Social Media & Text Analysis

lecture 4 - Paraphrase Data Sources

Follow @cocoweixu

CSE 5539-0010 Ohio State University

Instructor: Wei Xu

Website: socialmedia-class.org

Natural Language Processing

Dan Jurafsky

Language Technology

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

making good progress

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 13th 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

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

making good progress

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 13th 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)

Rahul Bhagat and Ed Hovy. "What Is a Paraphrase?". Computational Linguistics (2013)

what is Paraphrase?

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

wealthy	**word**	*rich*

Rahul Bhagat and Ed Hovy. "What Is a Paraphrase?". Computational Linguistics (2013)

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*

Rahul Bhagat and Ed Hovy. "What Is a Paraphrase?". Computational Linguistics (2013)

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 …*

Rahul Bhagat and Ed Hovy. "What Is a Paraphrase?". Computational Linguistics (2013)

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

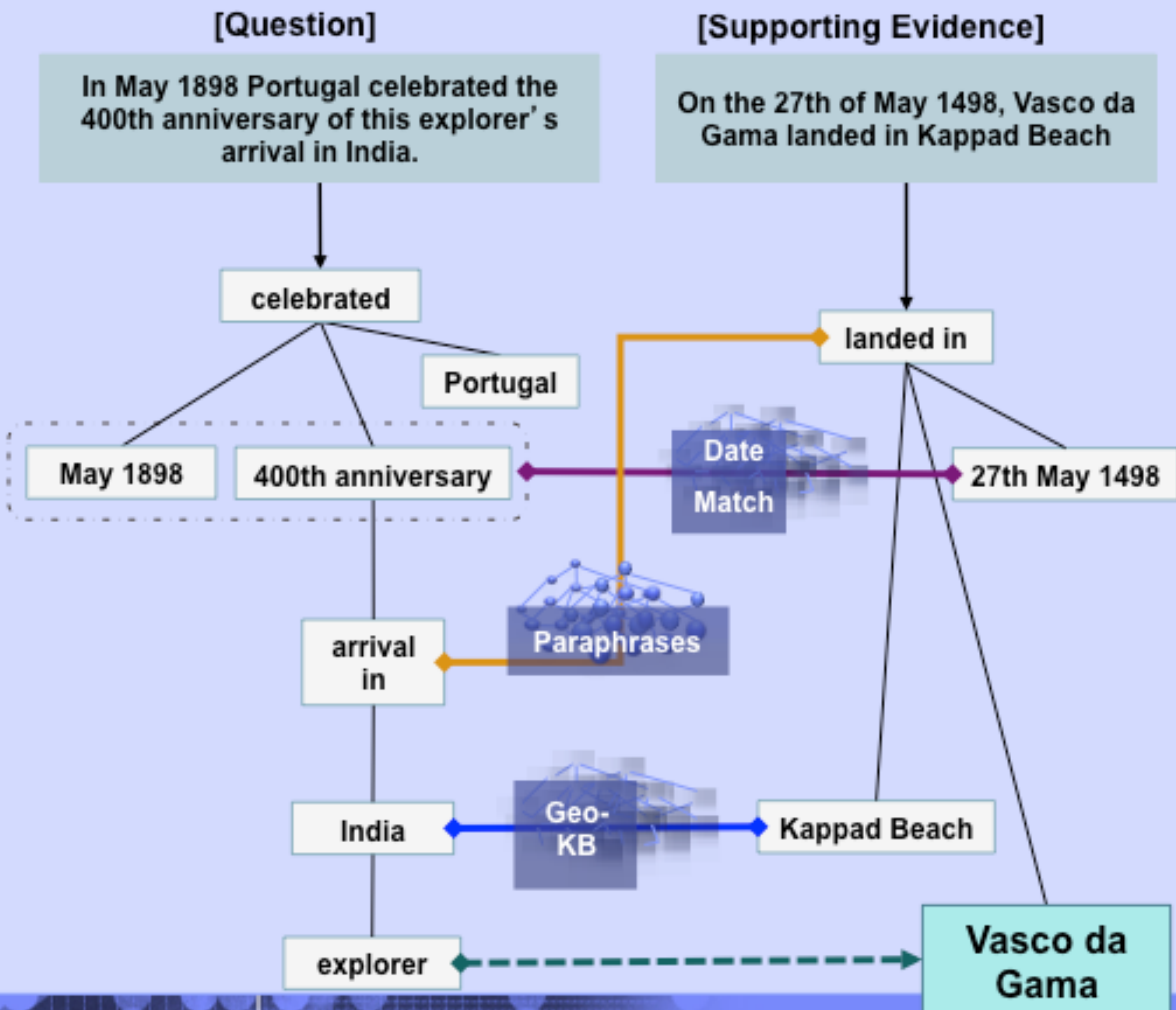
e.g. Question Answering

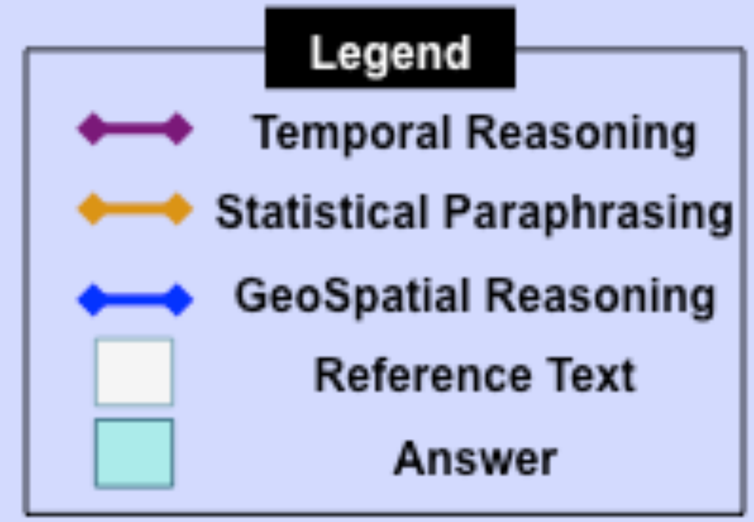
Who is the CEO stepping down from Boeing?

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



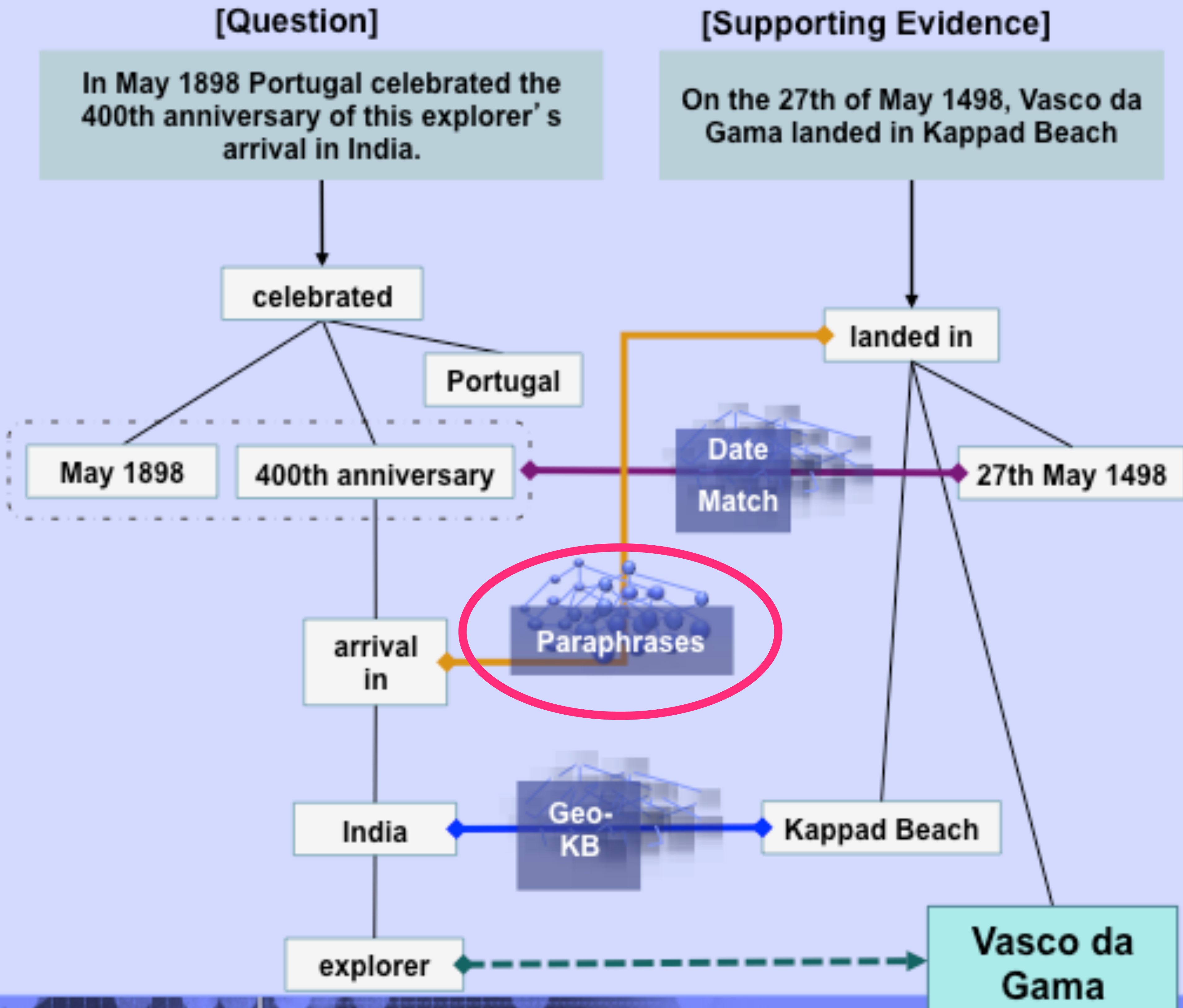


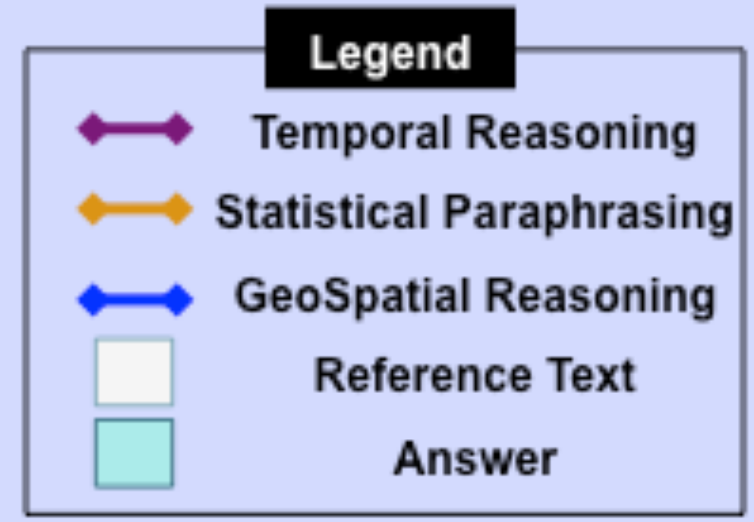
Stronger evidence can be much harder to find and score…

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

(courtesy: Salim Roukos)

Watson leverages multiple algorithms to perform deeper analysis





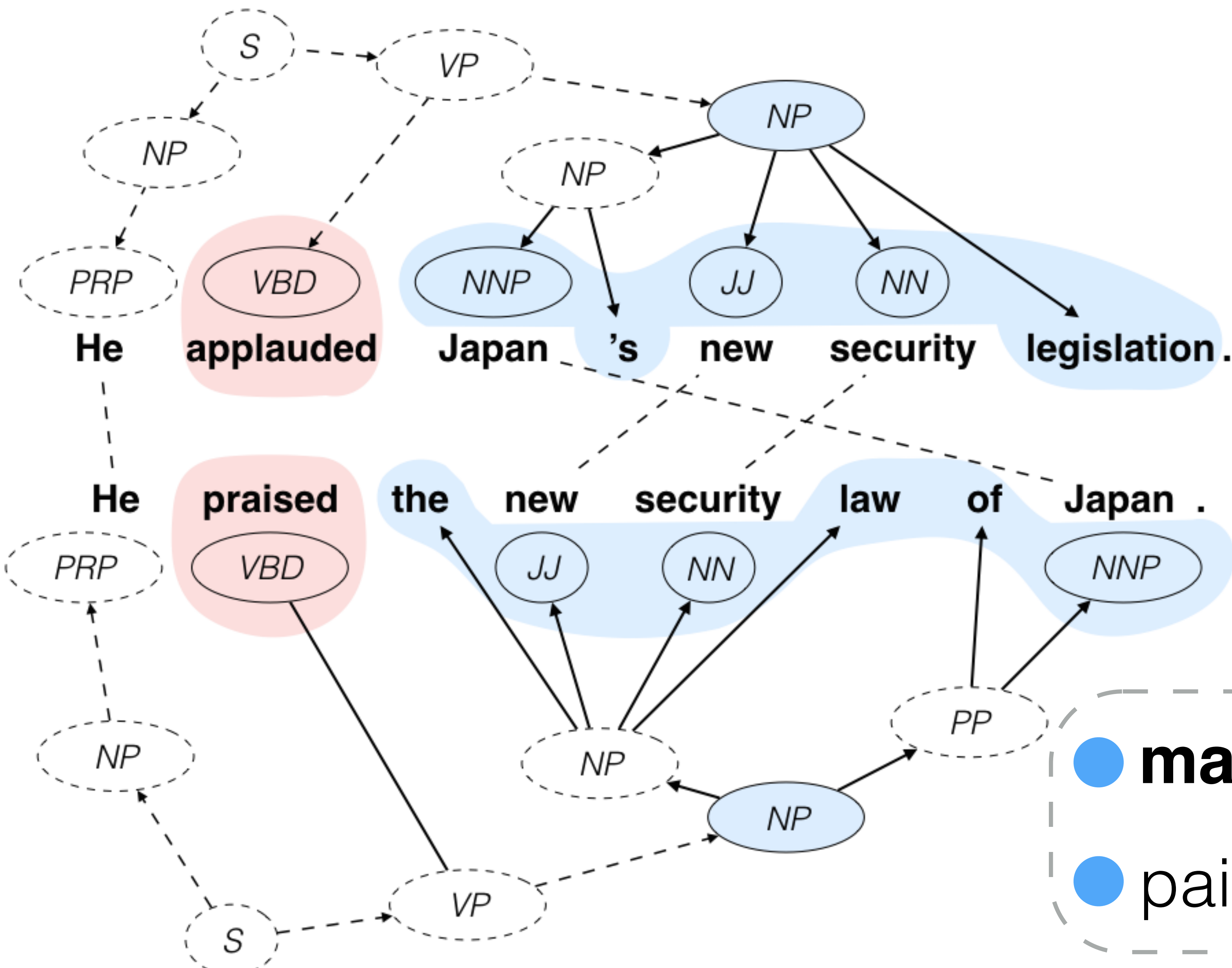
Stronger evidence can be much harder to find and score…

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

(courtesy: Salim Roukos)

Natural Language Generation

e.g. Text Simplification



Techniques

- **machine translation**
- pairwise ranking optimization

Wei Xu, Chris Callison-Burch, Courtney Napoles. "Problems in Current Text Simplification Research: New Data Can Help" in TACL (2015)
Wei Xu, Courtney Napoles, Ellie Pavlick, Chris Callison-Burch. "Optimizing Statistical Machine Translation for Simplification" in TACL (2016)

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!

