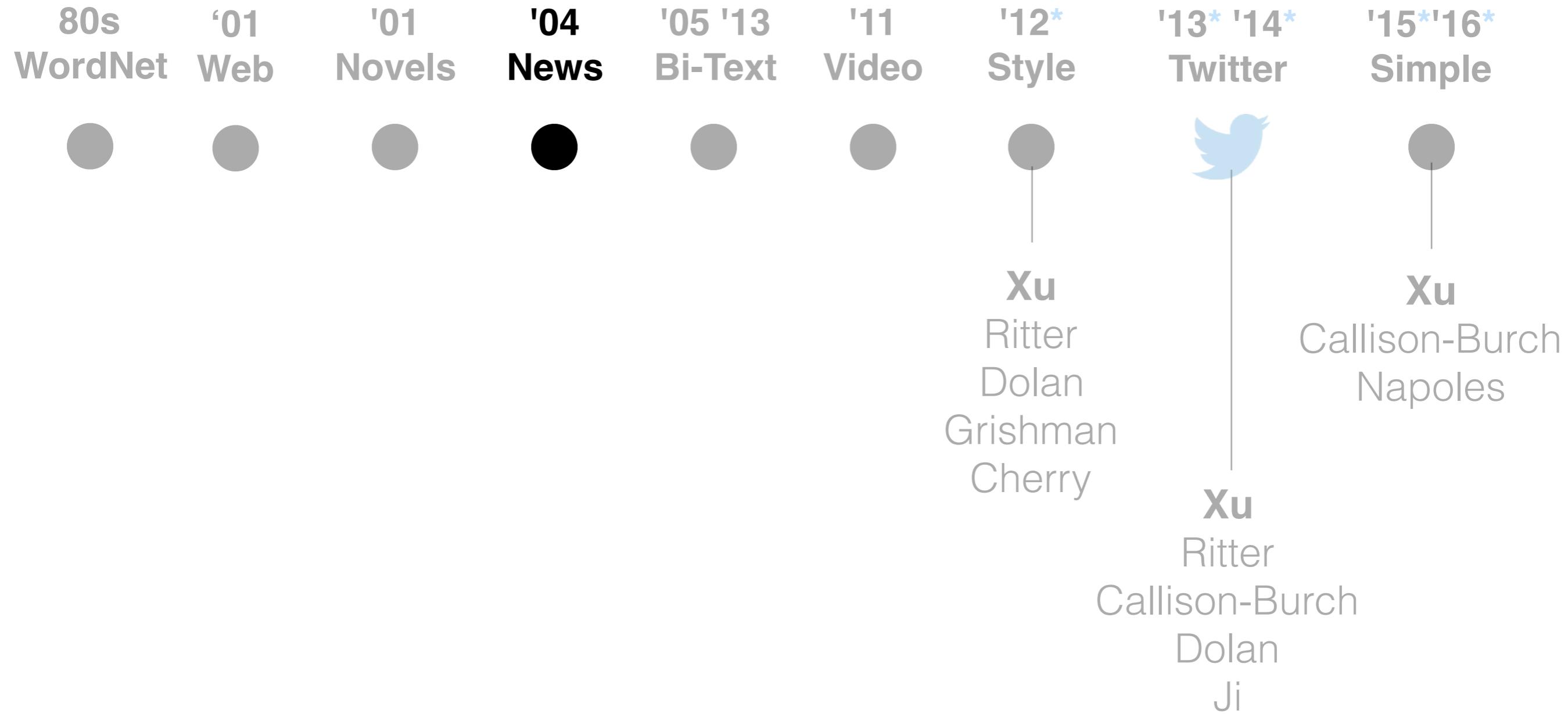
Emma cried and he tried to console her, adorning  
his words with puns.

burst into tears = cried and comfort = console

# Paraphrase Research



# Paraphrase Research



# News



Microsoft Research Paraphrase Corpus

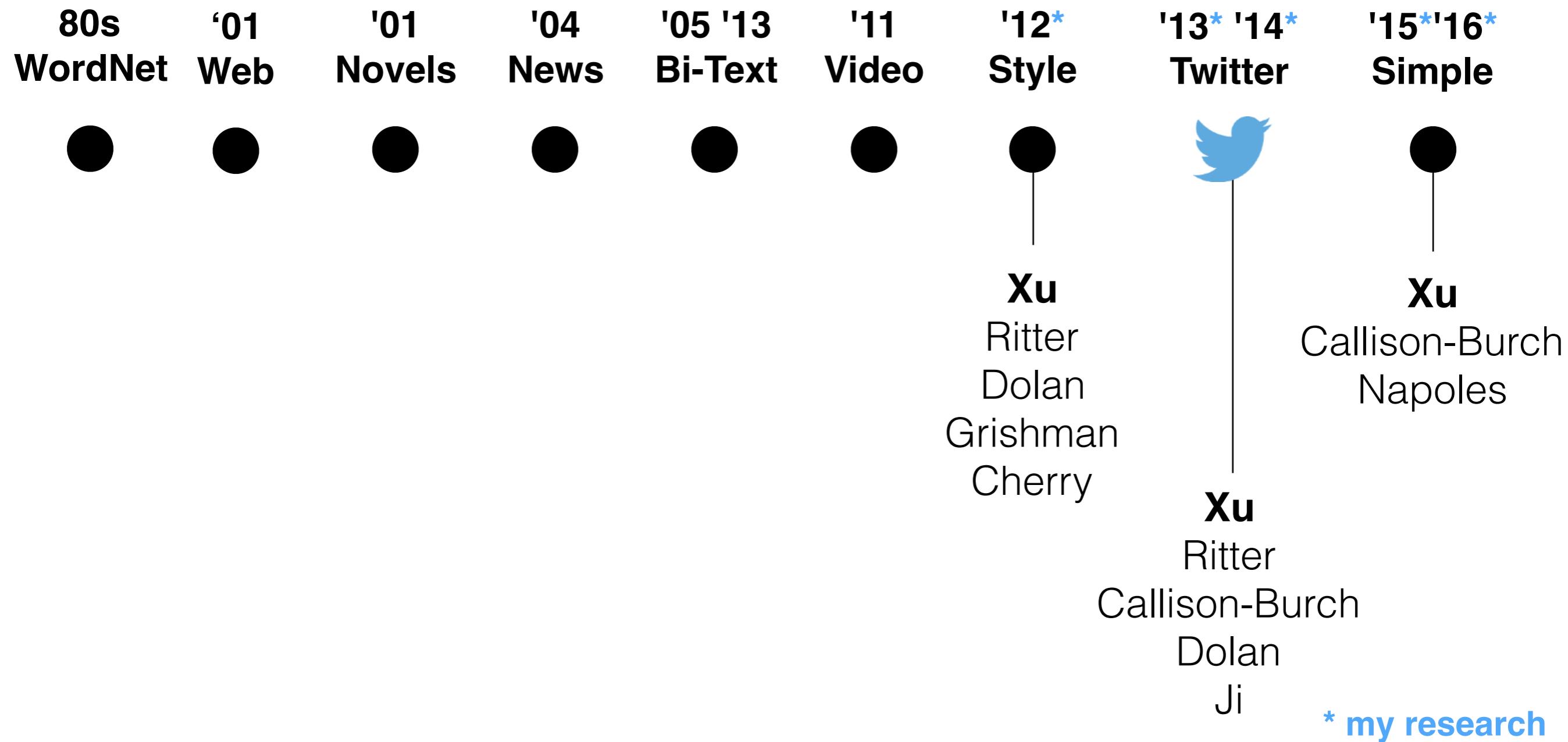
# News (comparable texts)

Dolan, Quirk, and Brockett (2004) extract sentential paraphrases from newspaper articles published on the same topic and date:

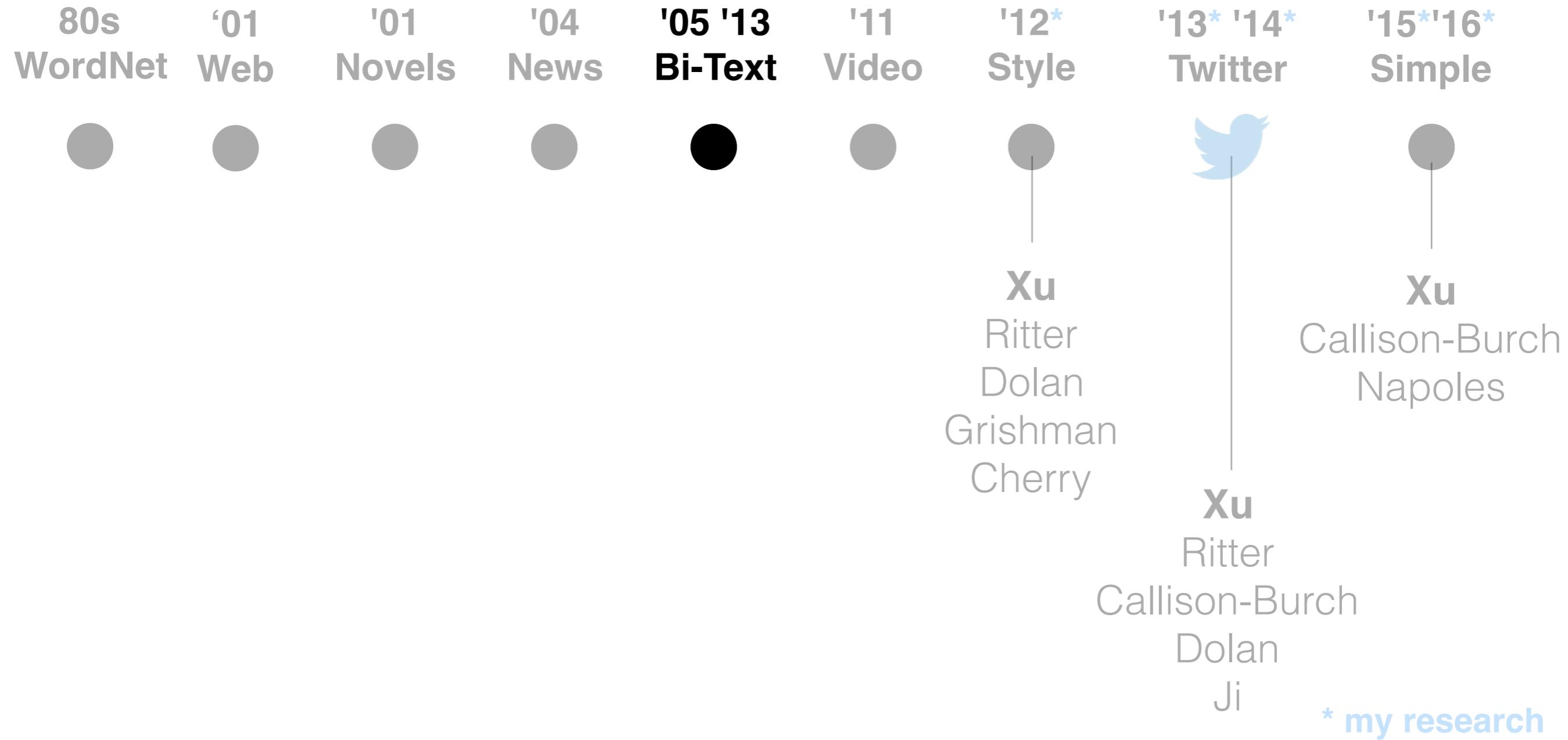
On its way to an extended mission at Saturn, the Cassini probe on Friday makes its closest rendezvous with Saturn's dark moon Phoebe.

The Cassini spacecraft, which is en route to Saturn, is about to make a close pass of the ringed planet's mysterious moon Phoebe.

# Paraphrase Research



# Paraphrase Research



# Data-Driven Paraphrasing

'01  
Novels

Monolingual parallel: English – English

Source: Chris Callison-Burch

Nitin Madnani and Bonnie Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods.  
In Computational Linguistics (2010)

# Data-Driven Paraphrasing

'01 Novels	Monolingual parallel:	English – English
'01 Web	Plain monolingual:	English

Source: Chris Callison-Burch

Nitin Madnani and Bonnie Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods.  
In Computational Linguistics (2010)

# Data-Driven Paraphrasing

'01 Novels	Monolingual parallel:	English – English
'01 Web	Plain monolingual:	English
'04 News	Monolingual comparable:	English ~ English

Source: Chris Callison-Burch

Nitin Madnani and Bonnie Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods.  
In Computational Linguistics (2010)

# Data-Driven Paraphrasing

Monolingual parallel: English – English

Plain monolingual: English

Monolingual comparable: English ~ English

Source: Chris Callison-Burch

Nitin Madnani and Bonnie Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods.  
In Computational Linguistics (2010)

# Data-Driven Paraphrasing

Monolingual parallel: English – English

Plain monolingual: English

Monolingual comparable: English ~ English

Bilingual parallel: English – French

Source: Chris Callison-Burch

Nitin Madnani and Bonnie Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods.  
In Computational Linguistics (2010)

# Data-Driven Paraphrasing

Monolingual parallel: English – English

Plain monolingual: English

Monolingual comparable: English ~ English

Bilingual parallel: English – French

Source: Chris Callison-Burch

Nitin Madnani and Bonnie Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods.  
In Computational Linguistics (2010)

# Paraphrasing & Translation

Translation is re-writing a text using words in a different language.

Paraphrasing is translation into the same language.

# Bilingual Data

Sentence-aligned parallel corpora in English and any foreign language

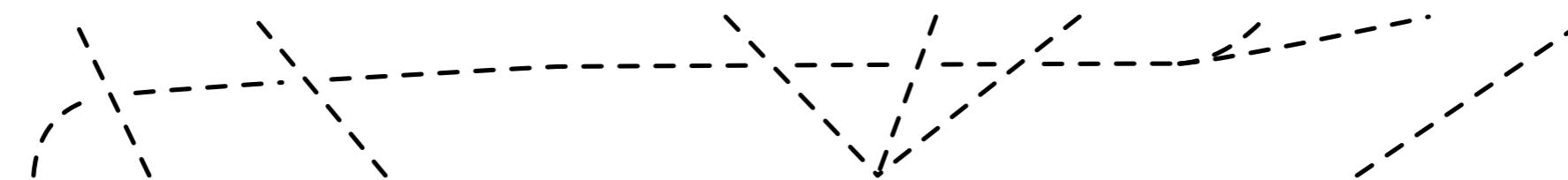
Available in large quantities

Strong meaning equivalence signal

... but different languages.

# Bilingual Pivoting

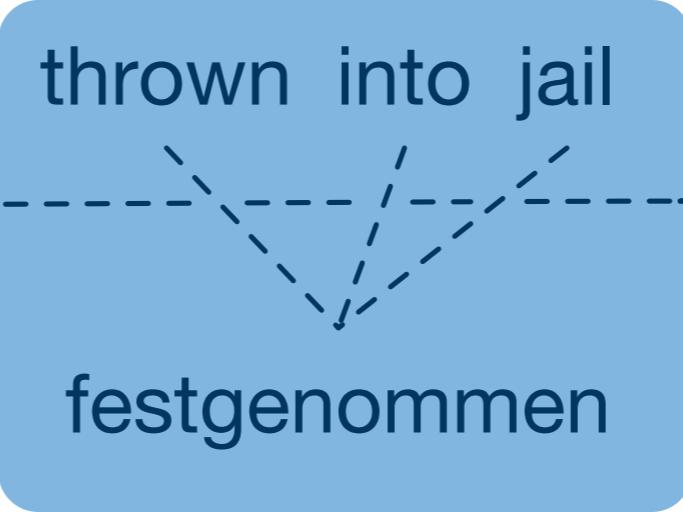
... 5 farmers were thrown into jail in Ireland ...



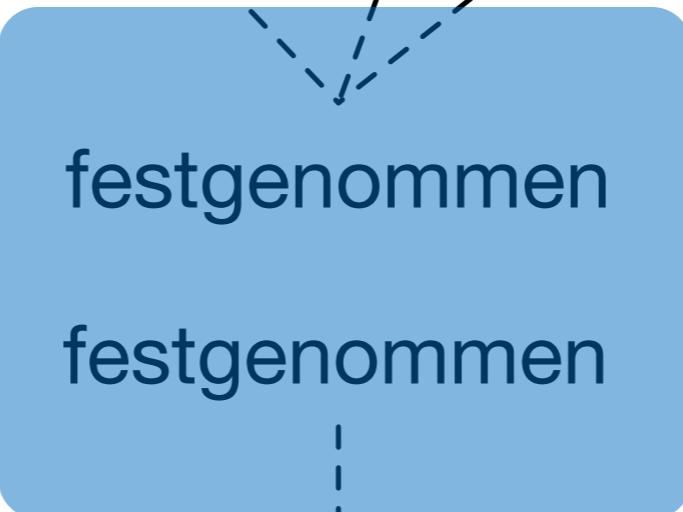
... fünf Landwirte festgenommen , weil ...

# Bilingual Pivoting

... 5 farmers were thrown into jail in Ireland ...  
... fünf Landwirte festgenommen, weil ...



# Bilingual Pivoting

... 5 farmers were thrown into jail in Ireland ...  
... fünf Landwirte  
... oder wurden  
... or have been        
festgenommen  
festgenommen  
imprisoned  
, weil ...  
, gefoltert ...  
, tortured ...

# Bilingual Pivoting

... 5 farmers were thrown into jail in Ireland ...

... fünf Landwirte festgenommen , weil ...