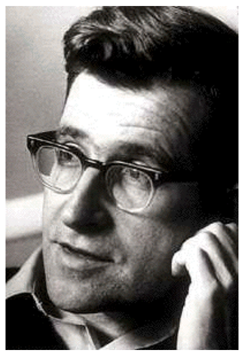
↓

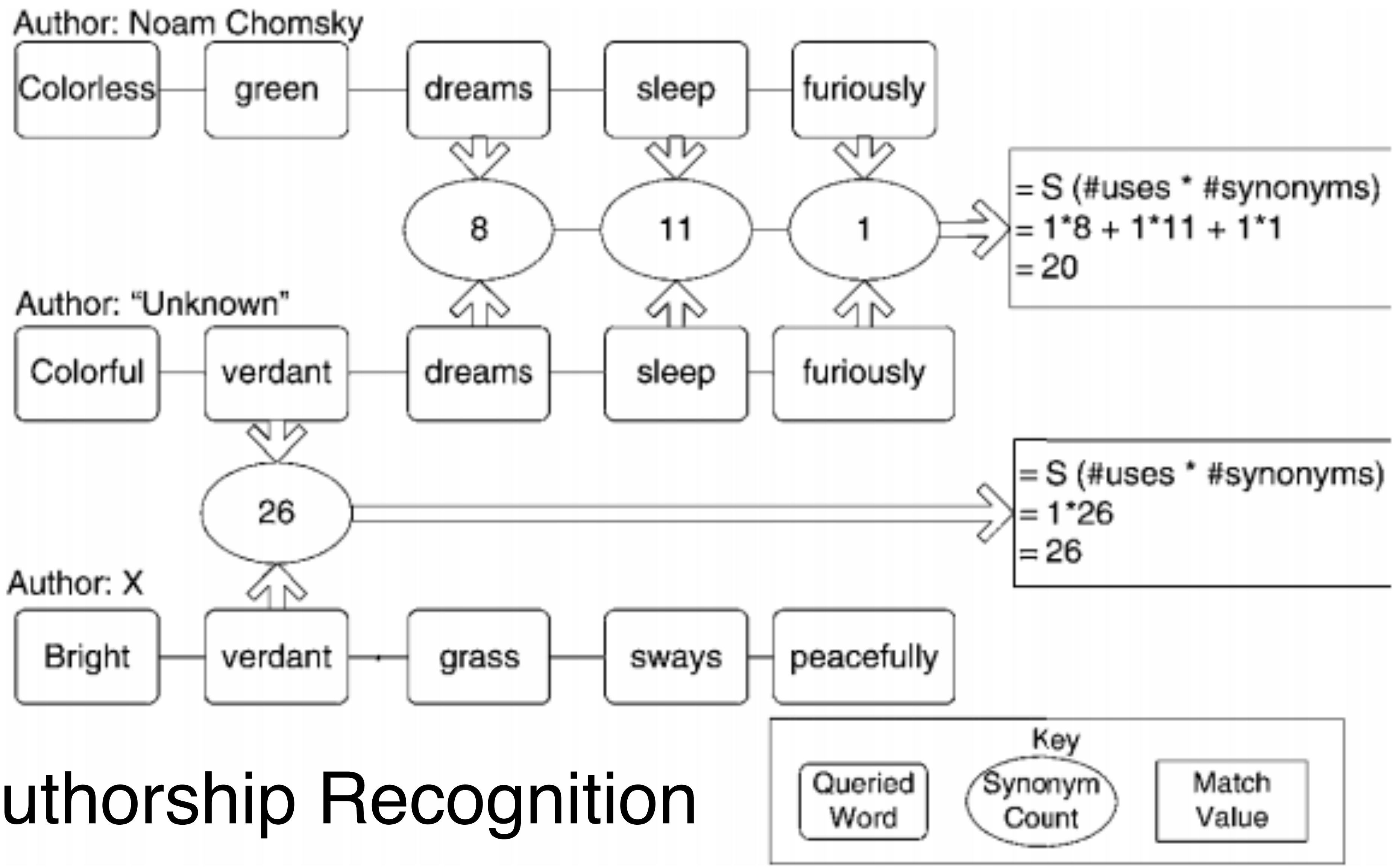
Father, I pray you! Help me!

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)

Plagiarism, Anonymity, Security

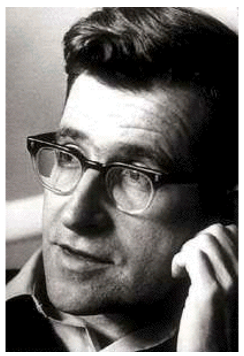




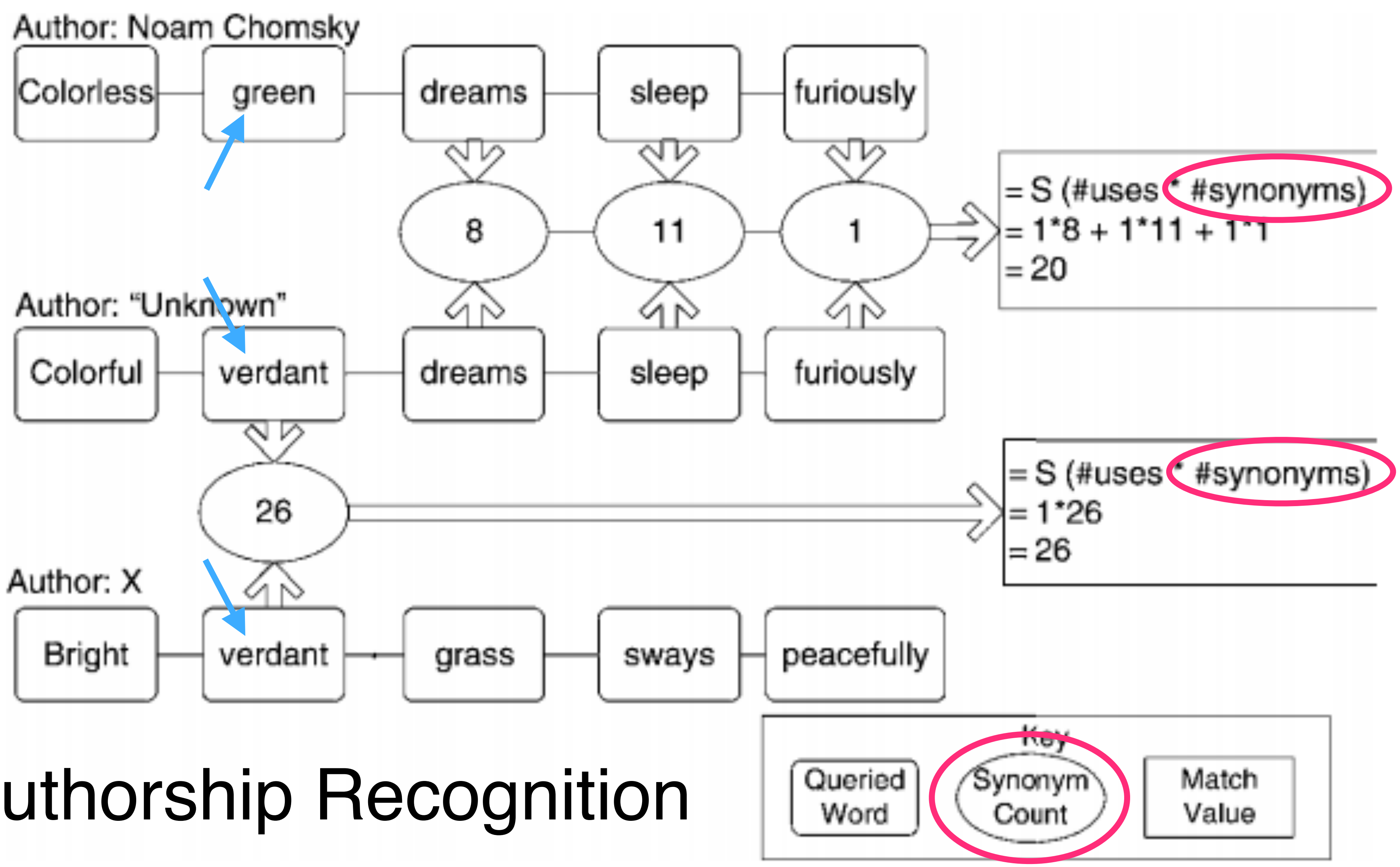
Authorship Recognition

(Clark and Hannon, 2007; Afroz et al. 2012; Brennan et al. 2012)

Plagiarism, Anonymity, Security



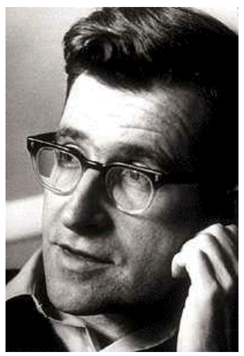
Paraphrases!!



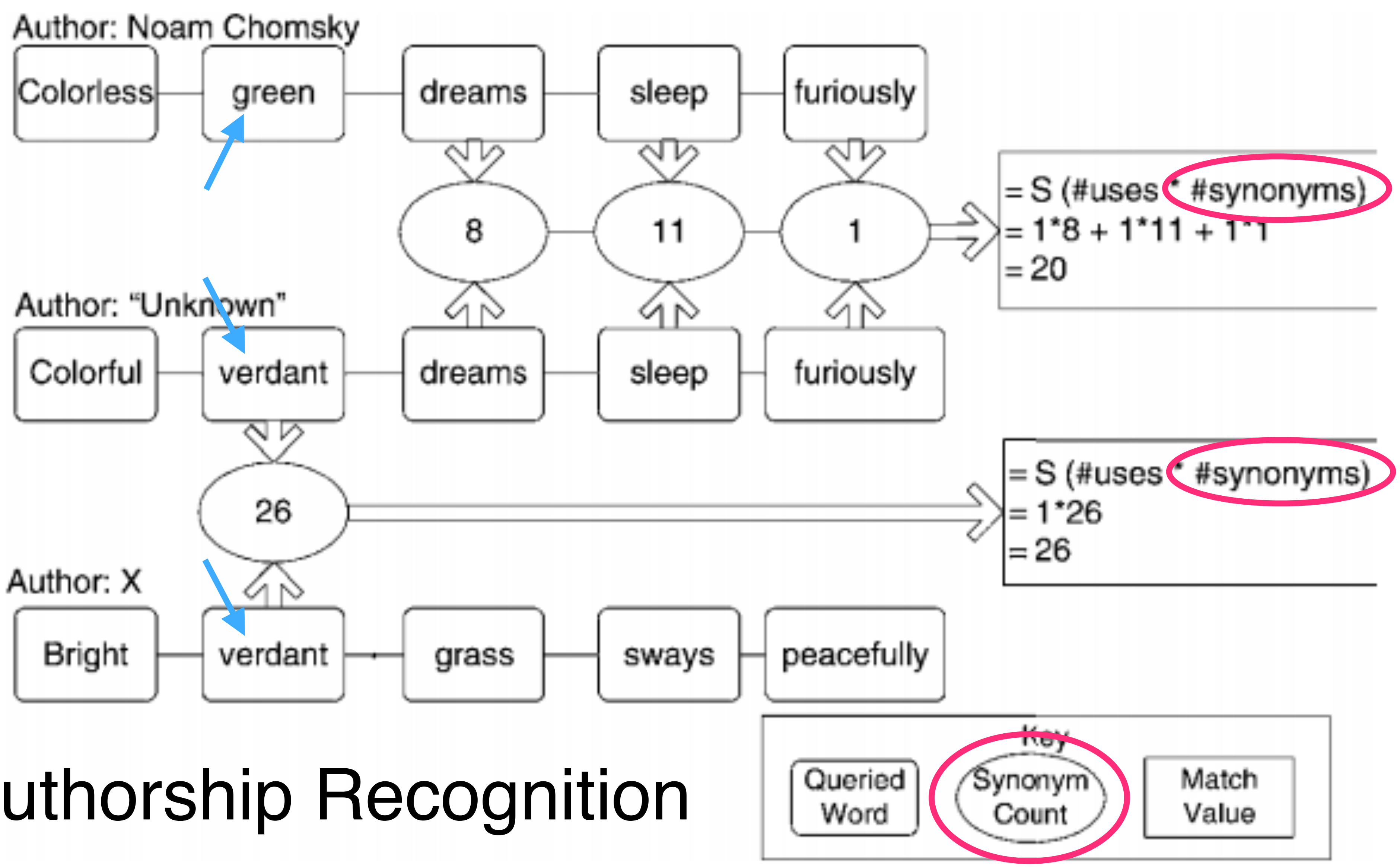
Authorship Recognition

(Clark and Hannon, 2007; Afroz et al. 2012; Brennan et al. 2012)

Plagiarism, Anonymity, Security



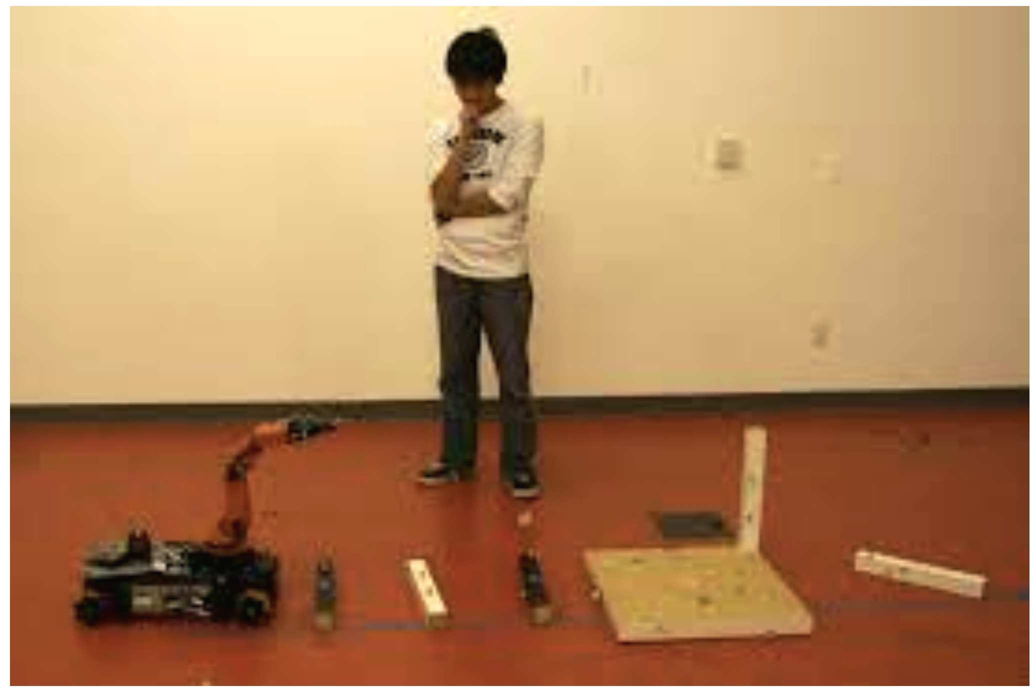
Paraphrases!!



Authorship Recognition

(Clark and Hannon, 2007; Afroz et al. 2012; Brennan et al. 2012)

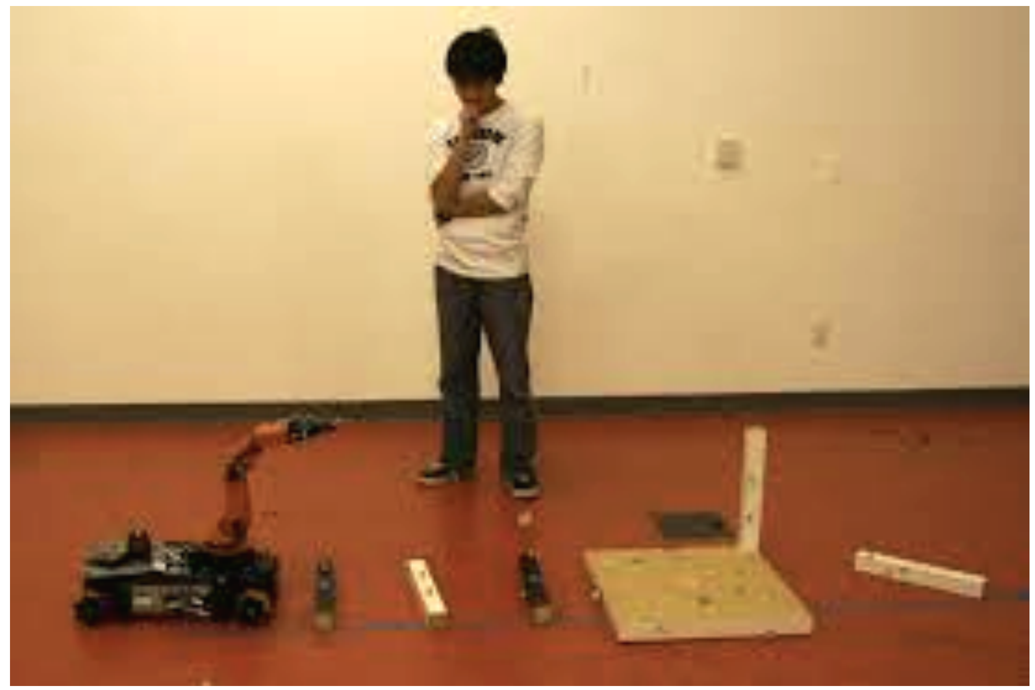
Language, Vision, Robotics, VR



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.

(Tellex et al. 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.

Paraphrases!!

Paraphrases!!

(Tellex et al. 2014)

Other Applications

fundamentally useful for a wide range of applications

*** my research**

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

80s WordNet

'01 Web

'01 Novels

'04 News

'05 '13 Bi-Text

'11 Video

'12* Style
Xu
Ritter
Dolan
Grishman
Cherry

'13* '14* Twitter
Xu
Ritter
Callison-Burch
Dolan
Ji

'15*'16* Simple
Xu
Callison-Burch
Napoles

* **my research**

Wei Xu. "Data-driven Approaches for Paraphrasing Across Language Variations" PhD Thesis (2014)

Paraphrase Research

80s WordNet

‘01 Web

'01 Novels

'04 News

'05 '13 Bi-Text

'11 Video

'12* Style

Xu
Ritter
Dolan
Grishman
Cherry

'13* '14* Twitter

Xu
Ritter
Callison-Burch
Dolan
Ji

'15*'16* Simple

Xu
Callison-Burch
Napoles

* my research

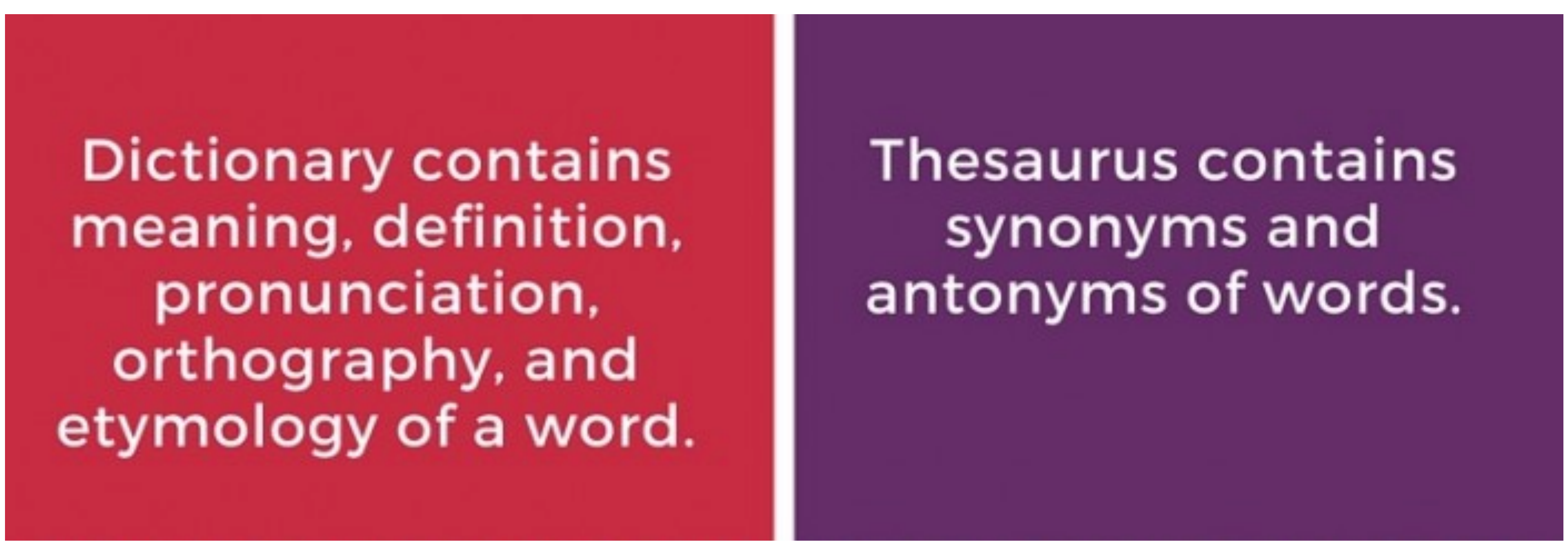
WordNet®

- 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

(Miller, Beckwith, Fellbaum, Gross, & Miller 1990; Miller & Fellbaum, 1991; Miller, 1995; Fellbaum, 1998)

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/



WordNet®

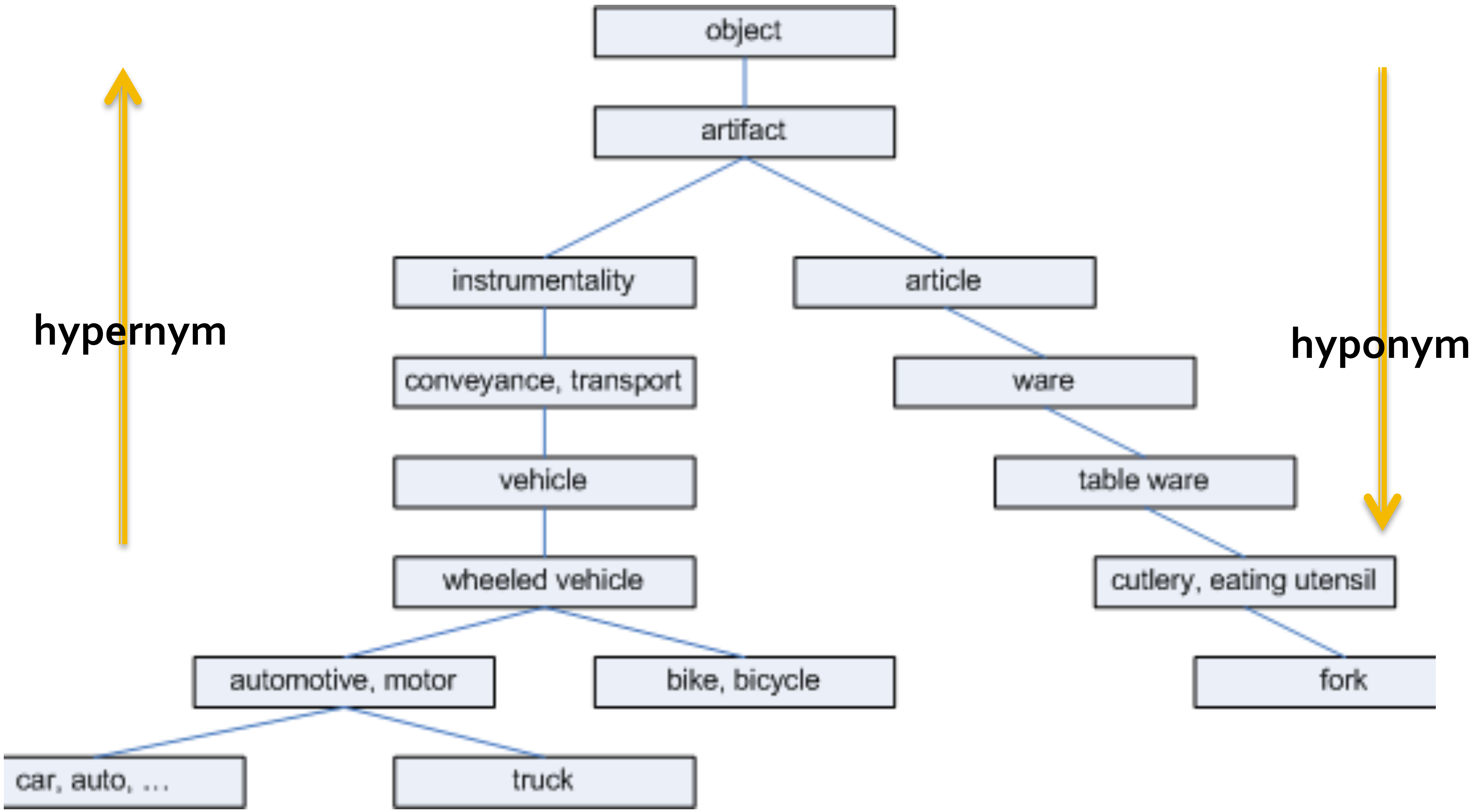
- 4 types of Parts of Speech (POS)
  - Noun, Verb, Adjective, Adverb
- Synset (synonym set)
  - the smallest unit in WordNet
  - represents a specific meaning of a word

- 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 anther 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®


object
artifact
instrumentality
article
hypernym
conveyance, transport
ware
hyponym
vehicle
table ware
wheeled vehicle
cutlery, eating utensil
automotive, motor
bike, bicycle
fork
car, auto, ...
truck

WordNet®

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

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

WordNet®

Google Scholar | wordnet

Articles

About 94,700 results (**0.08** sec)

Any time
Since 2017
Since 2016
Since 2013
Custom range...

Sort by relevance
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
Cited by 9594 Related articles All 34 versions Web of Science: 2440

[BOOK] **WordNet**
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,
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
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
Cited by 1504 Related articles All 37 versions

ImageNet

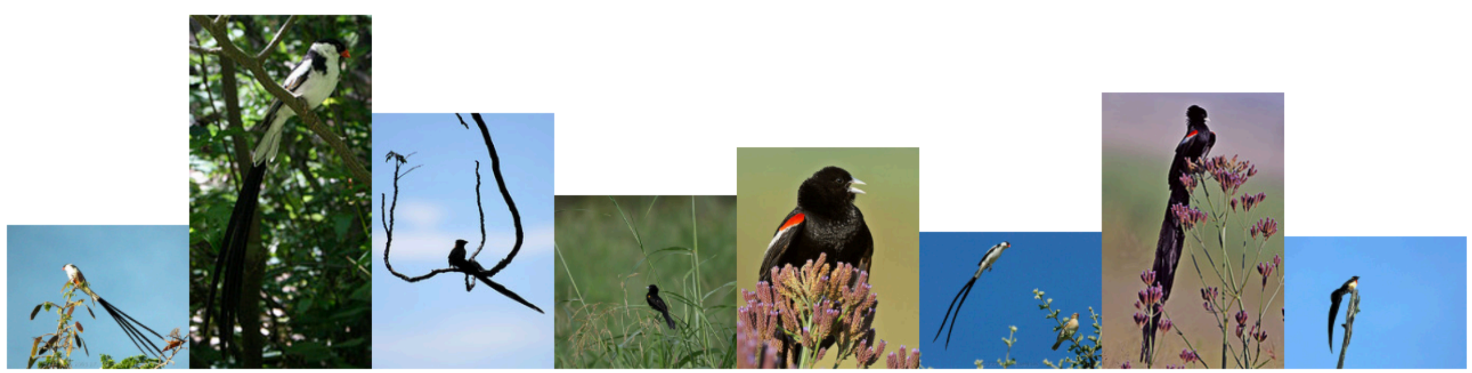
IMAGENET

14,197,122 images, 21841 synsets indexed

Explore Download Challenges Publications CoolStuff About

Not logged in. Login I Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.
Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.



What do these images have in common? *Find out!*

Check out the ImageNet Challenge on Kaggle!

More info: https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures

Paraphrase Research

80s WordNet

‘01 Web

'01 Novels

'04 News

'05 '13 Bi-Text

'11 Video

'12* Style
Xu
Ritter
Dolan
Grishman
Cherry

'13* '14* Twitter
Xu
Ritter
Callison-Burch
Dolan
Ji

'15*'16* Simple
Xu
Callison-Burch
Napoles

*** my research**

Paraphrase Research

80s WordNet

‘01 Web

'01 Novels

'04 News

'05 '13 Bi-Text

'11 Video

'12* Style

Xu
Ritter
Dolan
Grishman
Cherry

'13* '14* Twitter

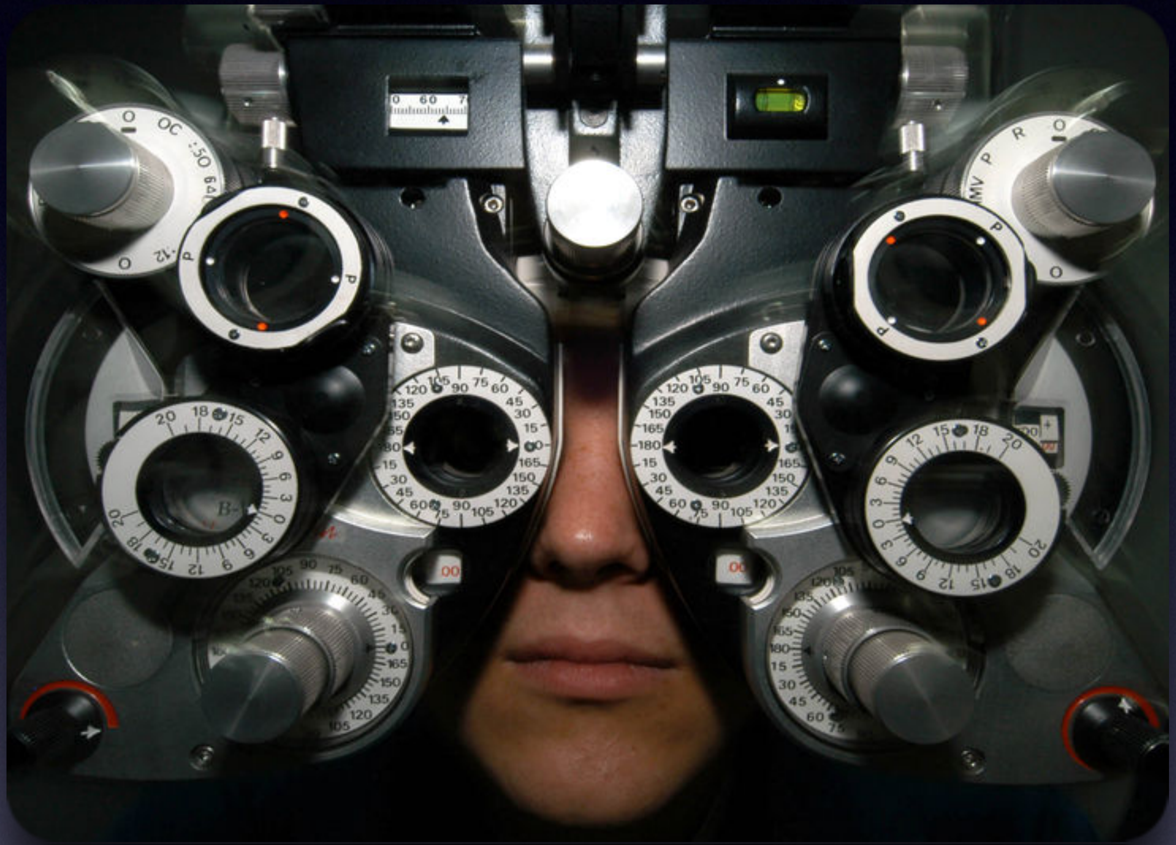
Xu
Ritter
Callison-Burch
Dolan
Ji

'15*'16* Simple

Xu
Callison-Burch
Napoles

* my research

Distributional Hypothesis



Source: Chris Callison-Burch

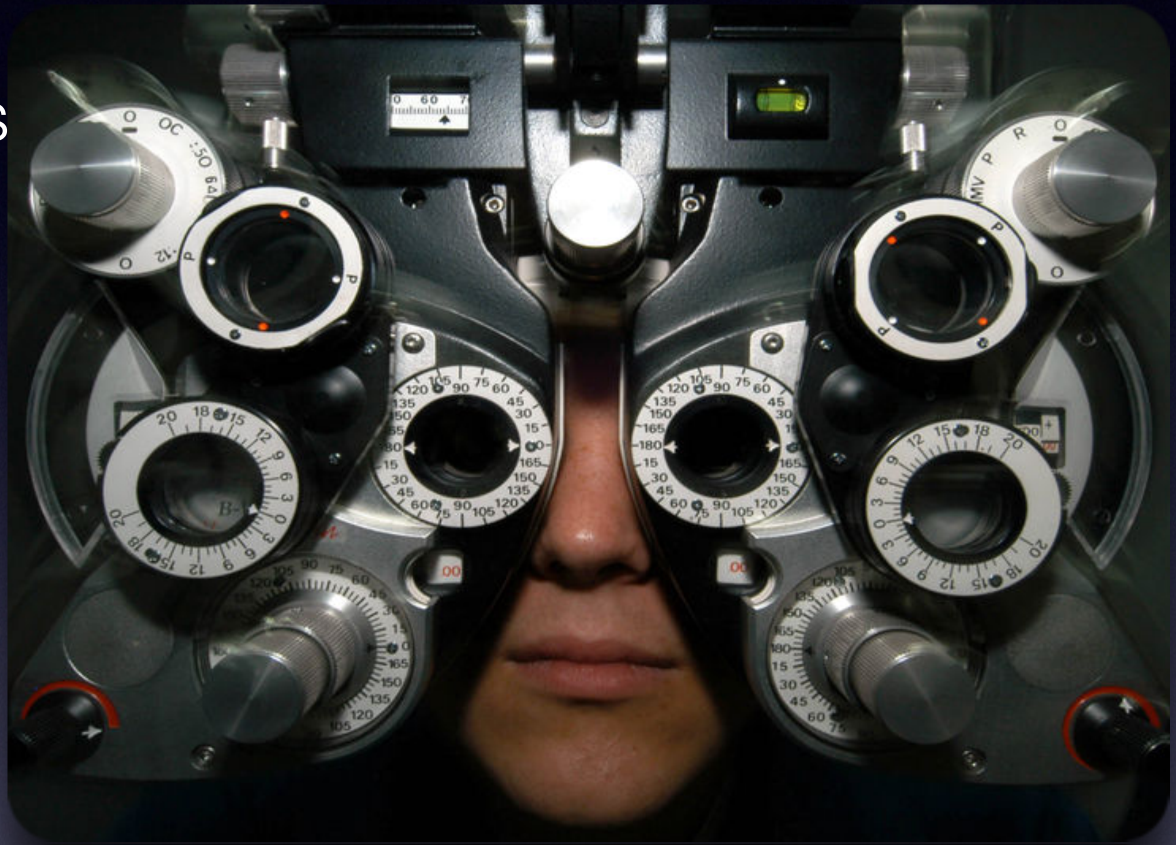
Distributional Hypothesis

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)



Source: Chris Callison-Burch

DIRT

(Discovery of Inference Rules from Text)

Lin and Panel (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

Decking Lin and Patrick Pantel. “DIRT - Discovery of Inference Rules from Text” In KDD (2001)