... oder wurden

... or have been

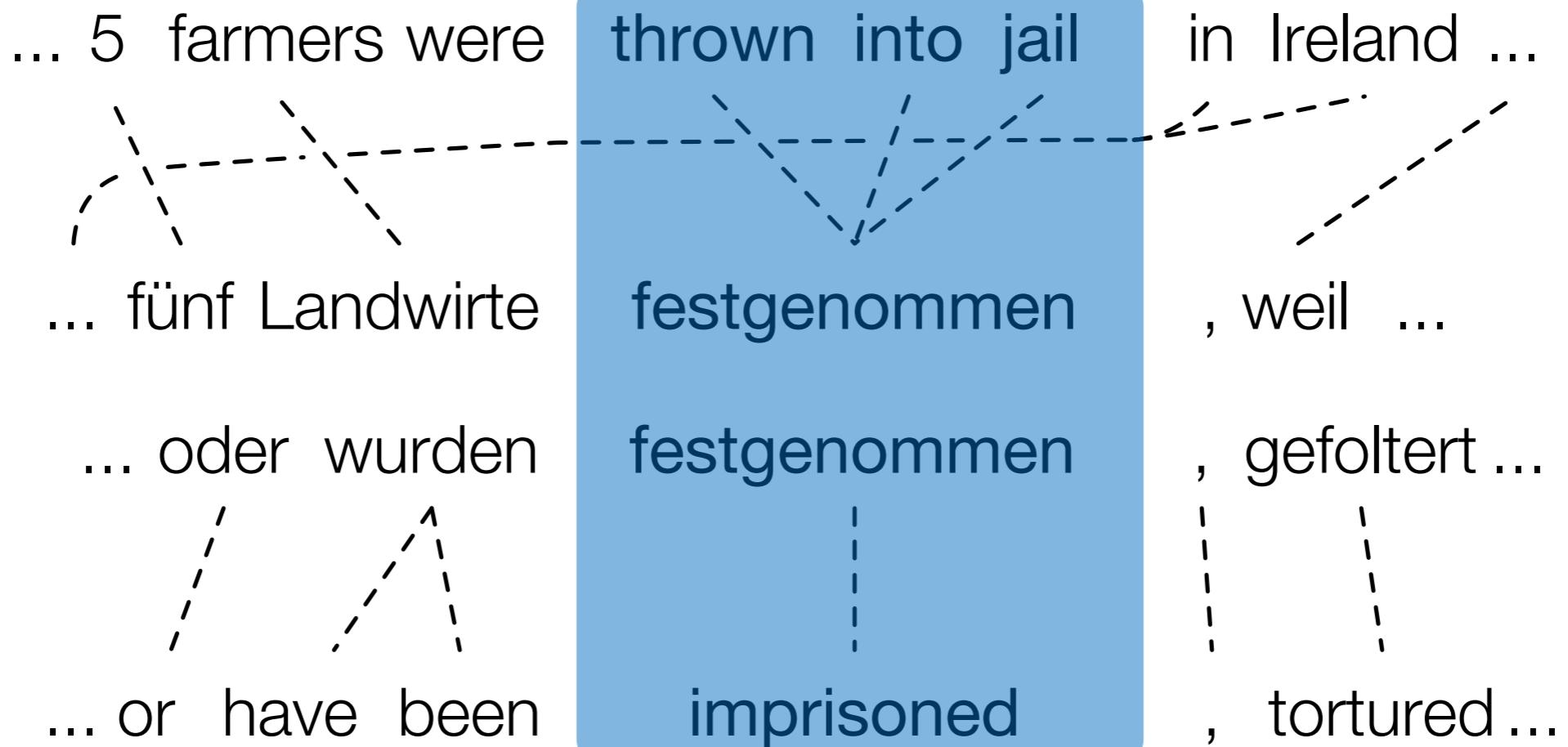
festgenommen

imprisoned

, gefoltert ...

, tortured ...

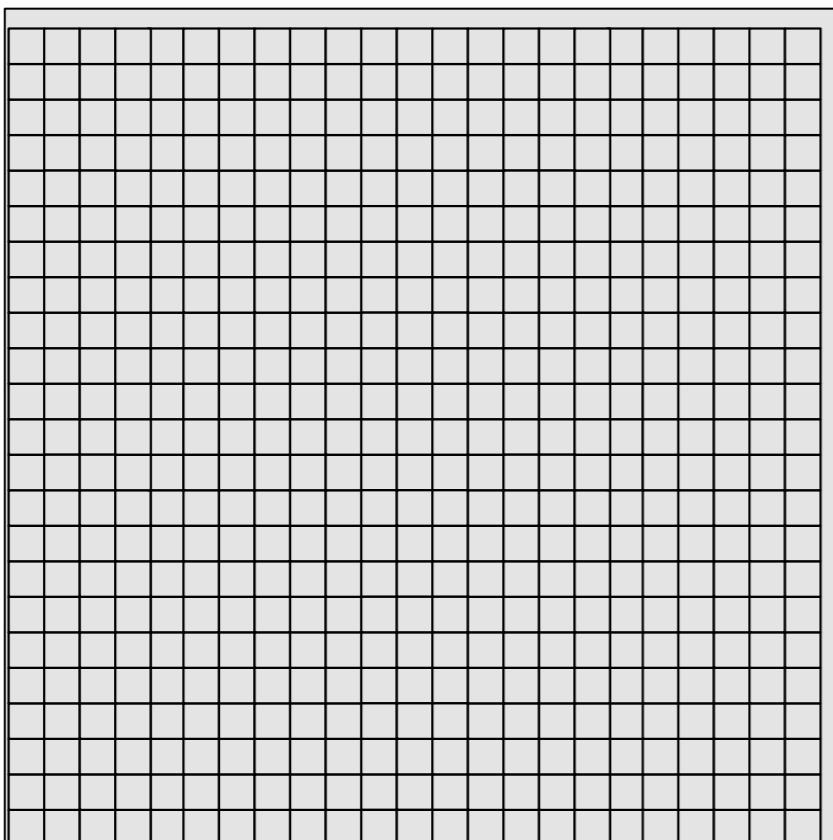
# Bilingual Pivoting



Large and diverse

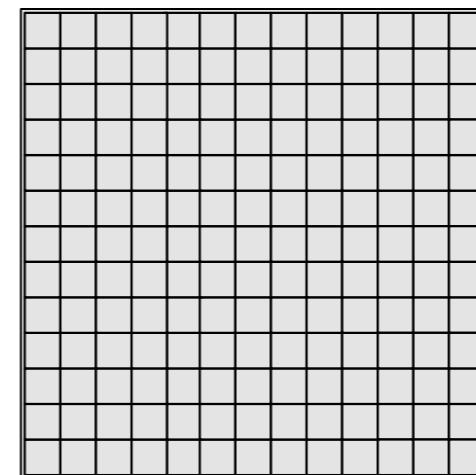
# Bilingual Data Sets

1000M



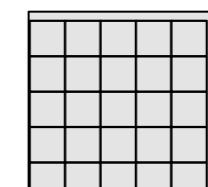
French-English  
 $10^9$  word webcrawl

2 languages @  
250M each



DARPA  
GALE Program

21 languages @  
50-80M each



European  
Parliament

Wide range of **Paraphrases**

**thrown into jail**

Wide range of **Paraphrases**

thrown into jail

arrested

detained

imprisoned

incarcerated

jailed

locked up

taken into custody

thrown into prison

Wide range of

# Paraphrases

thrown into jail

arrested

be thrown in prison

detained

been thrown into jail

imprisoned

being arrested

incarcerated

in jail

jailed

in prison

locked up

put in prison for

taken into custody

were thrown into jail

thrown into prison who are held in detention

Wide range of

# Paraphrases

thrown into jail

arrested

be thrown in prison

arrest

detained

been thrown into jail

cases

imprisoned

being arrested

custody

incarcerated

in jail

maltreated

jailed

in prison

owners

locked up

put in prison for

protection

taken into custody

were thrown into jail

thrown

thrown into prison who are held in detention

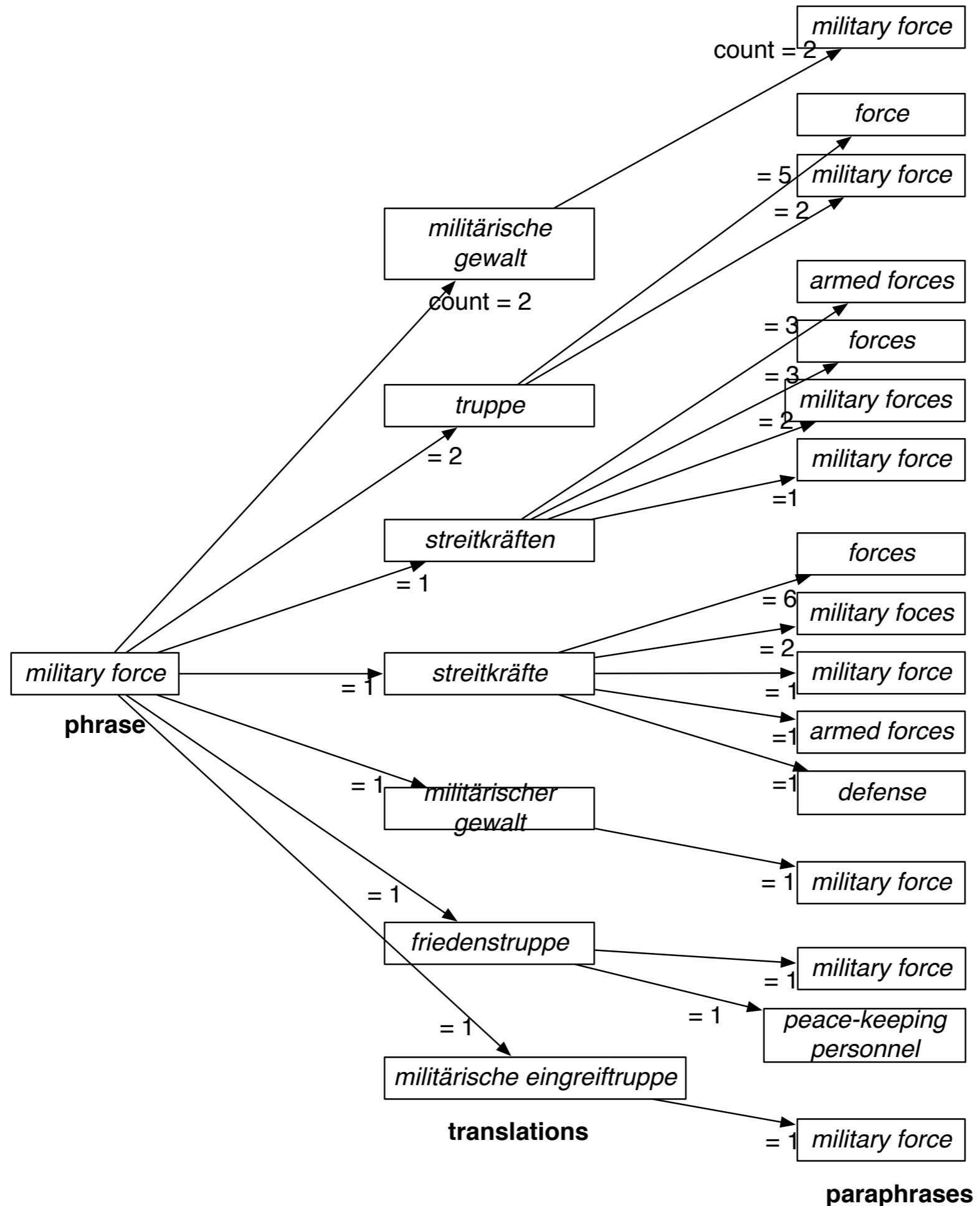
# Paraphrase Probability

$$\begin{aligned} p(e_2|e_1) &= \sum_f p(e_2, f|e_1) \\ &= \sum_f p(e_2|f, e_1)p(f|e_1) \\ &\approx \sum_f p(e_2|f)p(f|e_1) \end{aligned}$$