Paraphrase Research

80s WordNet

‘01 Web

'01 Novels

'04 News

'05 '13 Bi-Text

'11 Video

'12* Style

Xu
Ritter
Dolan
Grishman
Cherry

'13* '14* Twitter

Xu
Ritter
Callison-Burch
Dolan
Ji

'15*'16* Simple

Xu
Callison-Burch
Napoles

* my research

Paraphrase Research

80s WordNet

'01 Web

'01 Novels

'04 News

'05 '13 Bi-Text

'11 Video

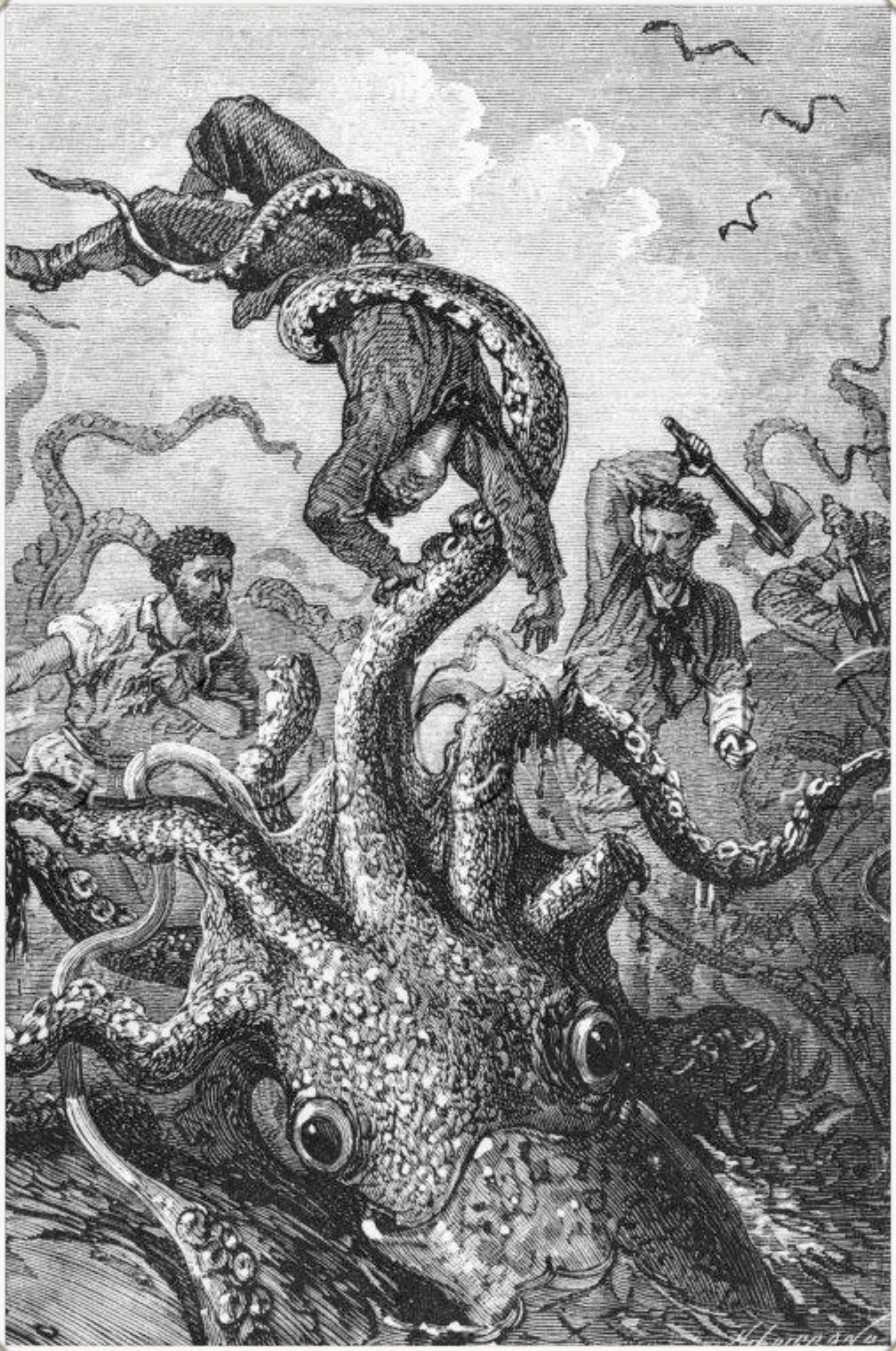
'12* Style
- **Xu**
- Ritter
- Dolan
- Grishman
- Cherry

'13* '14* Twitter
- **Xu**
- Ritter
- Callison-Burch
- Dolan
- Ji

'15*'16* Simple
- **Xu**
- Callison-Burch
- Napoles

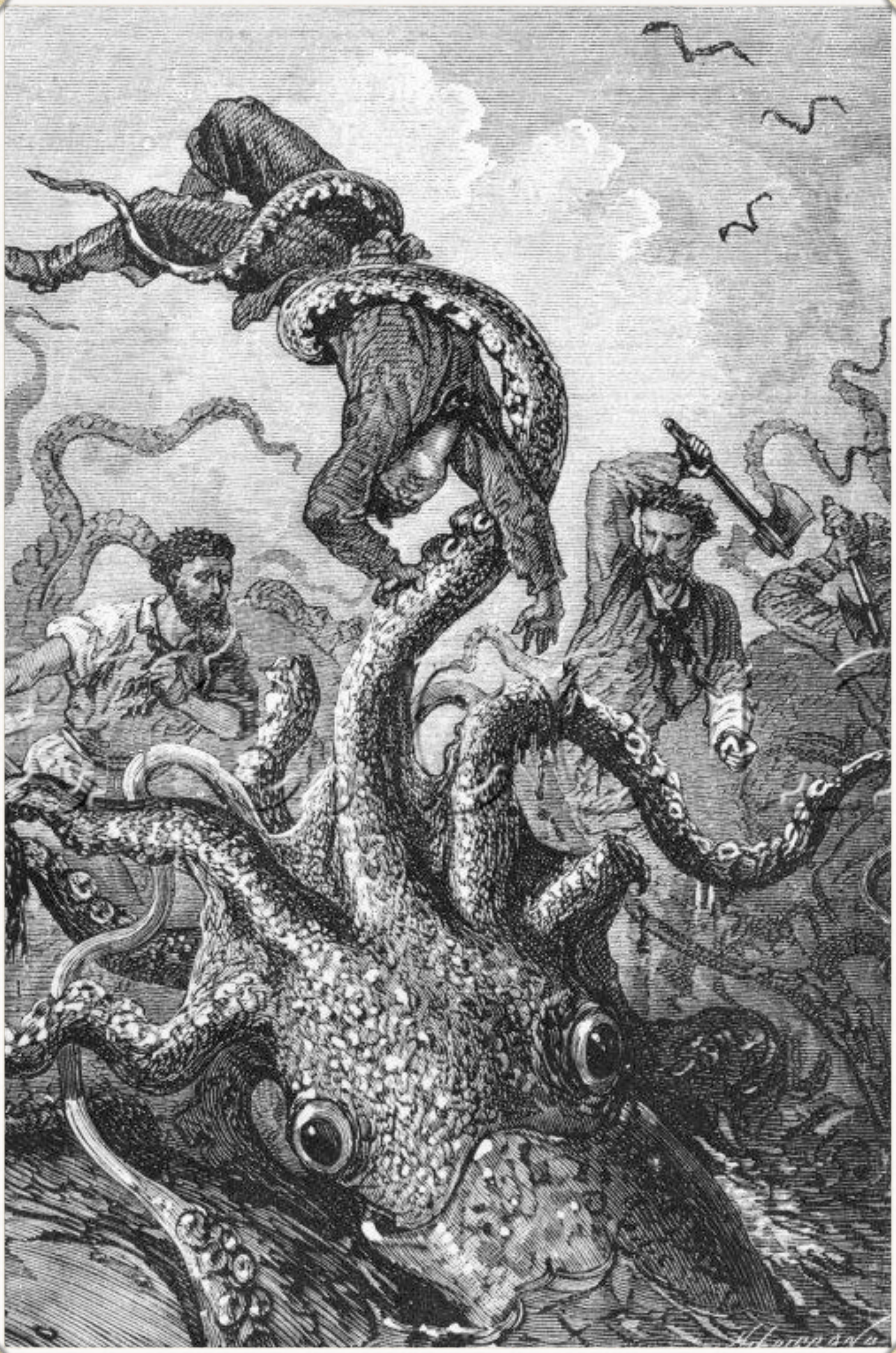
* my research

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**.

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! 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**.

Source: Chris Callison-Burch

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

Source: Chris Callison-Burch

Regina Barzilay and Kathleen McKeown. "Extracting Paraphrases from a Parallel Corpus" In ACL (2001)

Paraphrase Research

- **80s WordNet**
- **‘01 Web**
- **'01 Novels**
- **'04 News**
- **'05 '13 Bi-Text**
- **'11 Video**
- **'12* Style**
  - **Xu**
  - Ritter
  - Dolan
  - Grishman
  - Cherry
- **'13* '14* Twitter**
  - **Xu**
  - Ritter
  - Callison-Burch
  - Dolan
  - Ji
- **'15*'16* Simple**
  - **Xu**
  - Callison-Burch
  - Napoles

*** my research**

Paraphrase Research

80s	‘01	'01	**'04**	'05 '13	'11	'12*	'13* '14*	'15*'16*
WordNet	Web	Novels	**News**	Bi-Text	Video	Style	Twitter	Simple

'12* Style:
Xu
Ritter
Dolan
Grishman
Cherry

'13* '14* Twitter:
Xu
Ritter
Callison-Burch
Dolan
Ji

'15*'16* Simple:
Xu
Callison-Burch
Napoles

* my research

Wei Xu. “Data-driven Approaches for Paraphrasing Across Language Variations” PhD Thesis (2014)

News

Microsoft Research Paraphrase Corpus

(Dolan, Quirk and Brockett, 2004; Dolan and Brockett, 2005; Brockett and Dolan, 2005)

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.

Source: Chris Callison-Burch

Bill Dolan, Chris Quirk, and Chris Brockett. "Extracting Paraphrases from a Parallel Corpus" In COLING (2004)

Paraphrase Research

- **80s WordNet**
- **‘01 Web**
- **'01 Novels**
- **'04 News**
- **'05 '13 Bi-Text**
- **'11 Video**
- **'12* Style**
  Xu
  Ritter
  Dolan
  Grishman
  Cherry
- **'13* '14* Twitter**
  Xu
  Ritter
  Callison-Burch
  Dolan
  Ji
- **'15*'16* Simple**
  Xu
  Callison-Burch
  Napoles

*** my research**

Paraphrase Research

80s WordNet

‘01 Web

'01 Novels

'04 News

'05 '13 Bi-Text

'11 Video

'12* Style
Xu
Ritter
Dolan
Grishman
Cherry

'13* '14* Twitter
Xu
Ritter
Callison-Burch
Dolan
Ji

'15*'16* Simple
Xu
Callison-Burch
Napoles

* my 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.

Source: Chris Callison-Burch

Bilingual Data

Sentence-aligned parallel corpora in English and any foreign language

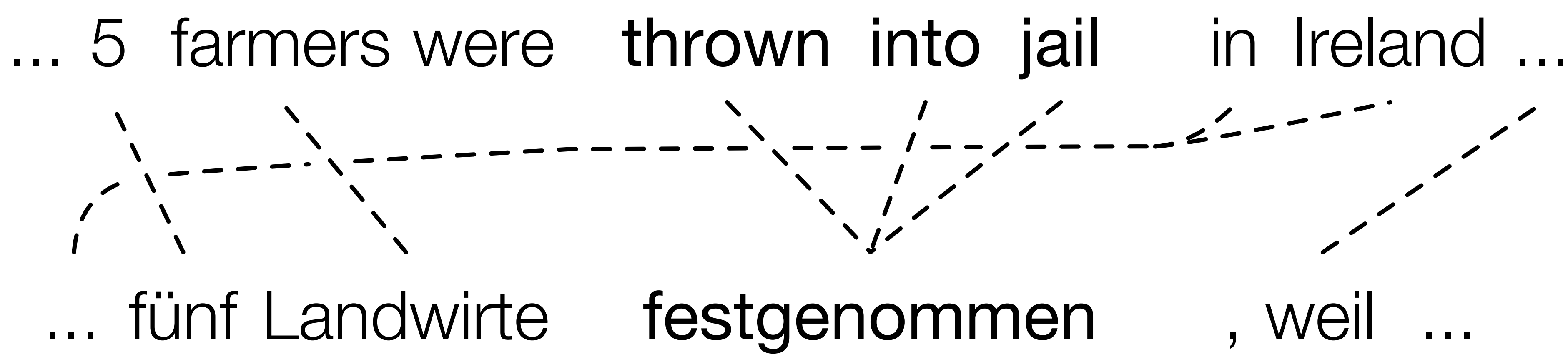
Available in large quantities

Strong meaning equivalence signal

... but different languages.

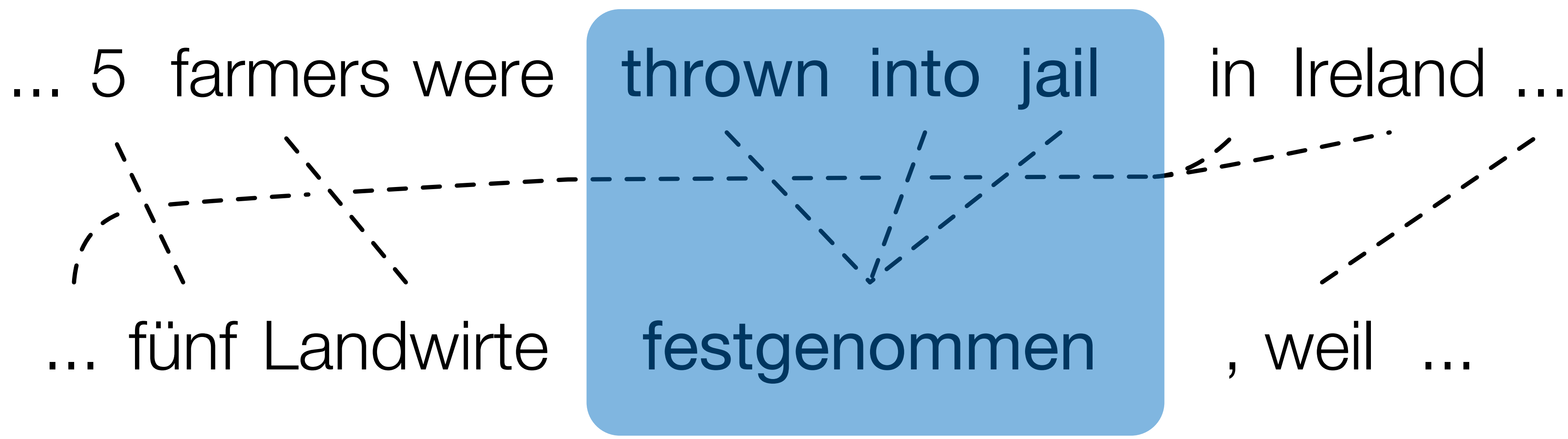
Source: Chris Callison-Burch

Bilingual Pivoting



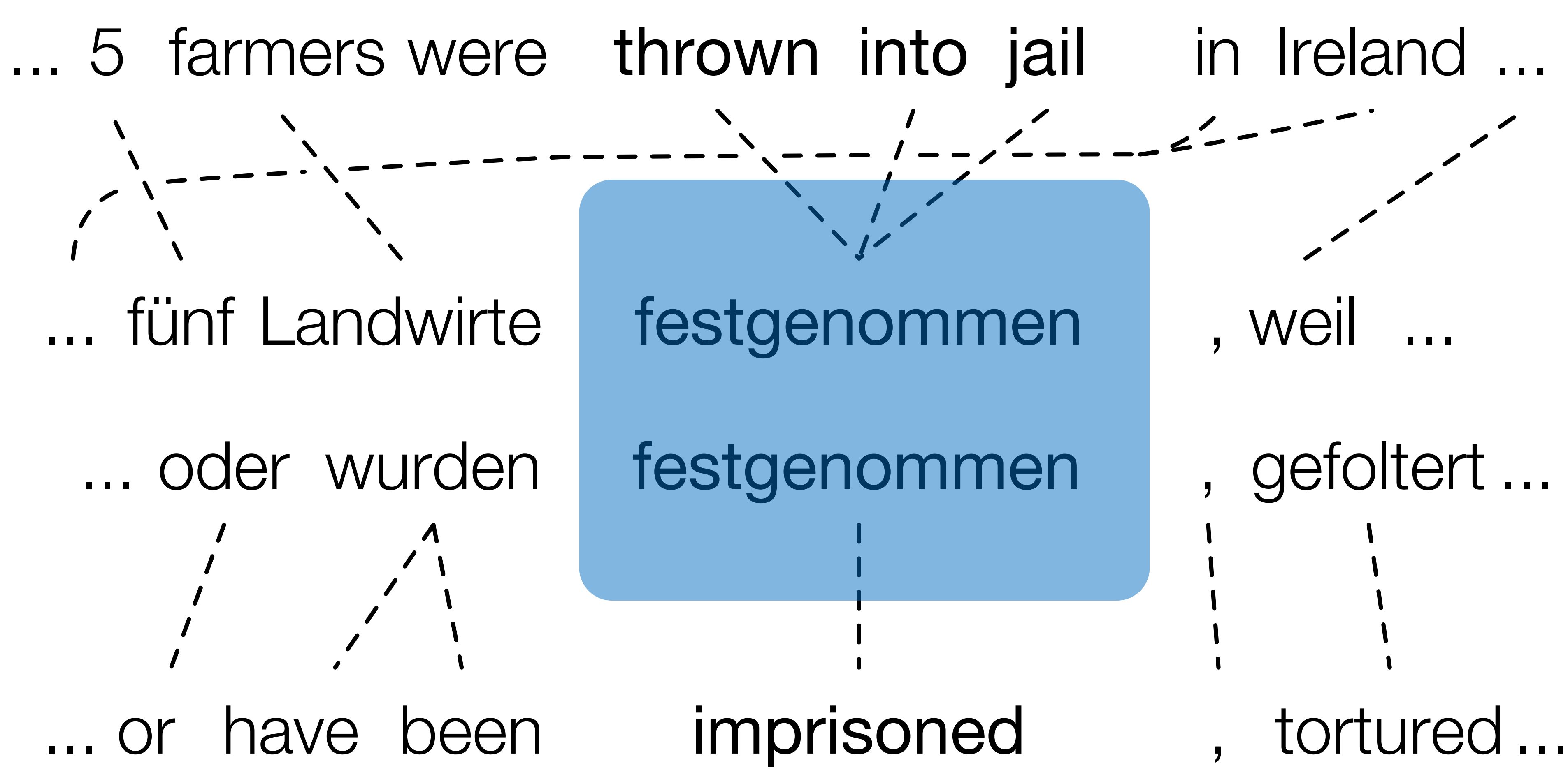
Source: Chris Callison-Burch

Bilingual Pivoting



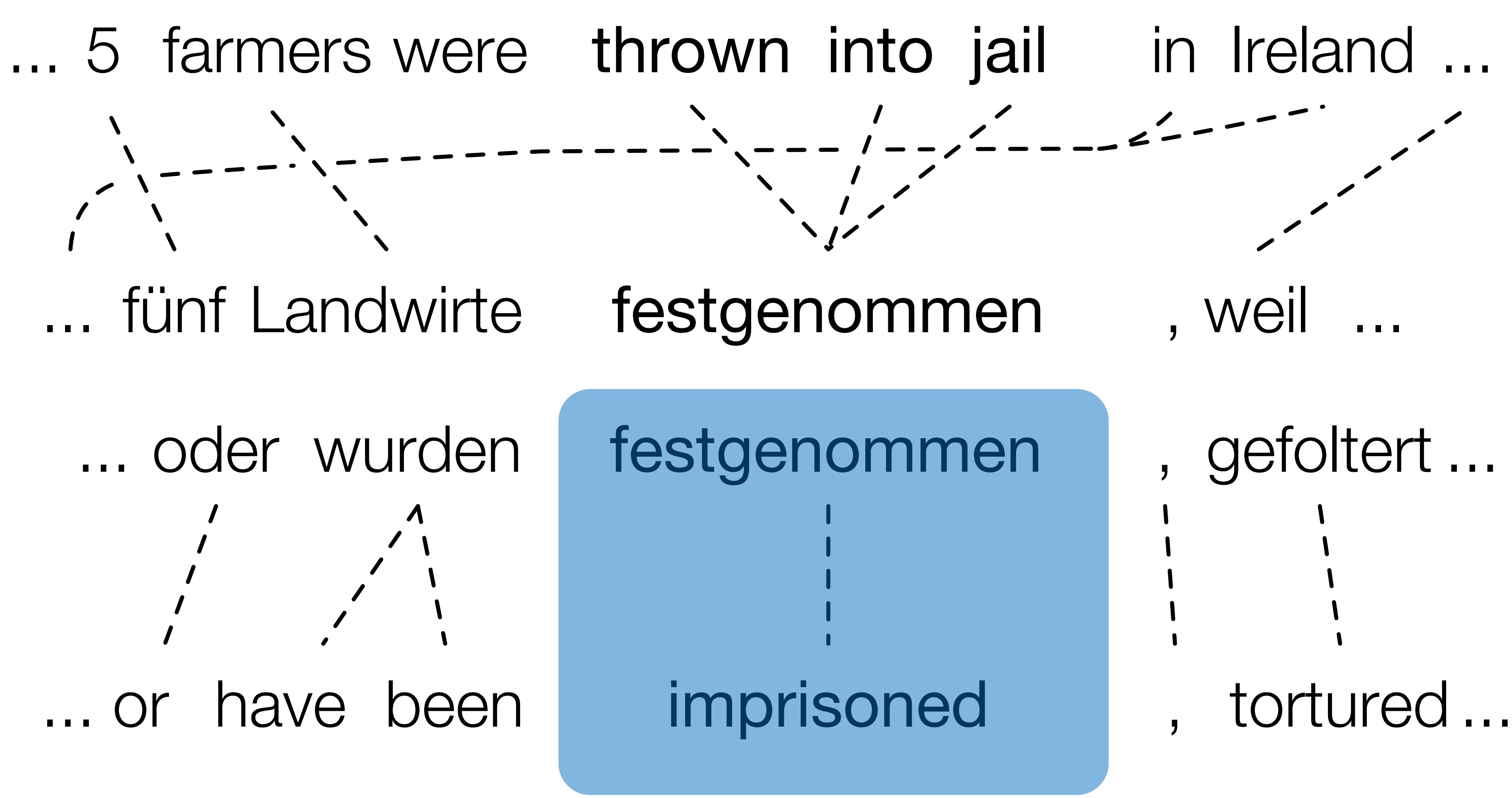
Source: Chris Callison-Burch

Bilingual Pivoting



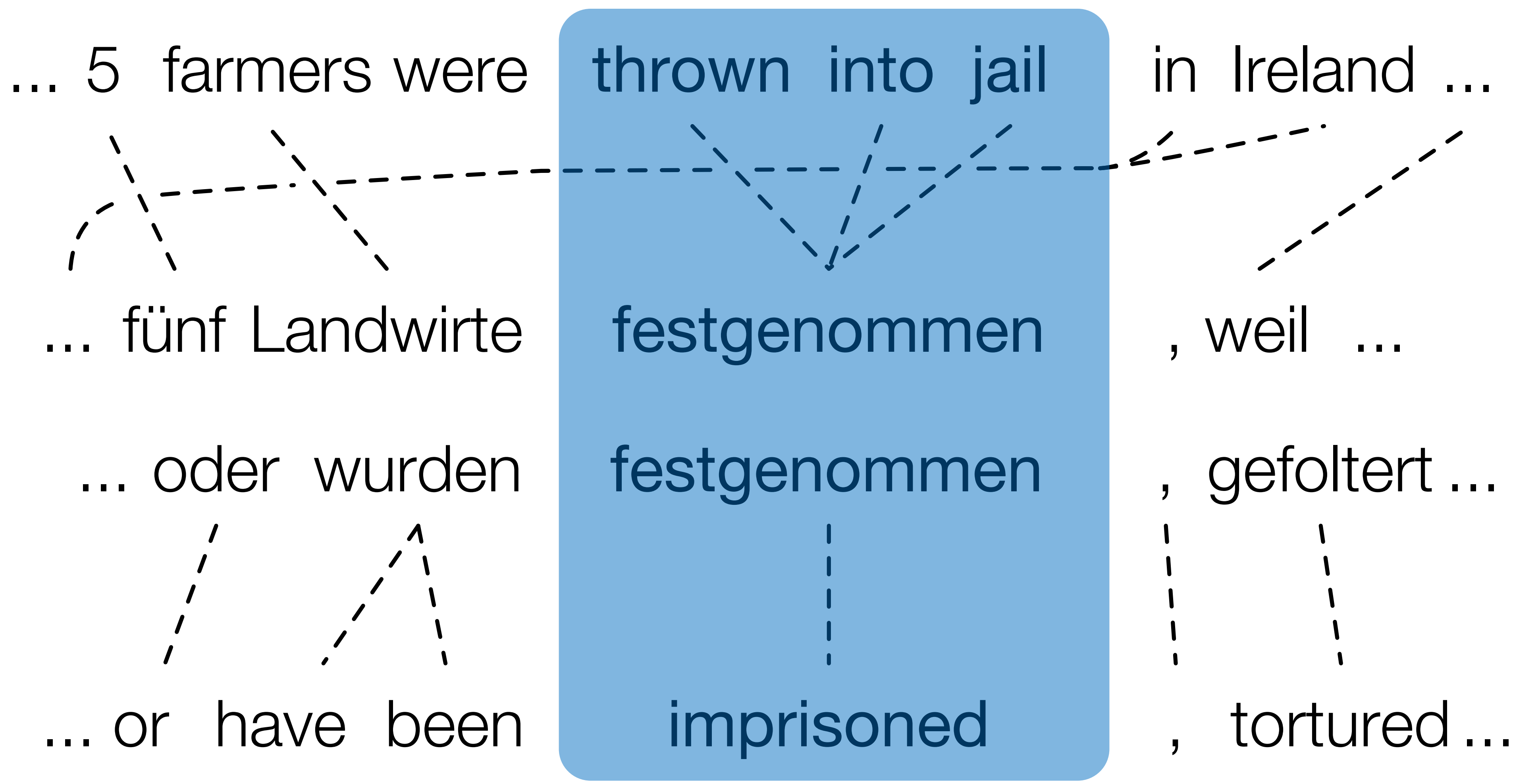
Source: Chris Callison-Burch

Bilingual Pivoting



Source: Chris Callison-Burch

Bilingual Pivoting



Source: Chris Callison-Burch

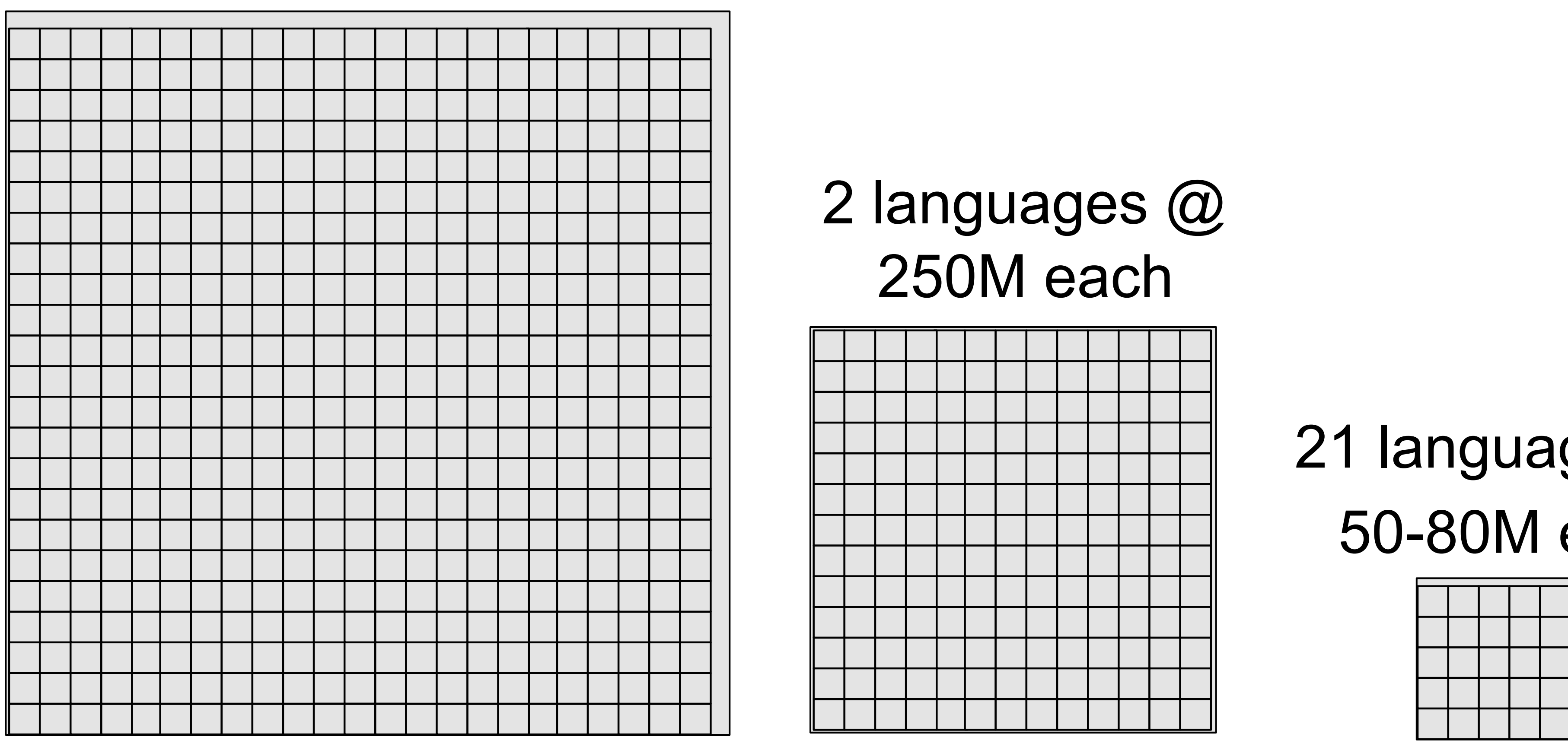
Large and diverse

Bilingual Data Sets

1000M

2 languages @ 250M each

21 languages @ 50-80M each



French-English
10^9 word webcrawl

DARPA
GALE Program

European
Parliament

Source: Chris Callison-Burch

Wide range of Paraphrases

thrown into jail

Source: Chris Callison-Burch

Wide range of Paraphrases

thrown into jail

arrested
detained
imprisoned
incarcerated
jailed
locked up
taken into custody
thrown into prison

Source: Chris Callison-Burch

Wide range of Paraphrases

thrown into jail

arrested
detained
imprisoned
incarcerated
jailed
locked up
taken into custody
thrown into prison

be thrown in prison
been thrown into jail
being arrested
in jail
in prison
put in prison for
were thrown into jail
who are held in detention