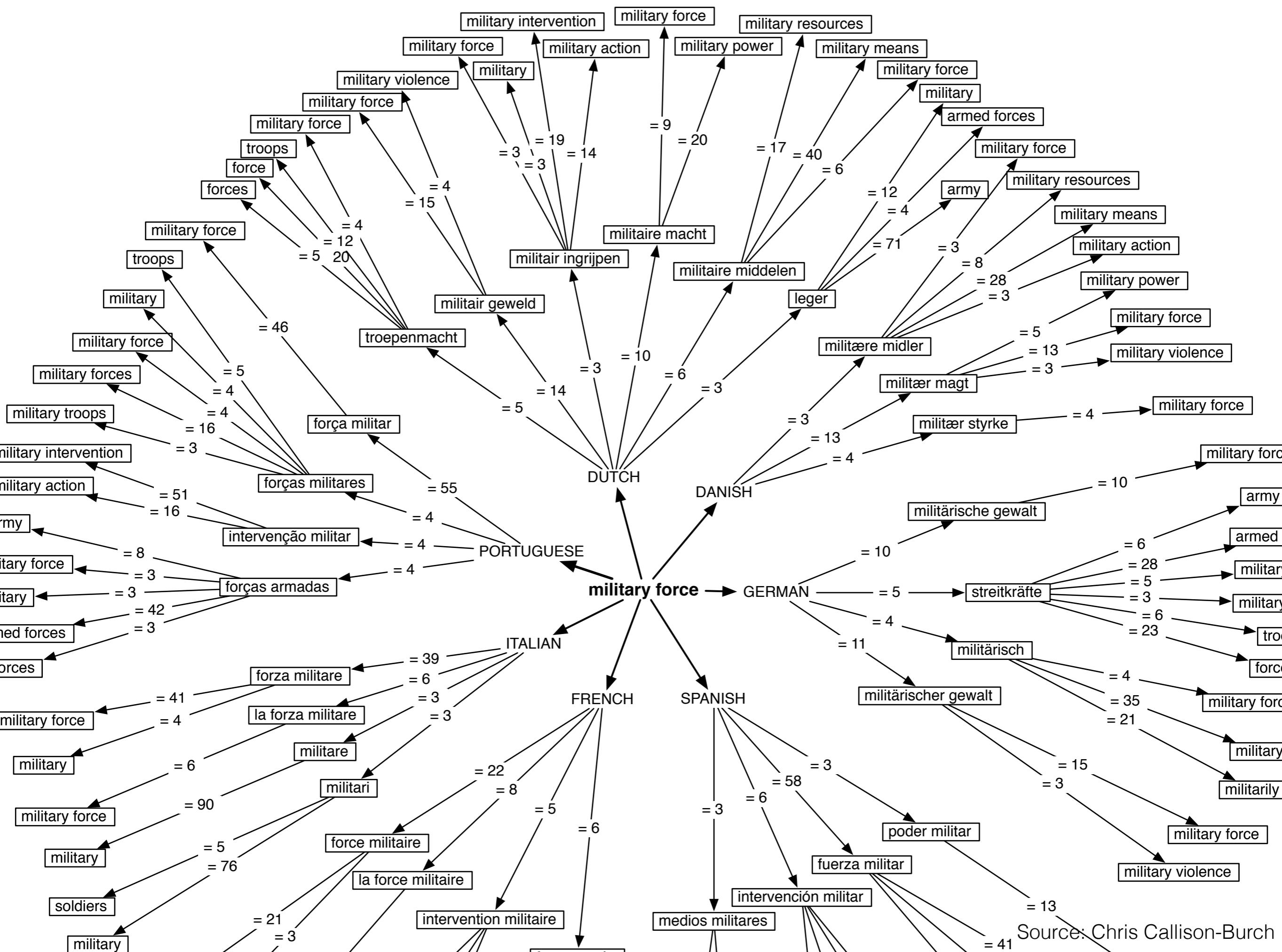
Source: Chris Callison-Burch

Colin Bannard and Chris Callison-Burch. Paraphrasing with Bilingual Parallel Corpora. ACL 2005.









Source: Chris Callison-Burch

# Syntactic Constraints

thrown into jail

arrested

be thrown in prison

arrest

detained

been thrown into jail

eases

imprisoned

being arrested

custody

incarcerated

in jail

maltreated

jailed

in prison

owners

locked up

put in prison for

protection

taken into custody

were thrown into jail

thrown

thrown into prison

who are held in detention

Source: Chris Callison-Burch

# Distributional Similarity

Idea: similar words occur in similar contexts.

Characterize words by their contexts

Contexts represented by co-occurrence vectors, similarity quantified by cosine

“Are these paraphrases substitutable?”

# Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

# Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

cup

↔

mug

# Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

..sip from a cup of cocoa..

..a cup of coffee.

cup

..sip from a mug of cocoa..

..a mug of coffee.



mug

# Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases



cup



mug

# Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases



cup



mug

the king's speech

↔ His Majesty's address

# Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases



cup



mug

..anxiously awaiting the king's  
speech..

the king's speech

..anxiously awaiting His  
Majesty's address..

His Majesty's address

# Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases



cup



mug



the king's speech



His Majesty's address

# Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases



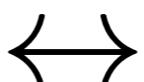
cup



mug

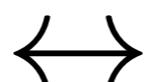


the king's speech



His Majesty's address

one JJ instance of NP



a JJ case of NP

# Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases



cup



mug



the king's speech



His Majesty's address



one JJ instance of NP



a JJ case of NP

# Syntactic Paraphrase Similarity

NN 's NP in the long term

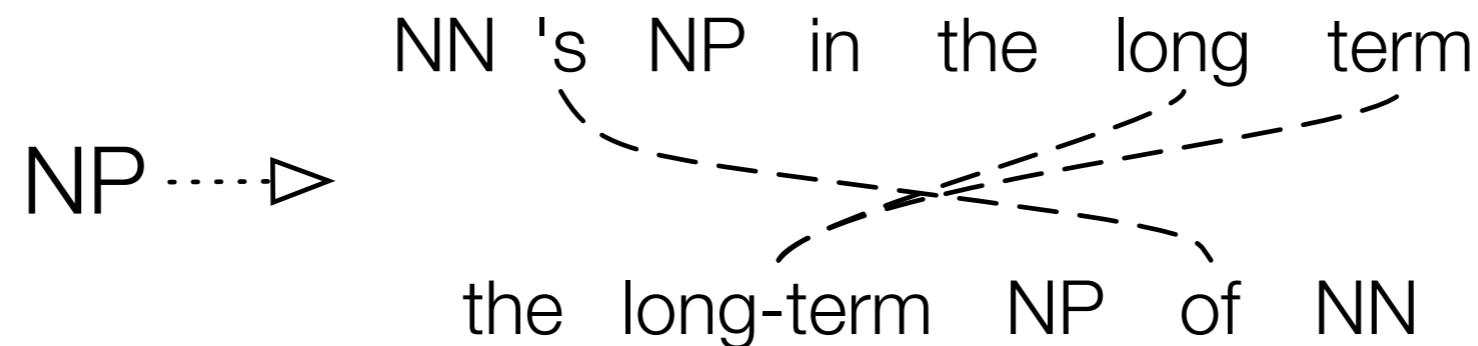
NP .....>

the long-term NP of NN

Source: Chris Callison-Burch

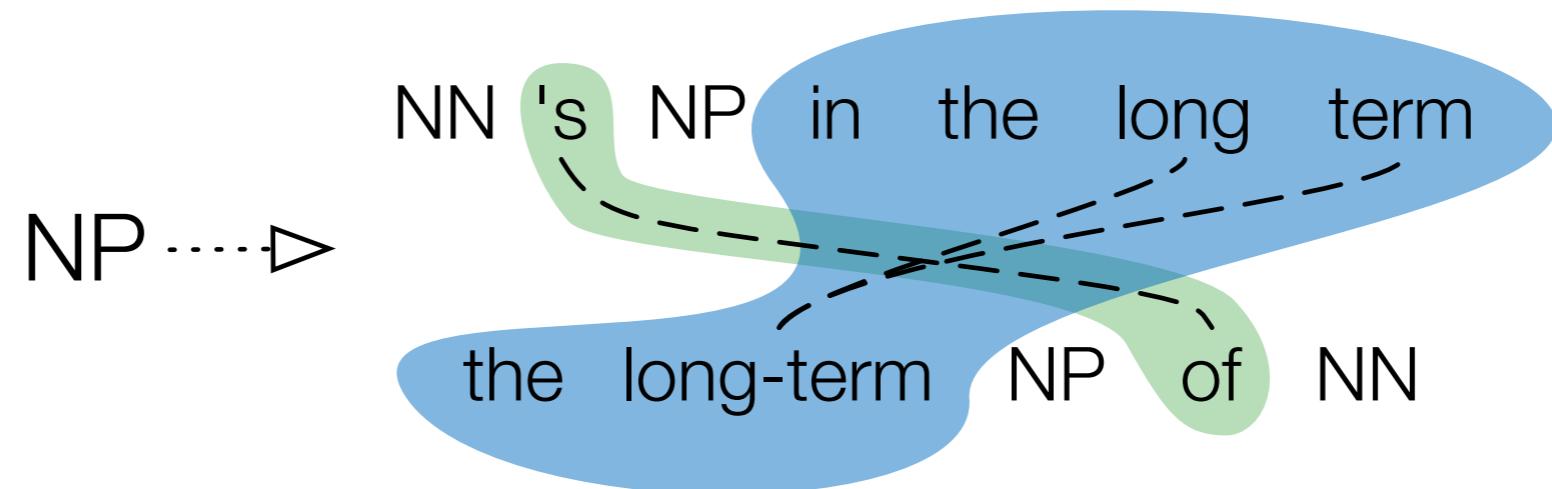
Juri Ganitkevitch, Ben Van Durme and Chris Callison-Burch. Monolingual Distributional Similarity for Text-to-Text Generation. \*SEM 2012.

# Syntactic Paraphrase Similarity



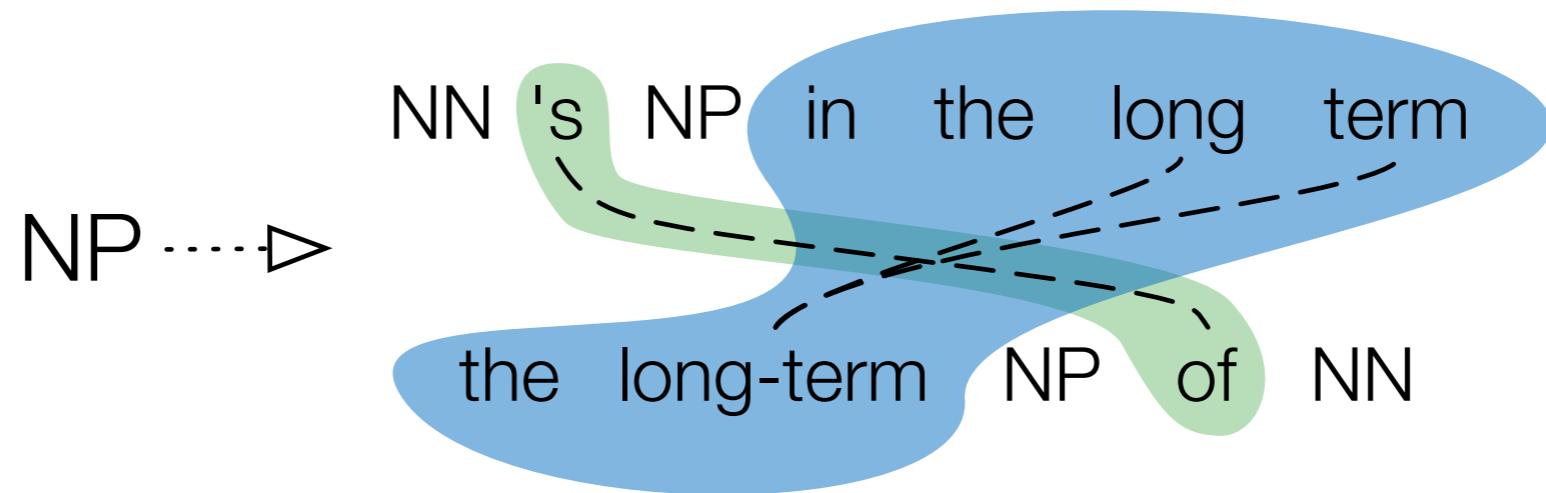
Source: Chris Callison-Burch

# Syntactic Paraphrase Similarity



Source: Chris Callison-Burch

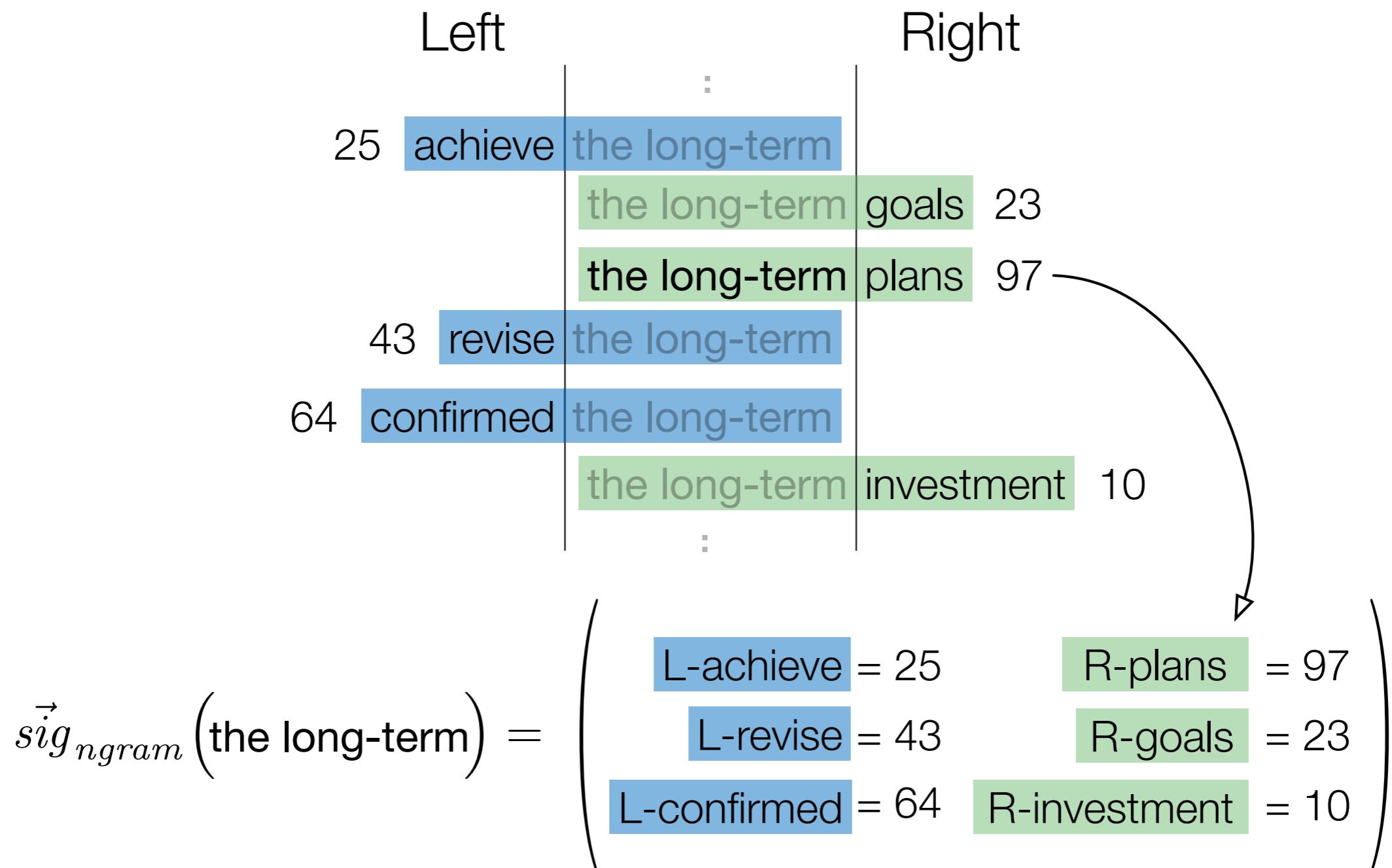
# Syntactic Paraphrase Similarity



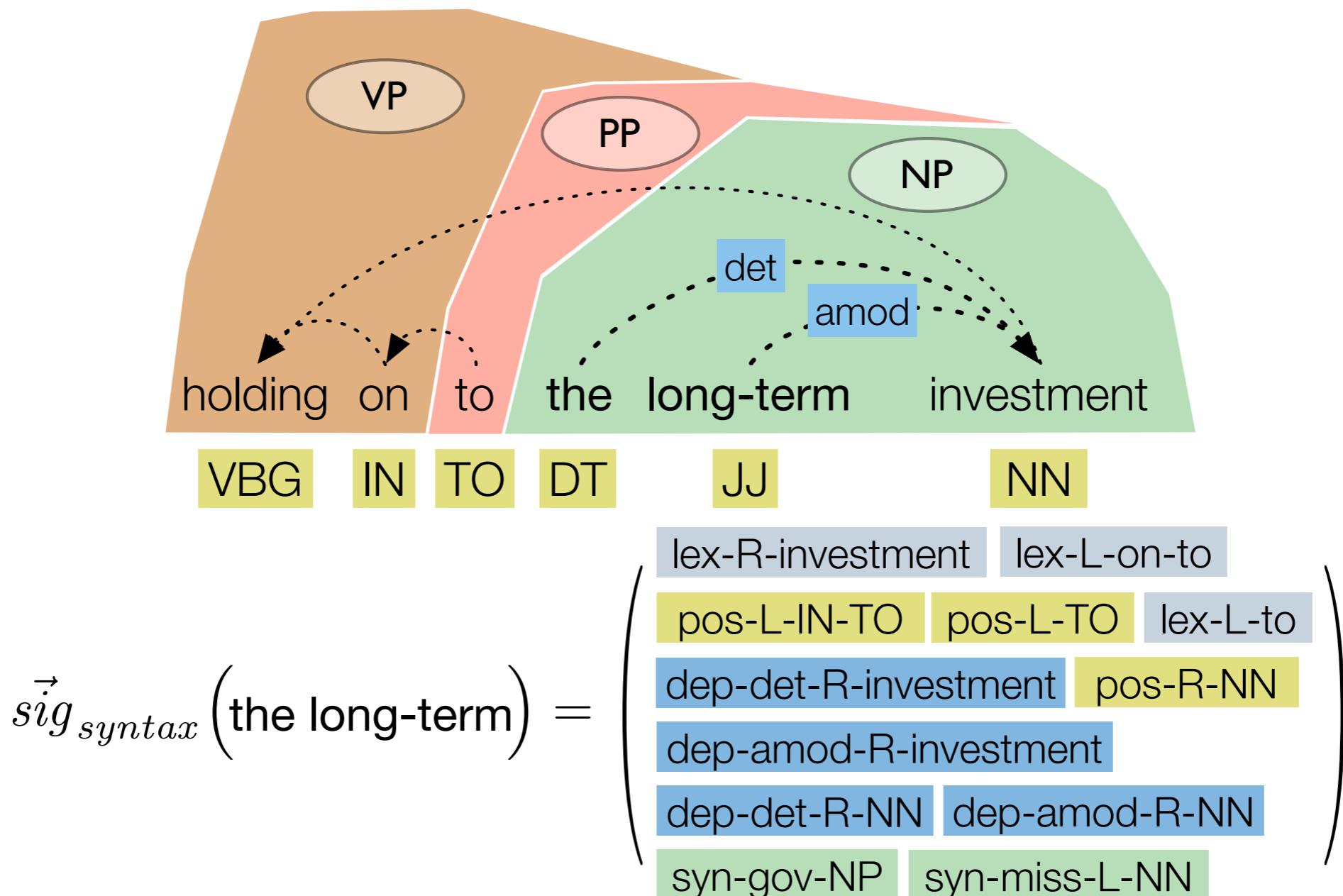
$$sim(\mathbf{r}) = \frac{1}{2} \left( sim\left( \text{the long-term} \atop \text{in the long term} \right) + sim\left( \text{'s} \atop \text{of} \right) \right)$$

Source: Chris Callison-Burch

# $n$ -gram Context



# Syntactic Context


$$\vec{sig}_{syntax}(\text{the long-term}) =$$

lex-R-investment	lex-L-on-to	
pos-L-IN-TO	pos-L-TO	lex-L-to
dep-det-R-investment	pos-R-NN	
dep-amod-R-investment		
dep-det-R-NN	dep-amod-R-NN	
syn-gov-NP	syn-miss-L-NN	

# Large Monolingual Data Sets

Google n-grams

Collection of 1 trillion tokens with counts

Based on vast amounts of text

Annotated Gigaword (AKBC-WEKEX '12)

Collection of 4 billion words, parsed and tagged

Source: Chris Callison-Burch

# PPDB: The Paraphrase Database

- A huge collection of paraphrases
- Extracted from 106 million sentence pairs, 2 billion English words, 22 pivot languages

	Paraphrases
Lexical	7.6 M
Phrasal	68.4 M
Syntactic	93.6 M
Total	169.6 M

# PPDB: The Paraphrase Database

Language	Code	Number of Paraphrases			
		Lexical	Phrasal	Syntactic	Total
Arabic	Ara	119.7M	45.1M	20.1M	185.7M