Source: Chris Callison-Burch

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	

Source: Chris Callison-Burch

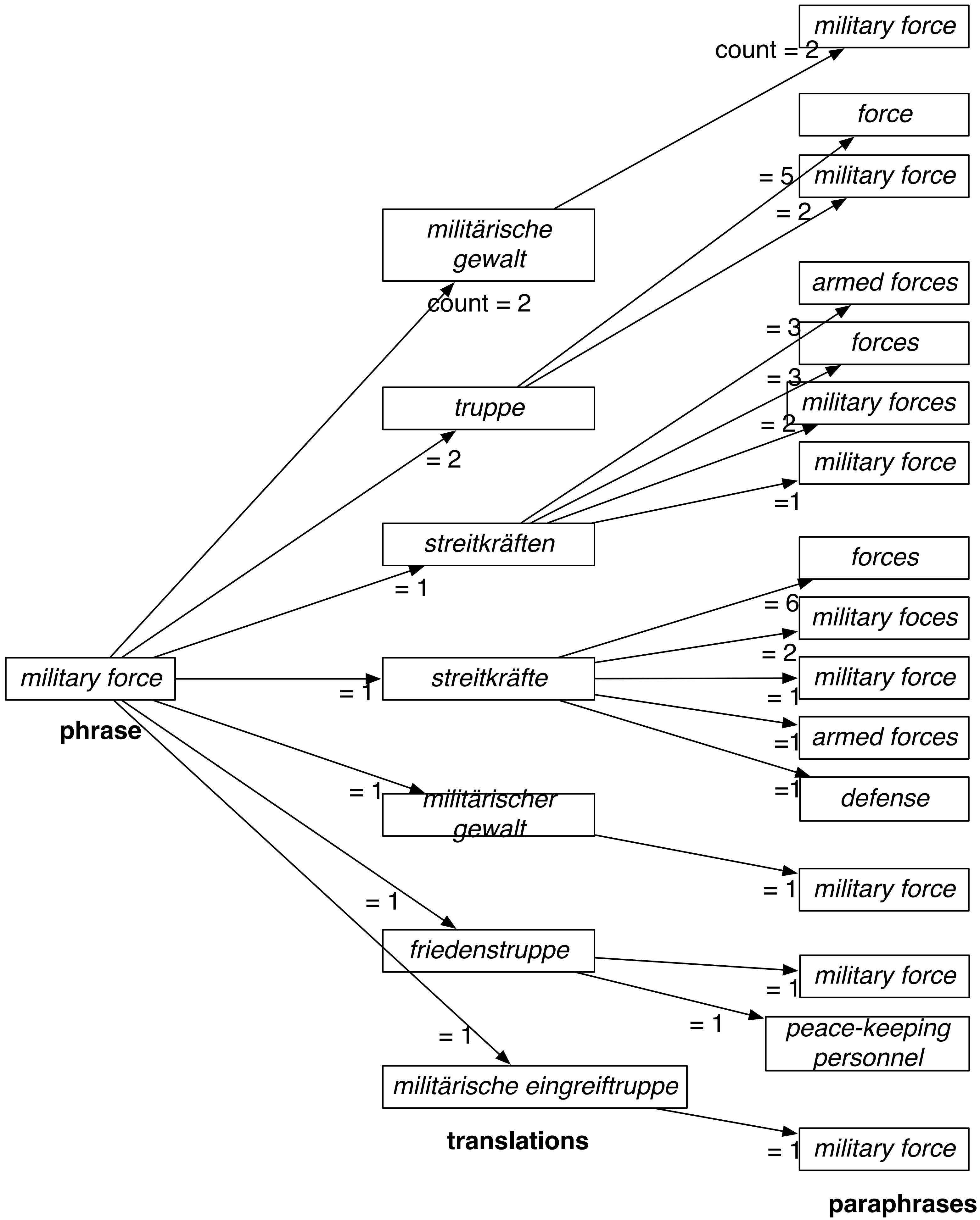
Paraphrase Probability

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

Source: Chris Callison-Burch

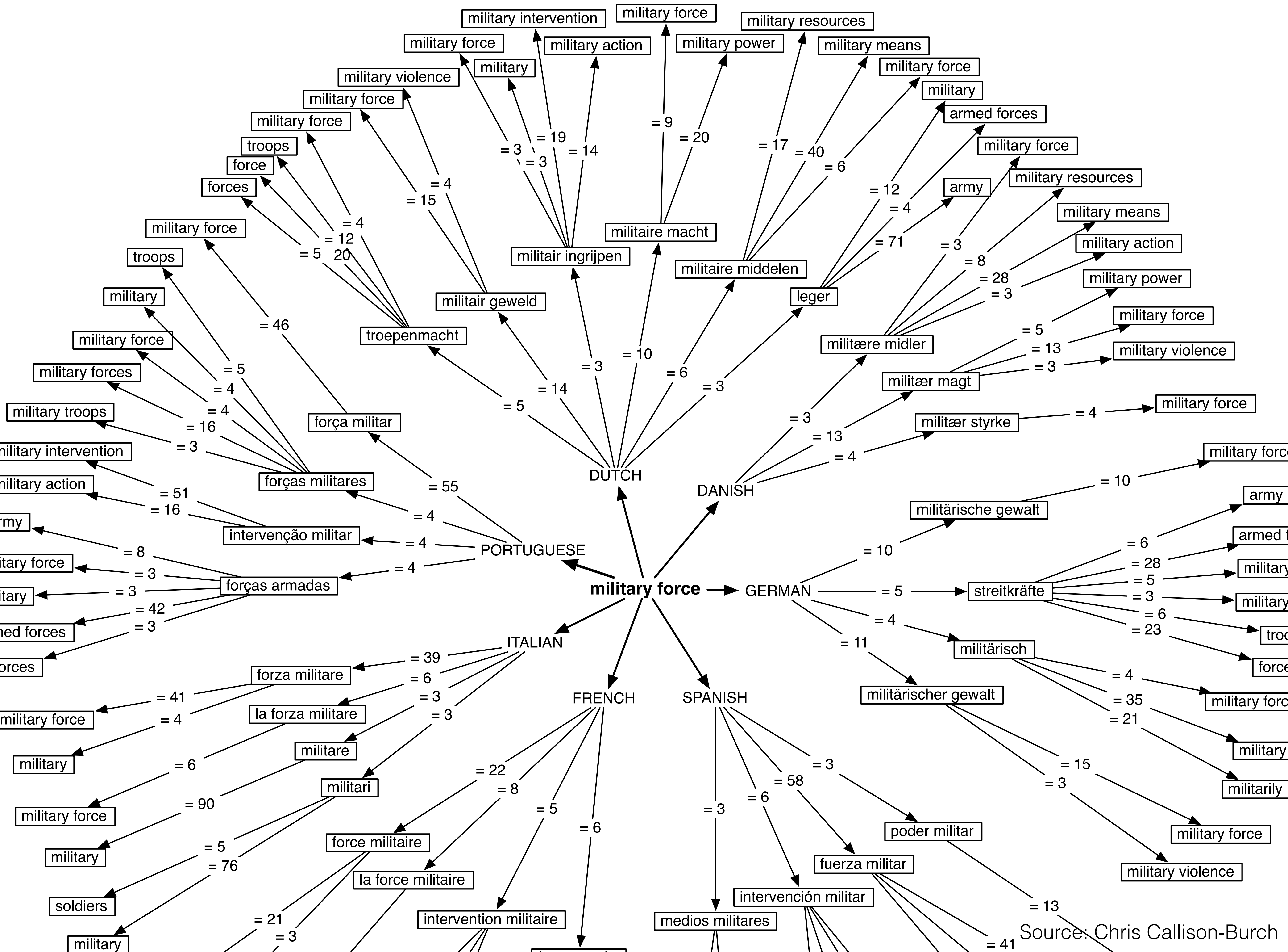
Colin Bannard and Chris Callison-Burch. Paraphrasing with Bilingual Parallel Corpora. ACL 2005.

Source: Chris Callison-Burch



Source: Chris Callison-Burch

Source: Chris Callison-Burch



Source: Chris Callison-Burch

Syntactic Constraints

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~~	

Source: Chris Callison-Burch

Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. Chris Callison-Burch. EMNLP 2008.

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?"

Source: Chris Callison-Burch

Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

Source: Chris Callison-Burch

Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

cup $\leftrightarrow$ mug

Source: Chris Callison-Burch

Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

..sip from a **cup** of cocoa.. ..a **cup** of coffee.		..sip from a **mug** of cocoa.. ..a **mug** of coffee.
cup	$\leftrightarrow$	**mug**

Source: Chris Callison-Burch

Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

✓ cup $\leftrightarrow$ mug

Source: Chris Callison-Burch

Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

✓	cup	$\leftrightarrow$	mug
	the king's speech	$\leftrightarrow$	His Majesty's address

Source: Chris Callison-Burch

Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

✓ cup ↔ mug

..anxiously awaiting **the king's speech**..

..anxiously awaiting **His Majesty's address**..

the king's speech ↔ His Majesty's address

Source: Chris Callison-Burch

Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

✓	cup	$\leftrightarrow$	mug
✓	the king's speech	$\leftrightarrow$	His Majesty's address

Source: Chris Callison-Burch

Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

✓	cup	$\leftrightarrow$	mug
✓	the king's speech	$\leftrightarrow$	His Majesty's address
	one JJ instance of NP	$\leftrightarrow$	a JJ case of NP

Source: Chris Callison-Burch

Similarity

Easy for lexical & phrasal paraphrases

More involved for syntactic paraphrases

✓	cup	$\leftrightarrow$	mug
✓	the king's speech	$\leftrightarrow$	His Majesty's address
?	one JJ instance of NP	$\leftrightarrow$	a JJ case of NP

Source: Chris Callison-Burch

Syntactic Paraphrase Similarity

NN 's NP in the long term

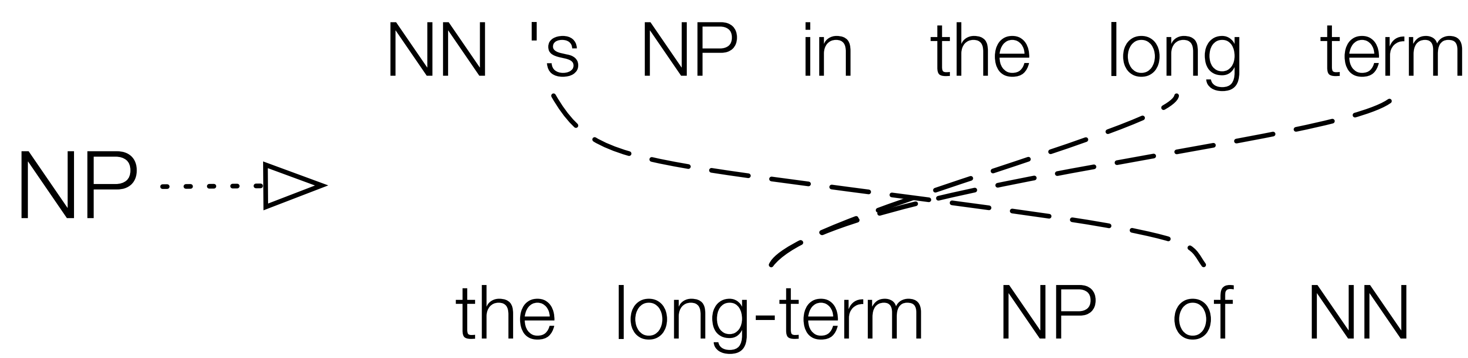
NP ⊳

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.

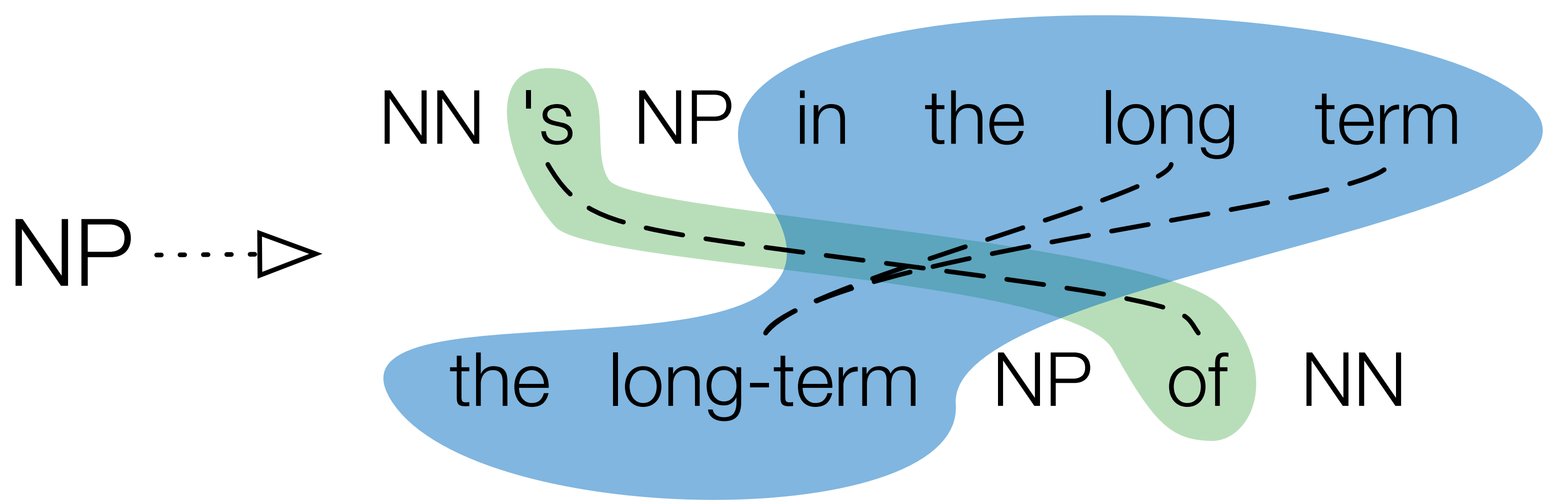
Syntactic Paraphrase Similarity



Source: Chris Callison-Burch

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

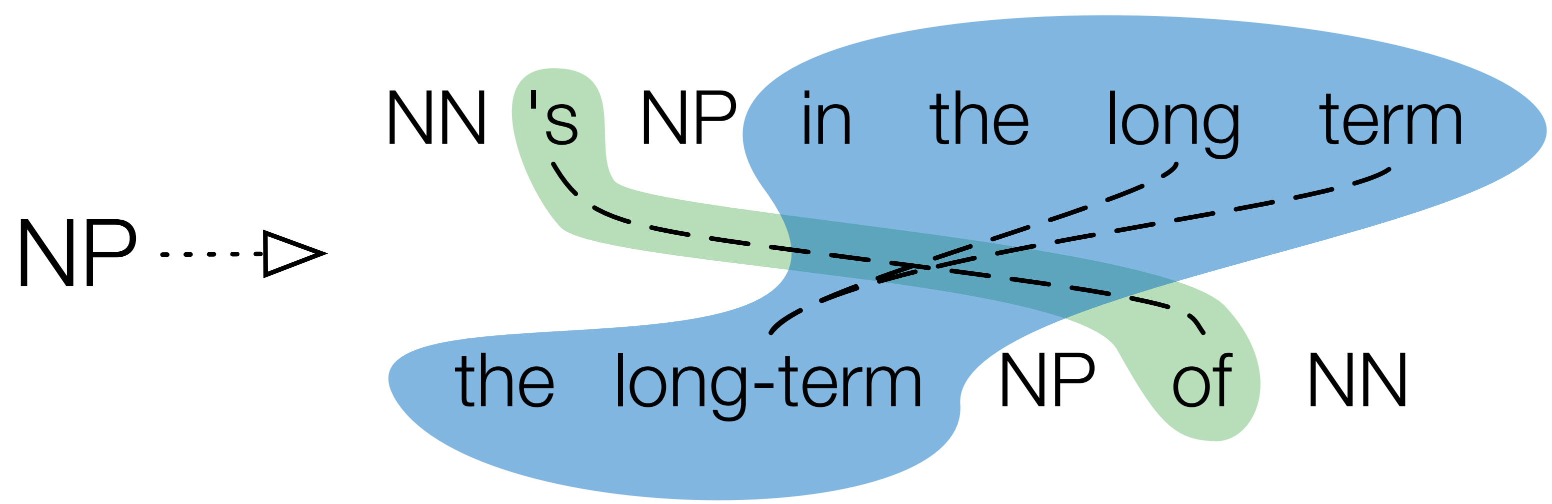
Syntactic Paraphrase Similarity



Source: Chris Callison-Burch

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

Syntactic Paraphrase Similarity

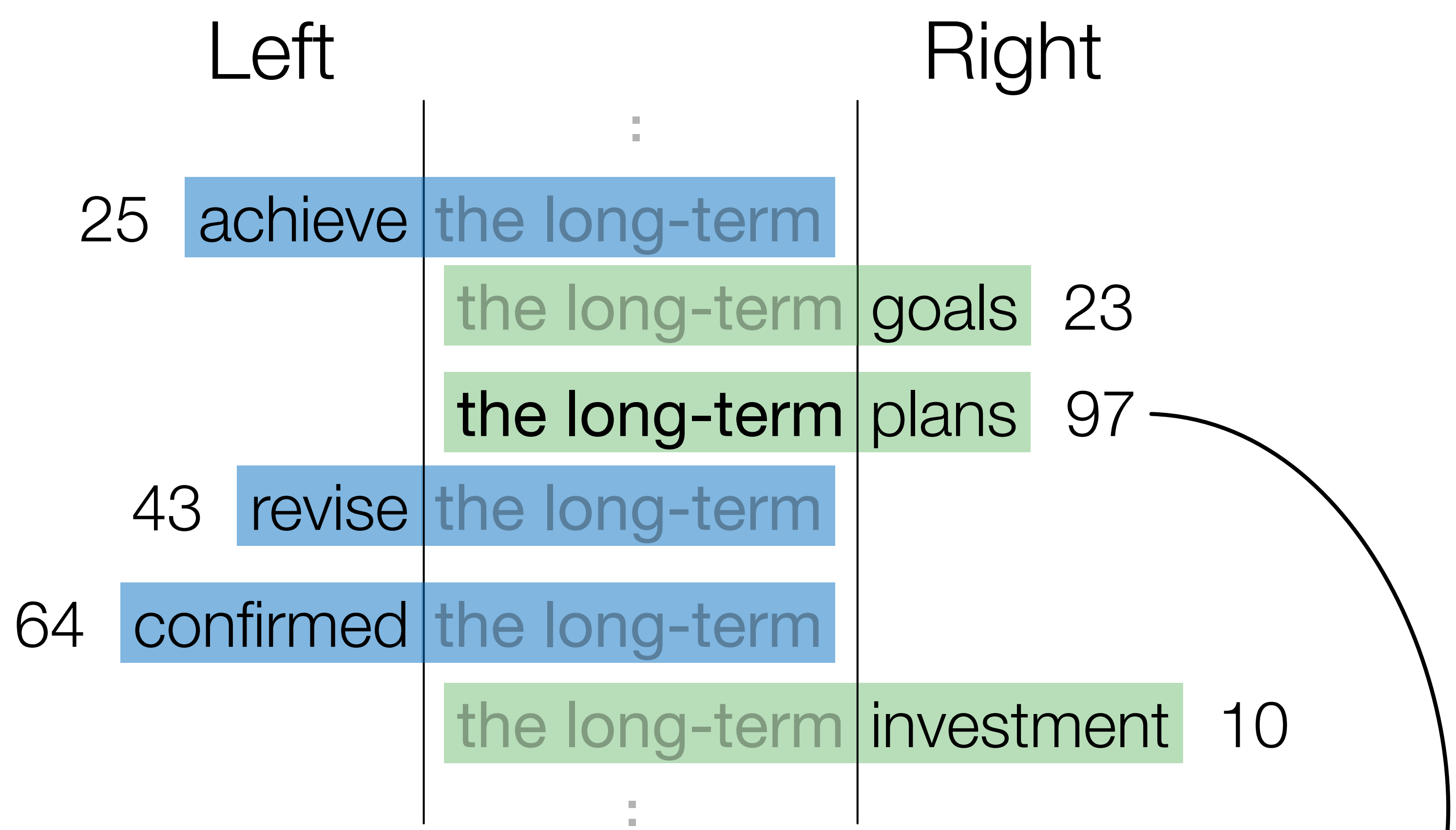


$$sim(\mathbf{r}) = \frac{1}{2}\left(sim\left(\begin{array}{c}\text{the long-term}\\ \text{in the long term}\end{array}\right) + sim\left(\begin{array}{c}\text{'s}\\ \text{of}\end{array}\right)\right)$$

Source: Chris Callison-Burch

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

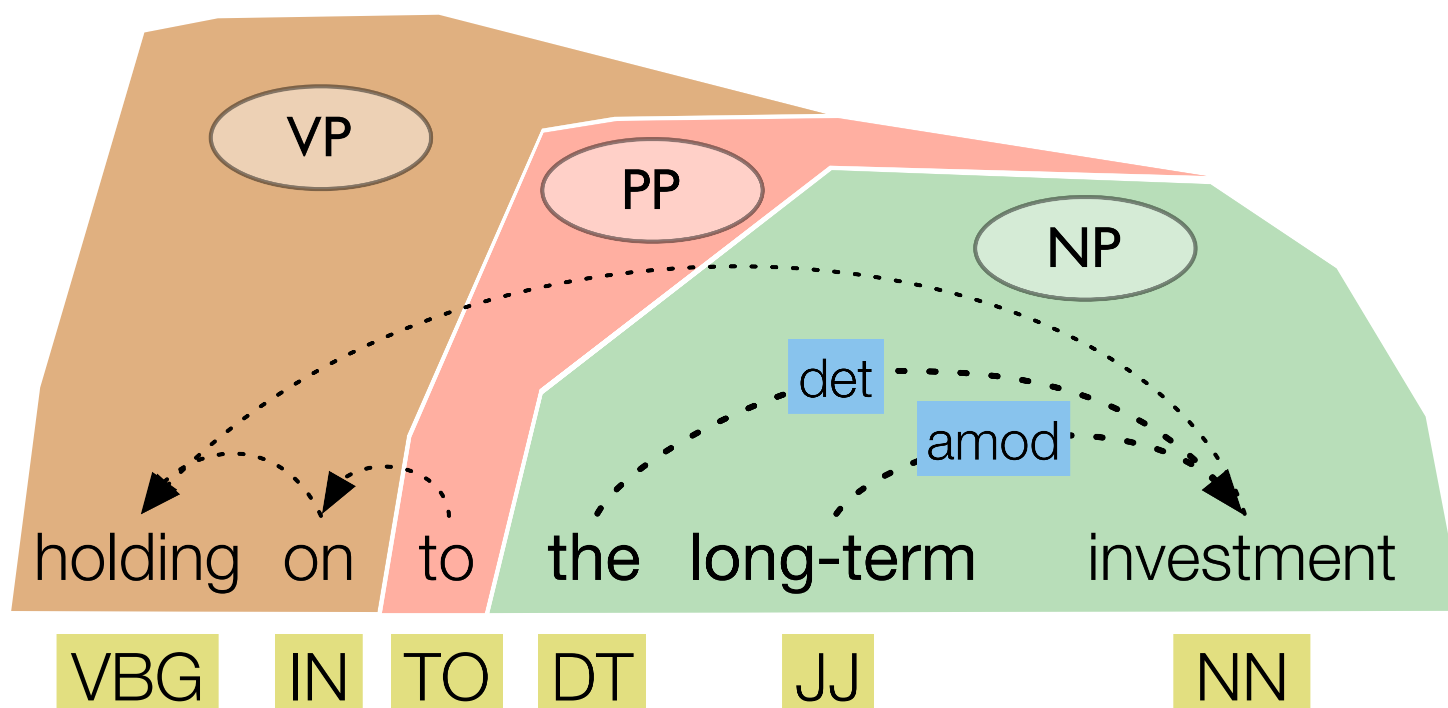
n-gram Context



$$\vec{sig}_{ngram}(\text{the long-term}) = \begin{pmatrix} \text{L-achieve} = 25 & \text{R-plans} = 97 \\ \text{L-revise} = 43 & \text{R-goals} = 23 \\ \text{L-confirmed} = 64 & \text{R-investment} = 10 \end{pmatrix}$$

Source: Chris Callison-Burch

Syntactic Context



$$\vec{sig}_{syntax}\big(\text{the long-term}\big) = \left(\begin{array}{l} \text{lex-R-investment} \quad \text{lex-L-on-to} \\ \text{pos-L-IN-TO} \quad \text{pos-L-TO} \quad \text{lex-L-to} \\ \text{dep-det-R-investment} \quad \text{pos-R-NN} \\ \text{dep-amod-R-investment} \\ \text{dep-det-R-NN} \quad \text{dep-amod-R-NN} \\ \text{syn-gov-NP} \quad \text{syn-miss-L-NN} \end{array}\right)$$

Source: Chris Callison-Burch

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

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

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

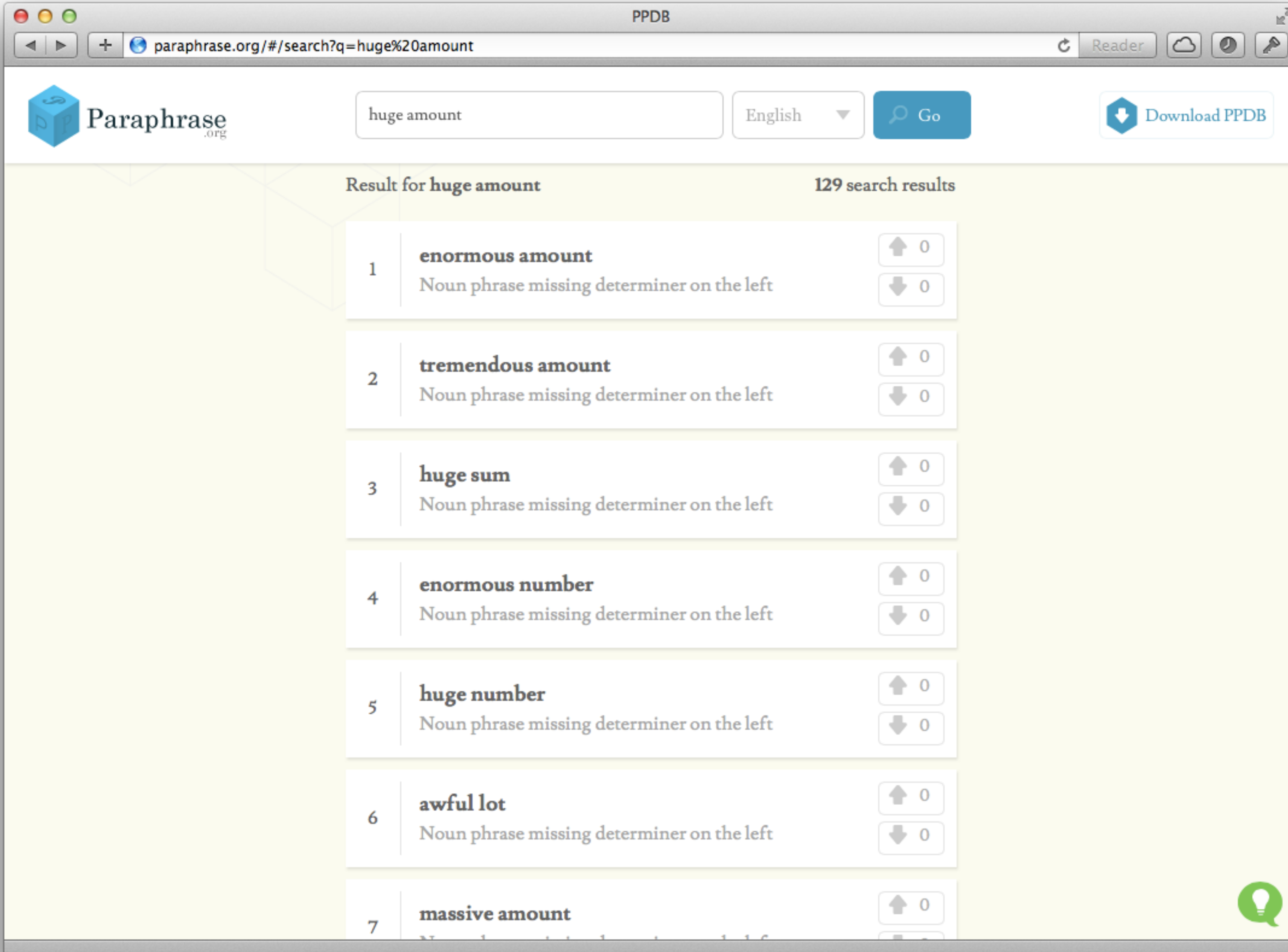
http://paraphrase.org

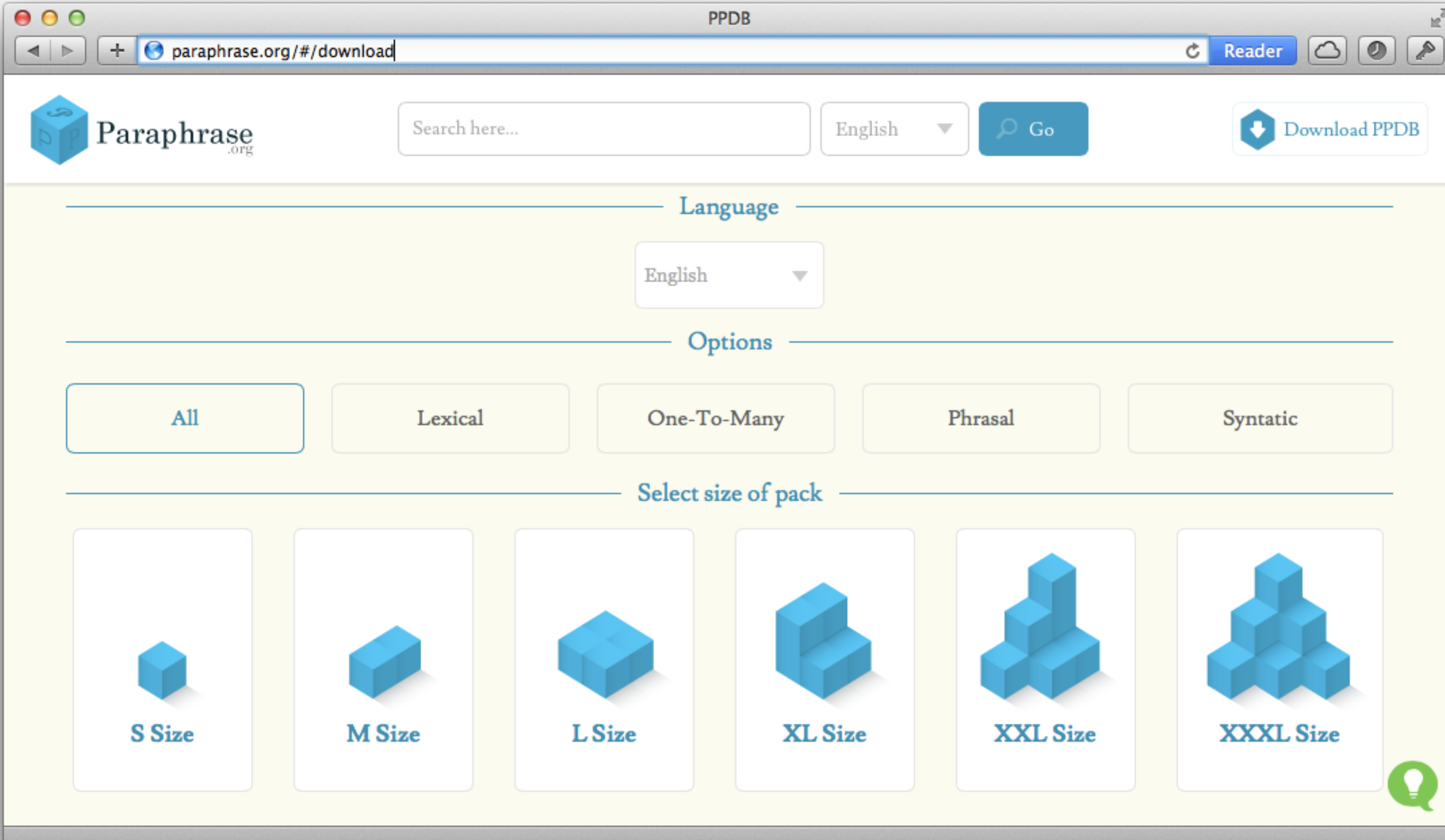
Source: Chris Callison-Burch

PPDB: The Paraphrase Database

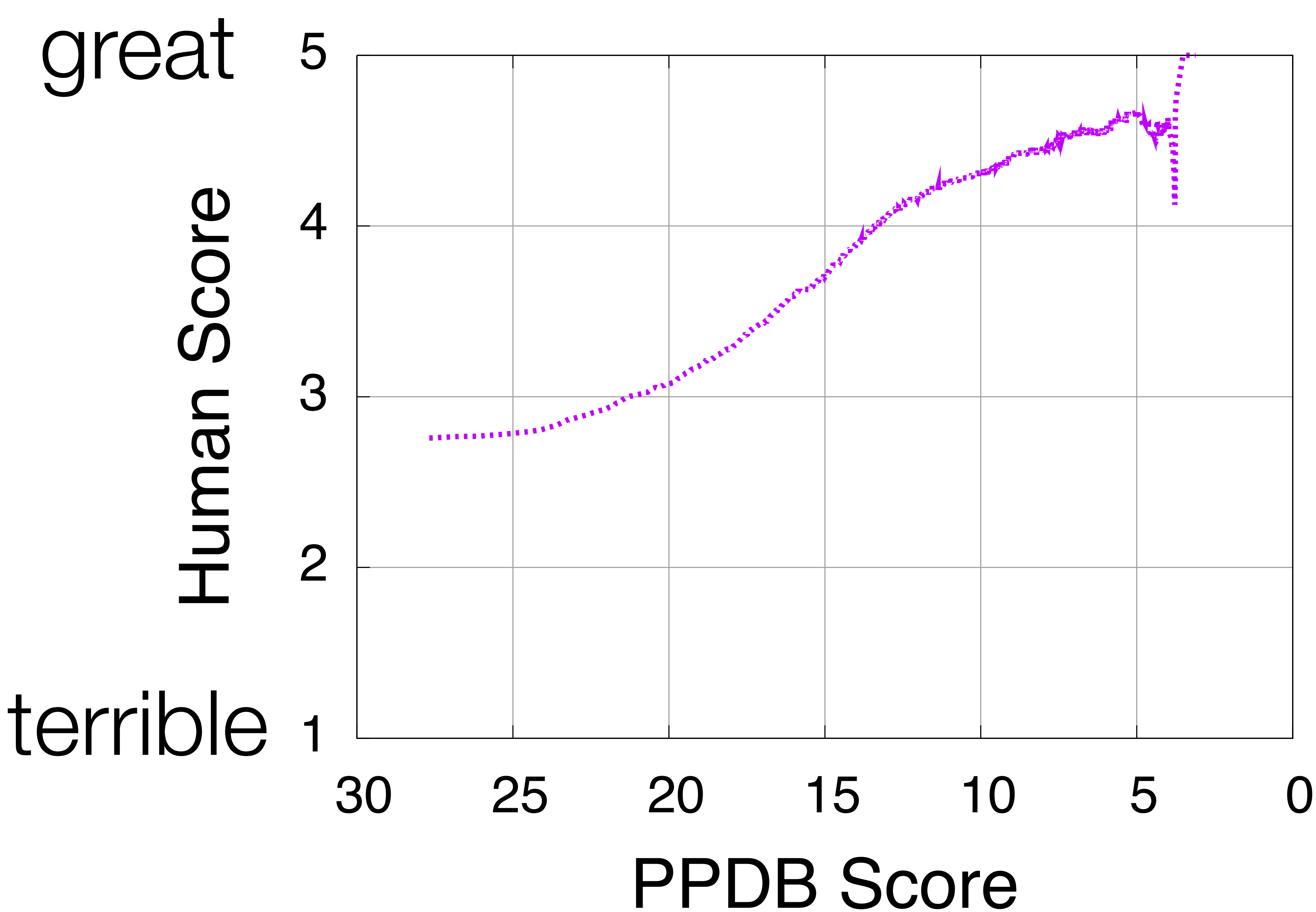
Language	Code	Number of Paraphrases			
		Lexical	Phrasal	Syntactic	Total
Arabic	Ara	119.7M	45.1M	20.1M	185.7M
Bulgarian	Bul	1.3M	1.4M	1.2M	3.9M
Czech	Ces	7.3M	2.7M	2.6	12.1M
German	Deu	7.9M	15.4M	4.9M	28.3M
Greek	Ell	5.4M	9.4M	7.4M	22.3M
Estonian	Est	7.9M	1.0M	0.4M	9.2M
Finnish	Fin	41.4M	4.9M	2.3M	48.6M
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


PPDB
paraphrase.org/#/search?q=huge%20amount
Reader
Paraphrase
.org
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

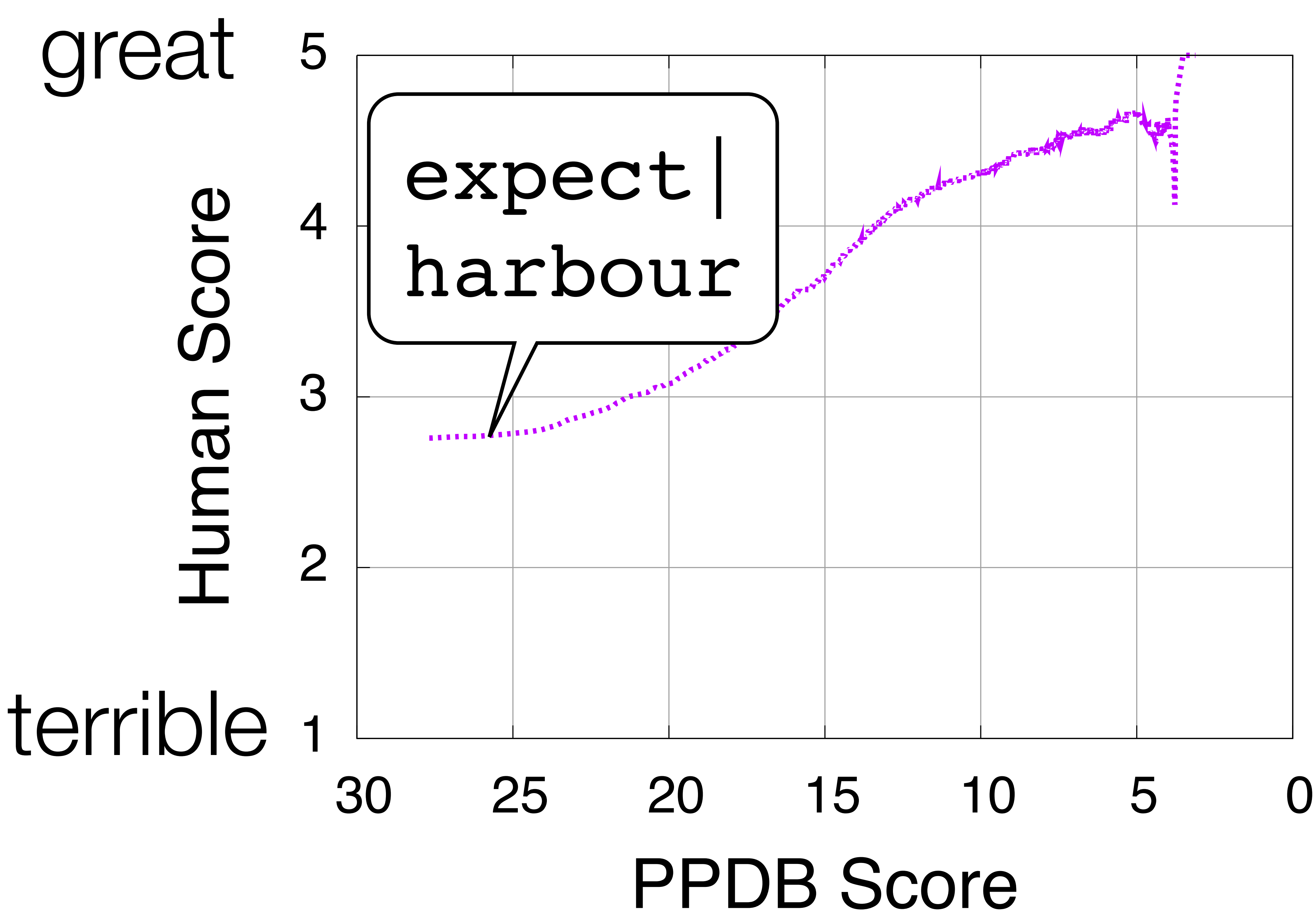

PPDB
paraphrase.org/#/download
Reader
Paraphrase
.org
Search here...
English
Go
Download PPDB
Language
English
Options
All
Lexical
One-To-Many
Phrasal
Syntatic
Select size of pack
S Size
M Size
L Size
XL Size
XXL Size
XXXL Size

Do the Scores Work?