Bulgarian	Bul	1.3M	1.4M	1.2M	3.9M
Czech	Ces	7.3M	2.7M	2.6	12.1M
German	Deu	7.9M	15.4M	4.9M	28.3M
Greek	Ell	5.4M	9.4M	7.4M	22.3M
Estonian	Est	7.9M	1.0M	0.4M	9.2M
Finnish	Fin	41.4M	4.9M	2.3M	48.6M
French	Fra	78.8M	254.2M	170.5M	503.5M
Hungarian	Hun	3.8M	1.3M	0.2M	5.3M
Italian	Ita	8.2M	17.9M	9.7M	35.8M
Lithuanian	Lit	8.7M	1.5M	0.8M	11.0M
Latvian	Lav	5.5M	1.4M	1.0M	7.9M
Dutch	Nld	6.1M	15.3M	4.5M	25.9M
Polish	Pol	6.5M	2.2M	1.4M	10.1M
Portuguese	Por	7.0M	17.0M	9.0M	33.0M
Romanian	Ron	1.5M	1.8M	1.1M	4.5M
Russian	Rus	81M	46M	16M	144.4M
Slovak	Slk	4.8M	1.8M	1.7M	8.2M
Slovenian	Slv	3.6M	1.6M	1.4M	6.7M
Swedish	Swe	6.2M	10.3M	10.3M	26.8M
Chinese	Zho	52.5M	46.0M	8.9M	107.4M

Source: Chris Callison-Burch



huge amount

English ▾

Go



Download PPDB

Result for **huge amount**

129 search results

1

**enormous amount**

Noun phrase missing determiner on the left



0



0

2

**tremendous amount**

Noun phrase missing determiner on the left



0



0

3

**huge sum**

Noun phrase missing determiner on the left



0



0

4

**enormous number**

Noun phrase missing determiner on the left



0



0

5

**huge number**

Noun phrase missing determiner on the left



0



0

6

**awful lot**

Noun phrase missing determiner on the left



0



0

7

**massive amount**

0



0



PPDB

paraphrase.org/#/download

Reader

Cloud

Download PPDB

Search here...

English

Go

Language

English

All

Lexical

One-To-Many

Phrasal

Syntactic

Select size of pack

S Size

M Size

L Size

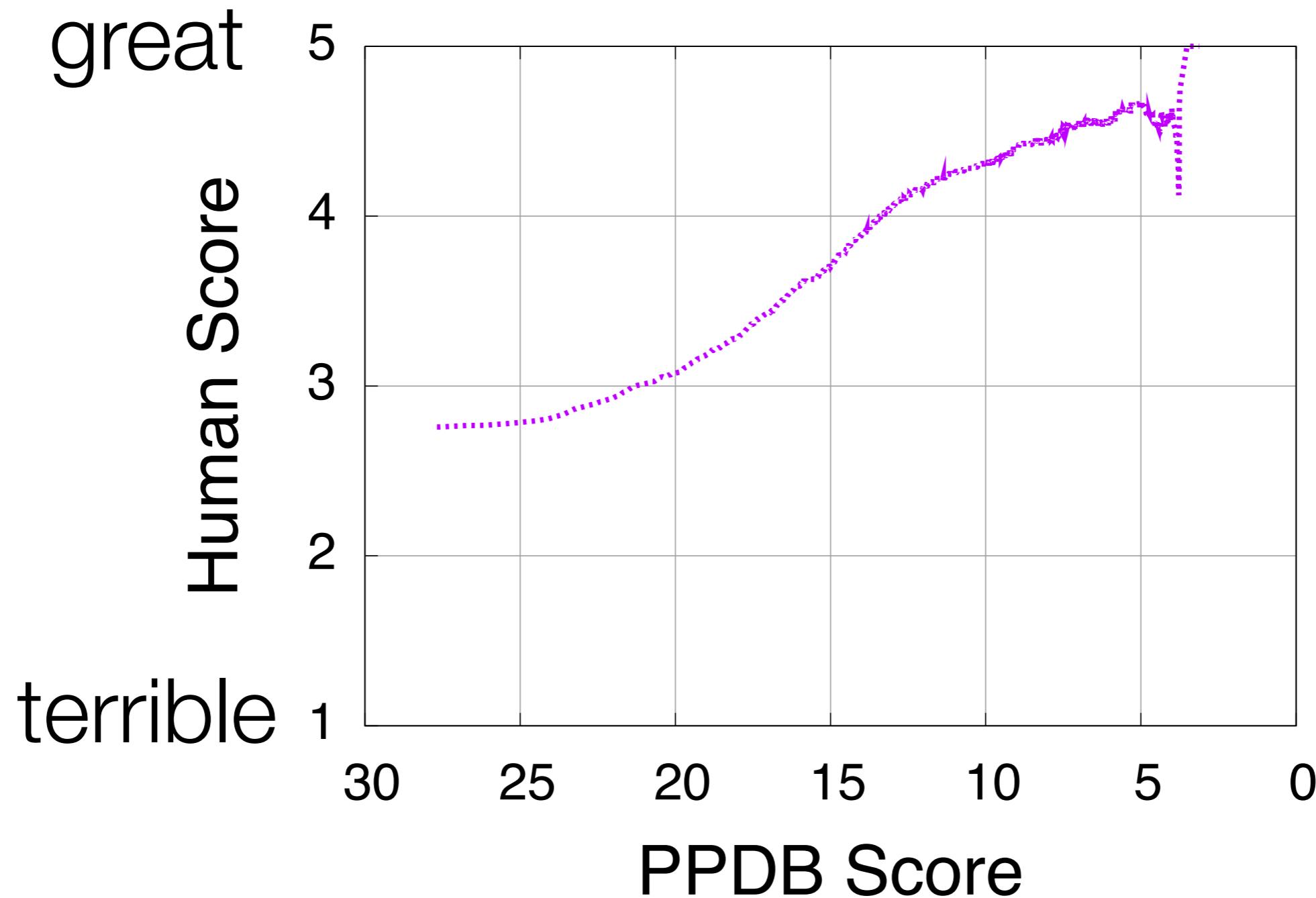
XL Size

XXL Size

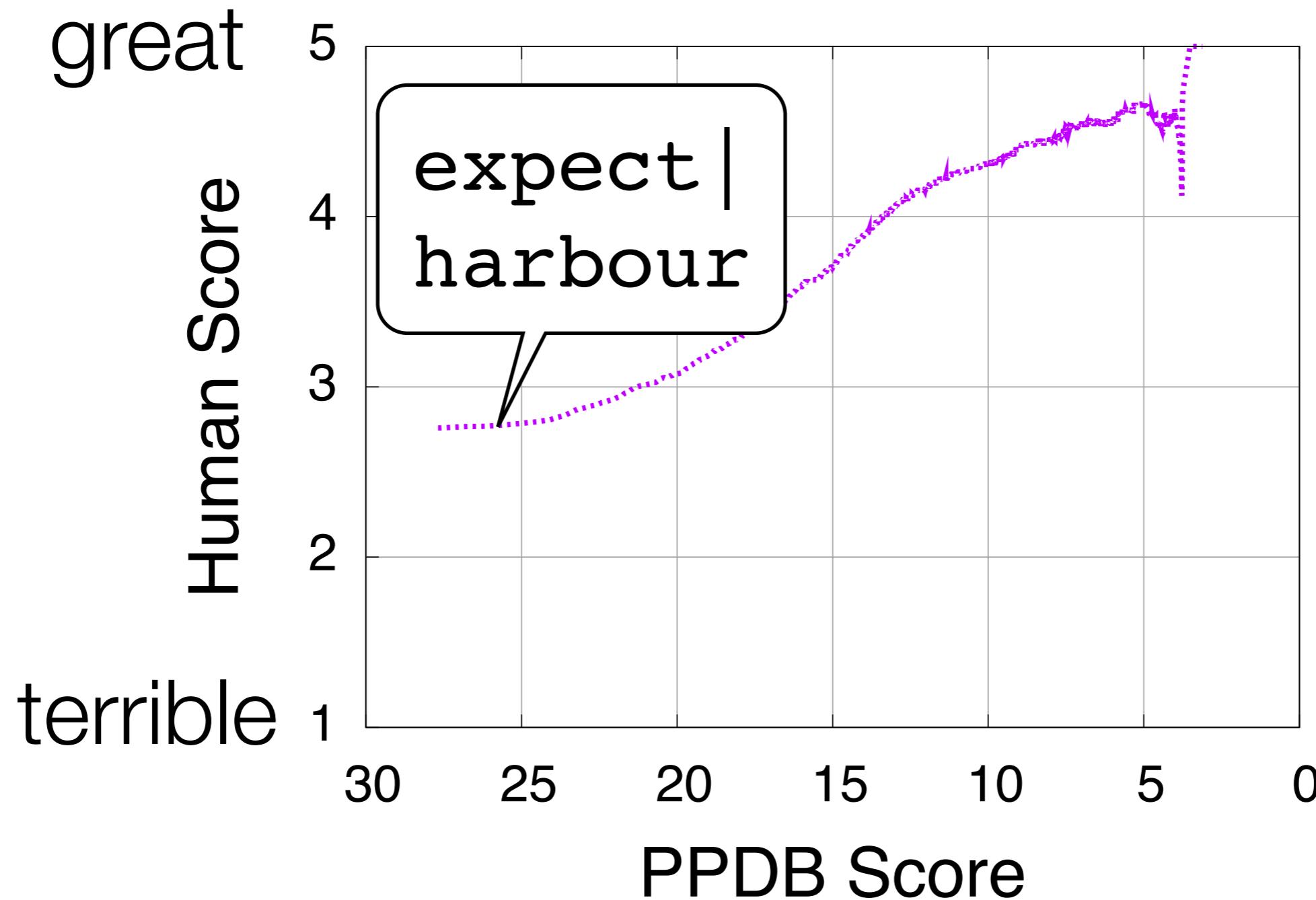
XXXL Size

💡

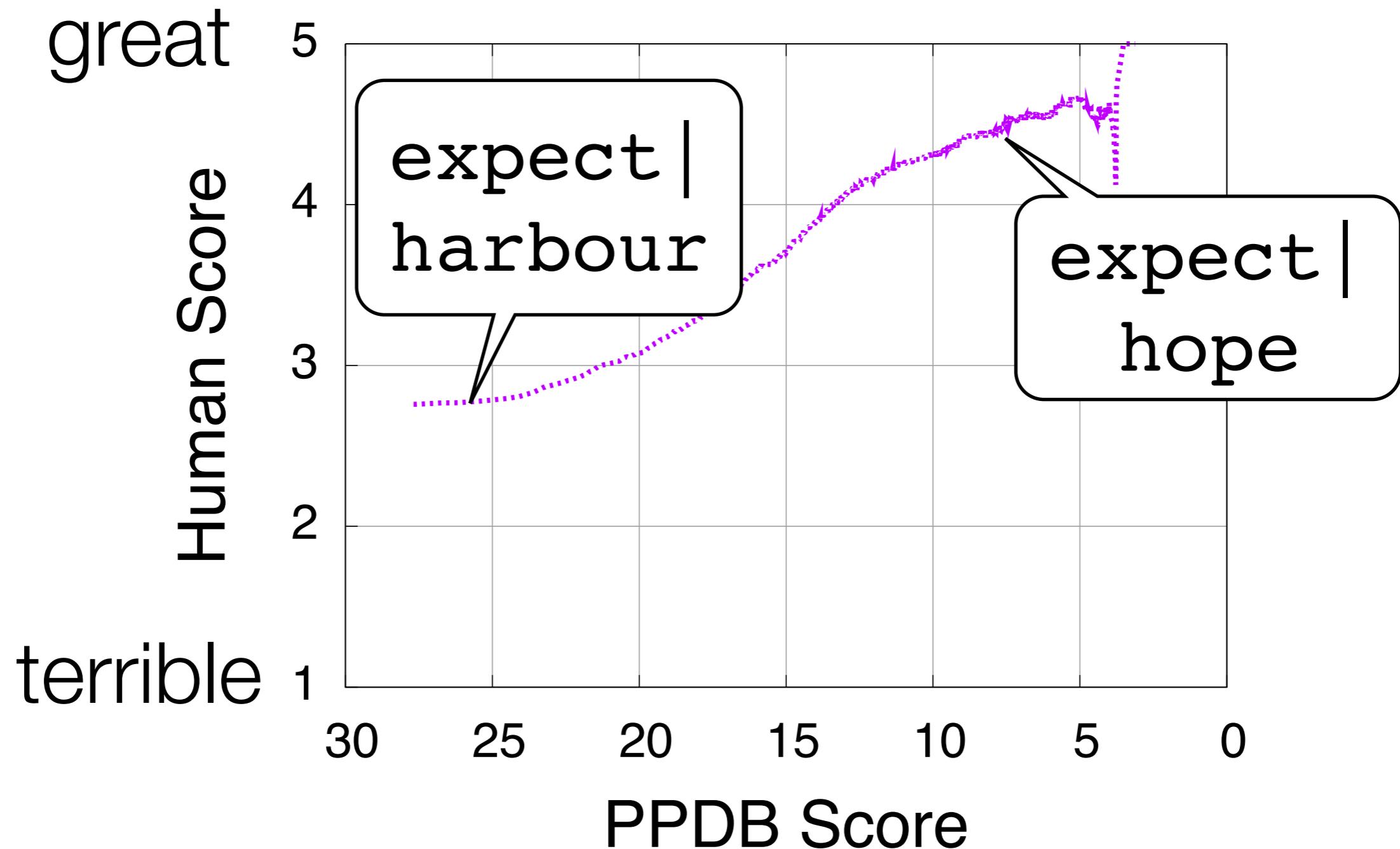
# Do the Scores Work?



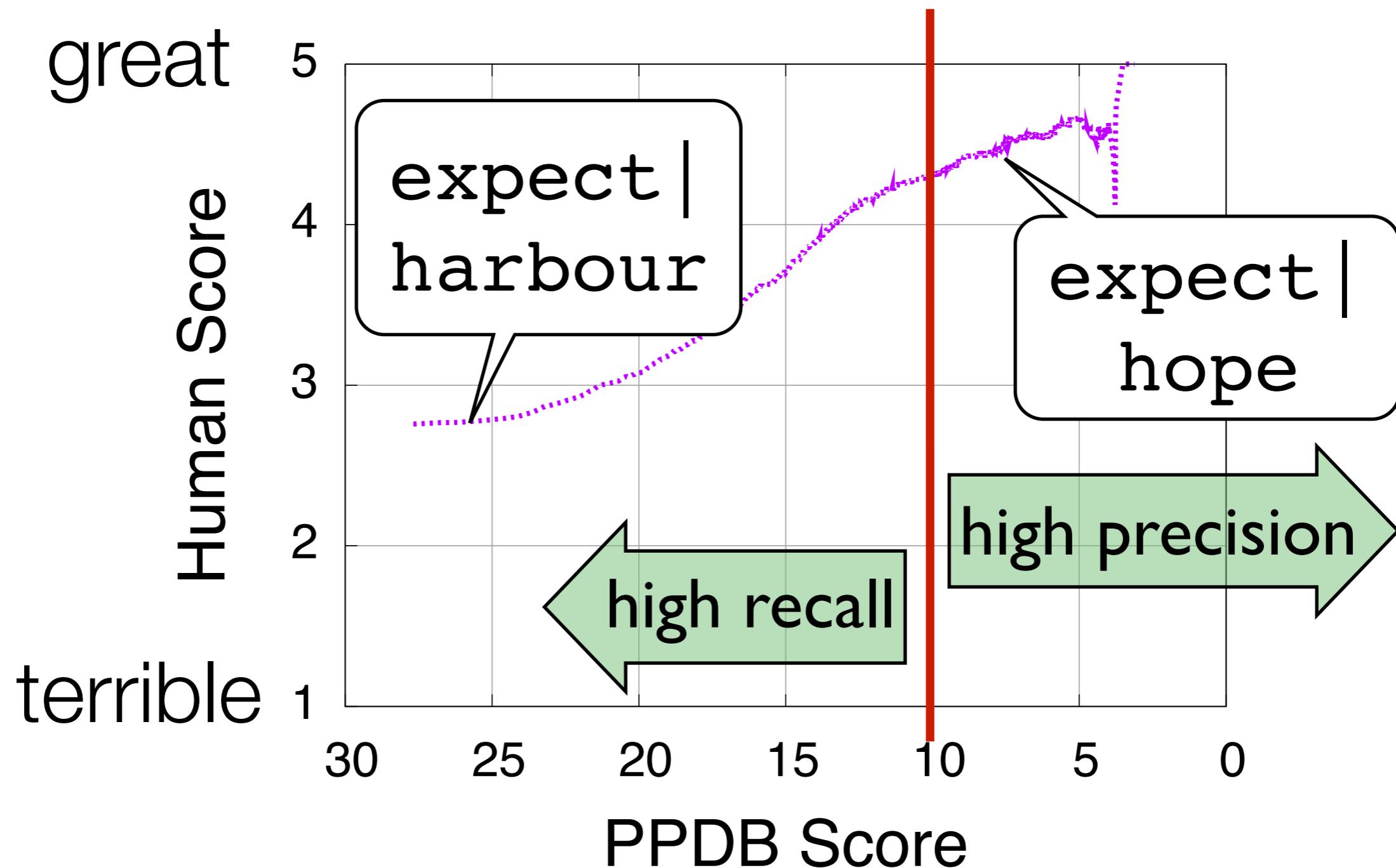
# Do the Scores Work?



# Do the Scores Work?



# Do the Scores Work?



# Fun PPDB Examples

**P A R E N T A L  
A D V I S O R Y  
E X P L I C I T   C O N T E N T**

# Fun PPDB Examples

munchies ||| hungry

hustle ||| scam

sexiest ||| hottest

dummies ||| losers

sheeit ||| dammit

abso-fucking-lutely ||| indeed

# Pivoting w/ Neural MT

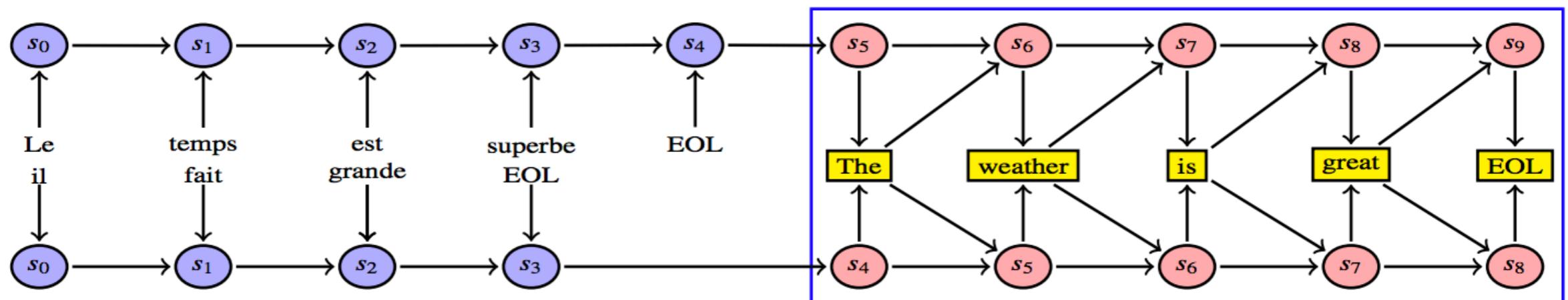


Figure 1: Late-weighted combination: two pivot sentences are simultaneously translated to one target sentence. Blue circles indicate the encoders, which individually encode the two source sentences. After the EOL token is seen, decoding starts (red circles). At each time step the two decoders produce a probability distribution over all words, which are then combined (in the yellow square) using Equation (6). From this combined distribution a word is chosen, which is then given as input to each decoder.

# Pivoting w/ Neural MT

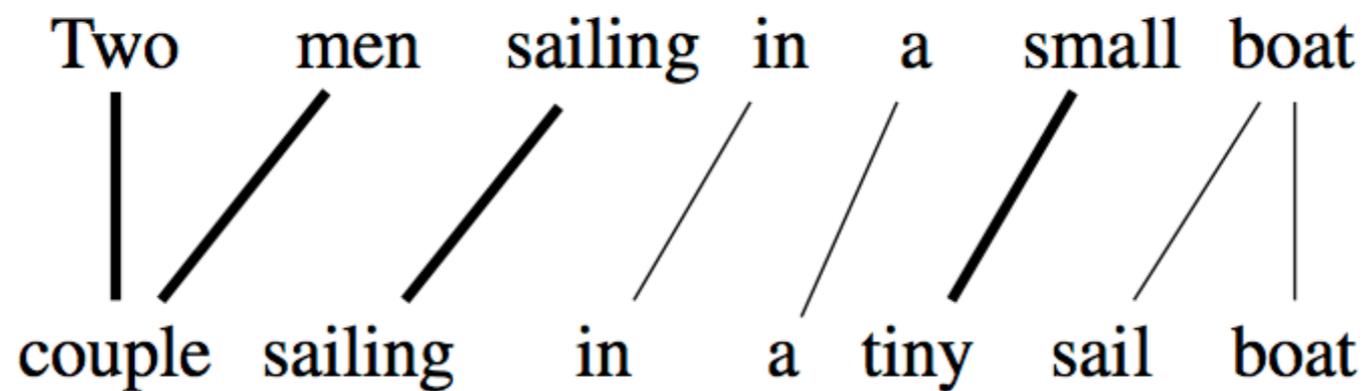


Figure 2: Attention between two sentences. Line thickness indicates the strength of the attention.

$$\alpha(E_2^i, E_1^j, \mathcal{F}) = \sum_F^{\mathcal{F}} (P(E_2|E_1, F) \cdot \sum_m^{T_F} (\alpha_{i,m}^{E_2,F} \cdot \alpha_{m,j}^{F,E_1}))$$

# Improve MT w/ PPDB

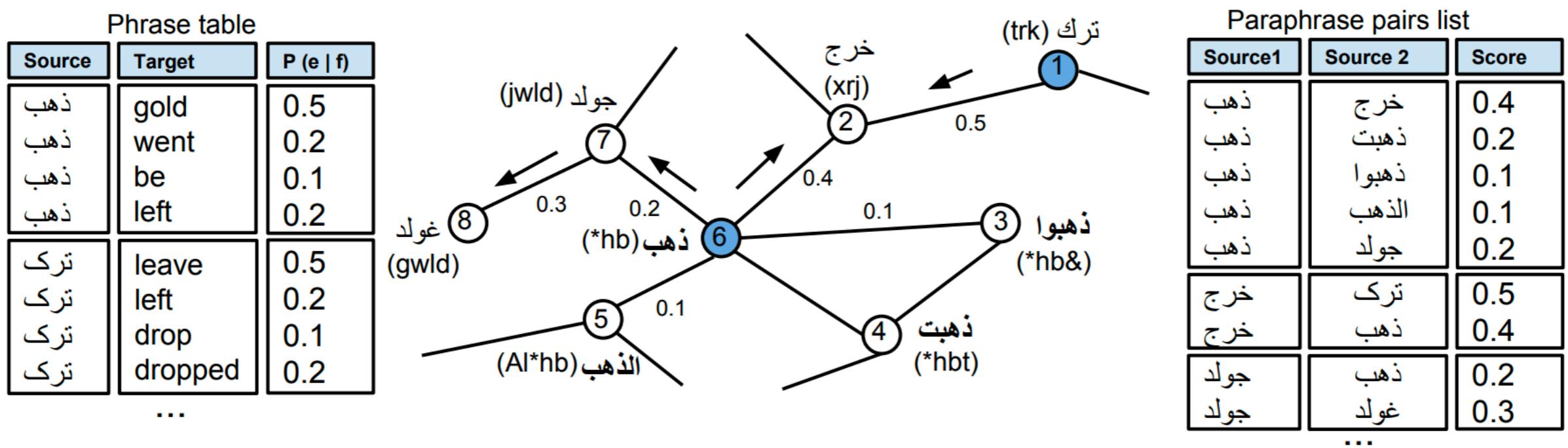


Figure 2: A small sample of the real graph constructed from the Arabic PPDB for Arabic to English translation. Filled nodes (1 and 6) are phrases from the SMT phrase table (unfilled nodes are not). Edge weights are set using a log-linear combination of scores from PPDB. Phrase #6 has different senses ('gold' or 'left'); and it has a paraphrase in phrase #7 for the 'gold' sense and a paraphrase in phrase #2 for the 'left' sense. After propagation, phrase #2 receives translation candidates from phrase #6 and phrase #1 reducing the probability of translation from unrelated senses (like the 'gold' sense). Phrase #8 is a misspelling of phrase #7 and is also captured as a paraphrase. Phrase #6 propagates translation candidates to phrase #8 through phrase #7. Morphological variants of phrase #6 (shown in bold) also receive translation candidates through graph propagation giving translation candidates for morphologically rich OOVs.