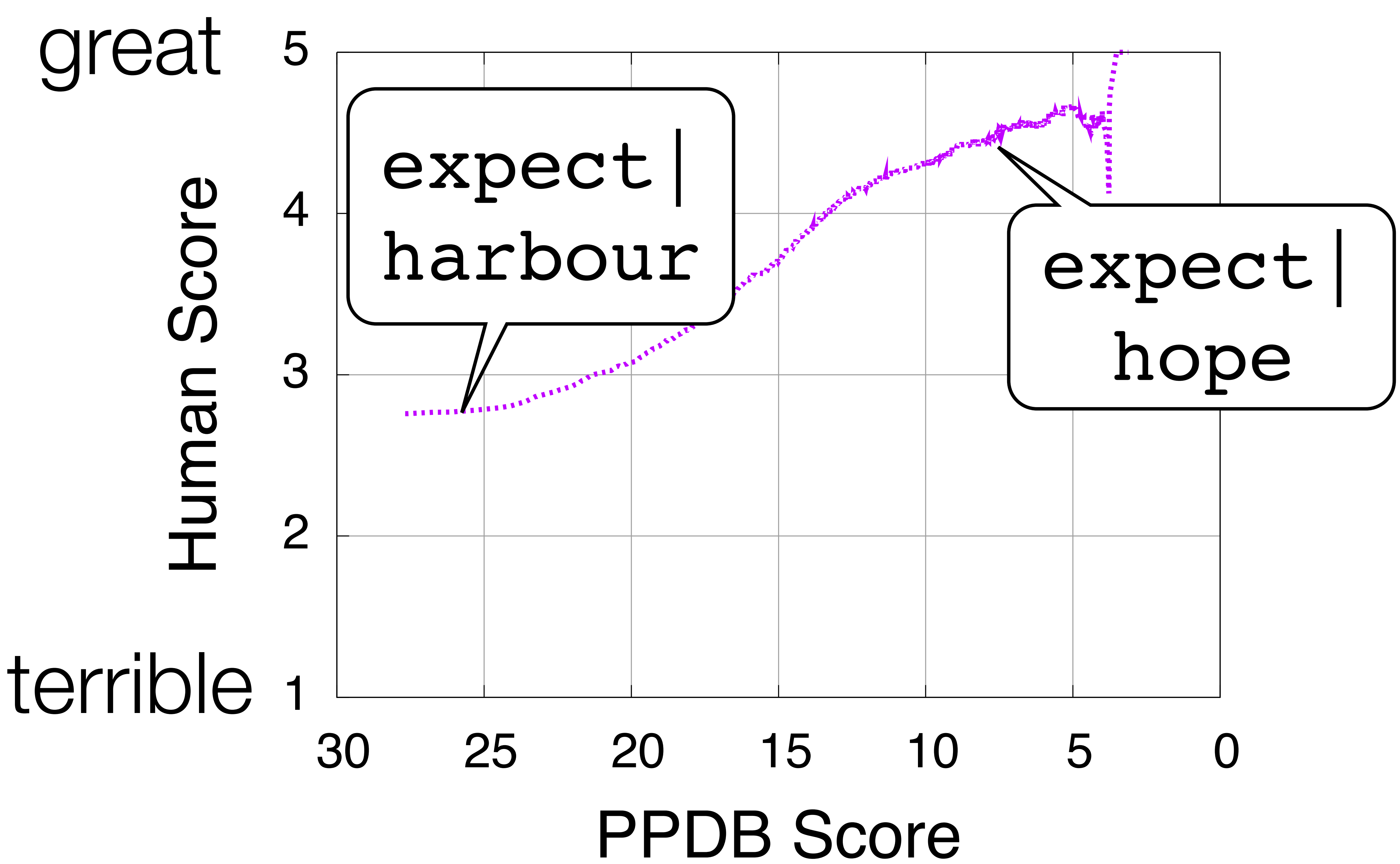
Source: Chris Callison-Burch

Do the Scores Work?



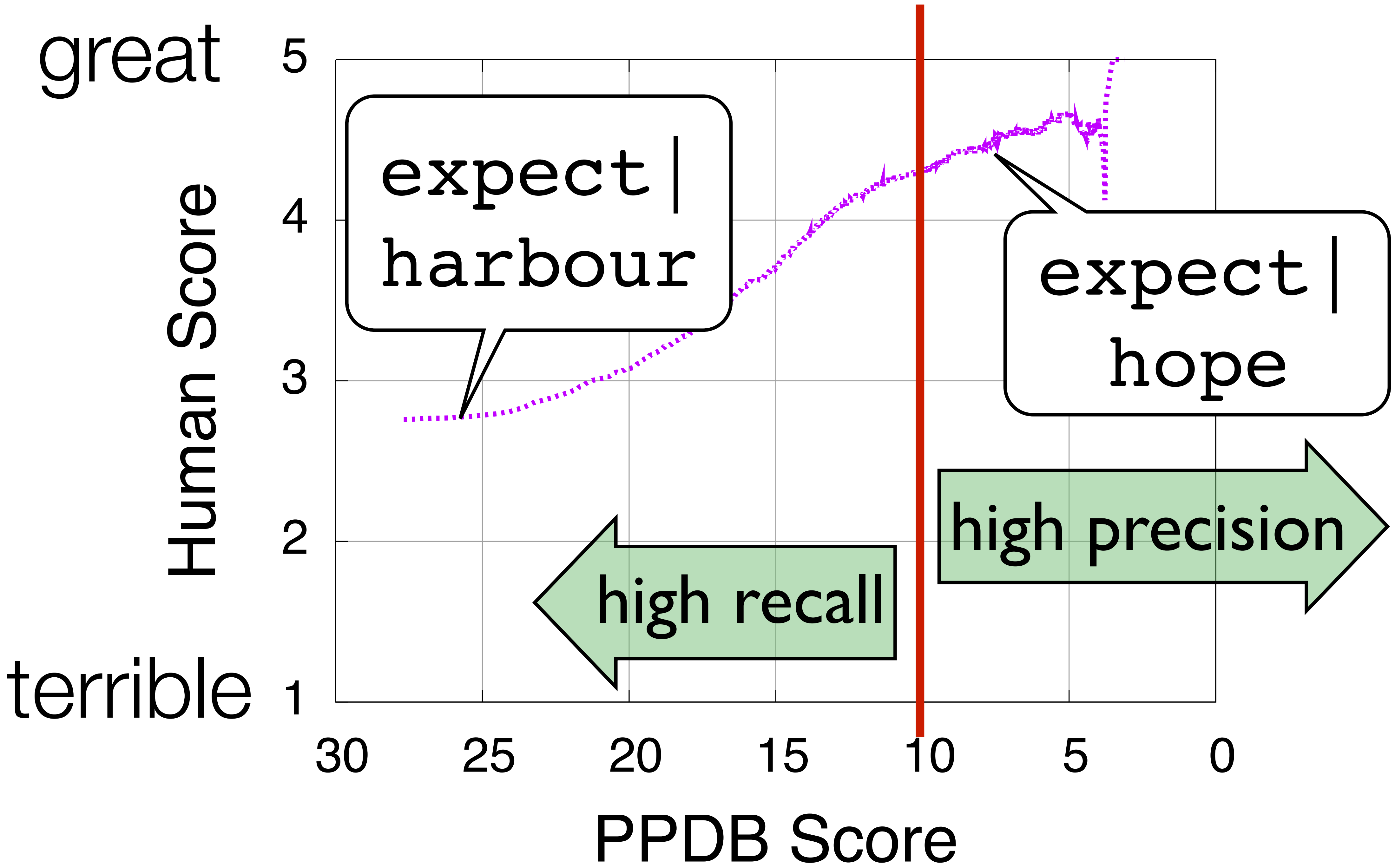
Source: Chris Callison-Burch

Do the Scores Work?



Source: Chris Callison-Burch

Do the Scores Work?



Source: Chris Callison-Burch

Fun PPDB Examples

PARENTAL
ADVISORY
EXPLICIT CONTENT

Source: Chris Callison-Burch

Juri Ganitkevitch, Ben Van Durme and Chris Callison-Burch. PPDB: The Paraphrase Database. NAACL 2013.

Fun PPDB Examples

`munchies ||| hungry`

`hustle ||| scam`

`sexiest ||| hottest`

`dummies ||| losers`

`sheeit ||| dammit`

`abso-fucking-lutely ||| indeed`

Source: Chris Callison-Burch

Juri Ganitkevitch, Ben Van Durme and Chris Callison-Burch. PPDB: The Paraphrase Database. NAACL 2013.

Pivoting w/ Neural MT

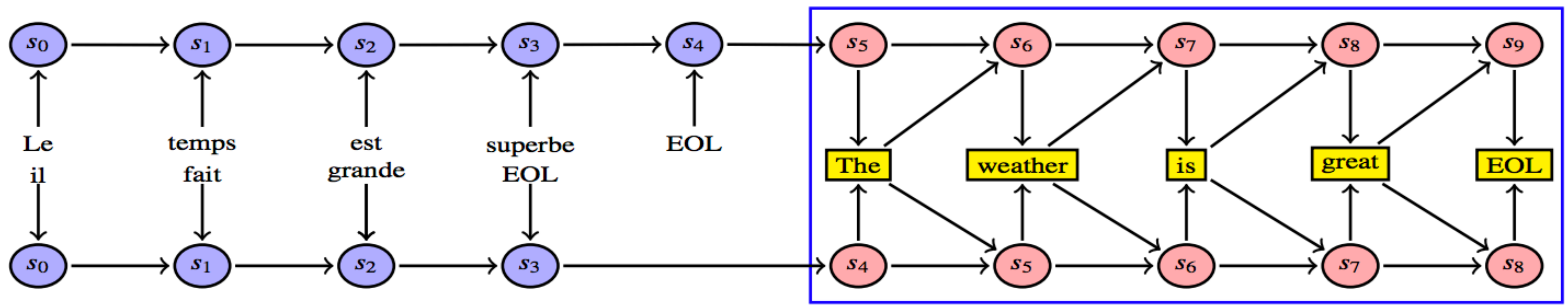


Figure 1: Late-weighted combination: two pivot sentences are simultaneously translated to one target sentence. Blue circles indicate the encoders, which individually encode the two source sentences. After the EOL token is seen, decoding starts (red circles). At each time step the two decoders produce 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.

Jonathan Mallinson, Rico Sennrich and Mirella Lapata. Paraphrasing Revisited with Neural Machine Translation. EACL 2017.

Pivoting w/ Neural MT

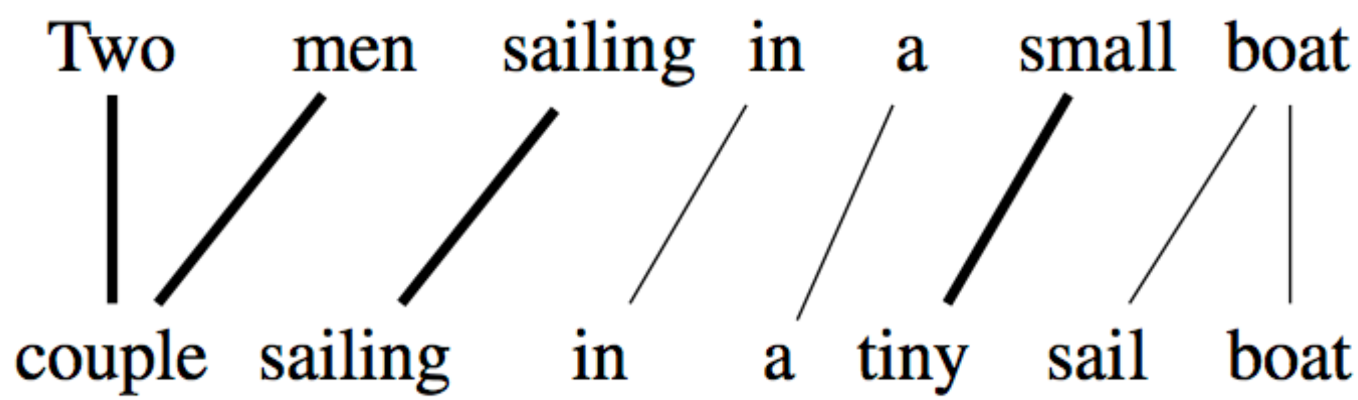


Figure 2: Attention between two sentences. Line thickness indicates the strength of the attention.

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

Neural MT tutorial: https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/

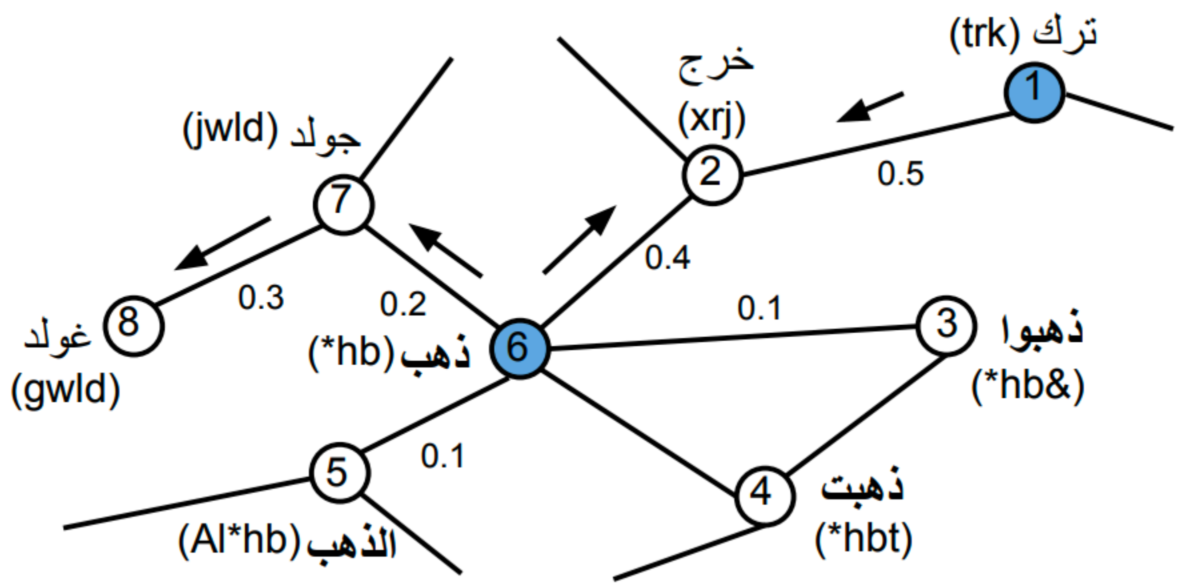
Jonathan Mallinson, Rico Sennrich and Mirella Lapata. Paraphrasing Revisited with Neural Machine Translation. EACL 2017.

Improve MT w/ PPDB

Phrase table

Source	Target	P (e | f)
ذهب	gold	0.5
ذهب	went	0.2
ذهب	be	0.1
ذهب	left	0.2
ترک	leave	0.5
ترک	left	0.2
ترک	drop	0.1
ترک	dropped	0.2

...



Paraphrase pairs list

Source1	Source 2	Score
ذهب	خرج	0.4
ذهب	ذهبت	0.2
ذهب	ذهبوا	0.1
ذهب	الذهب	0.1
ذهب	جولد	0.2
خرج	ترک	0.5
خرج	ذهب	0.4
جولد	ذهب	0.2
جولد	غولد	0.3

...

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.

Ramtin Mehdizadeh Seraj, Maryam Siahbani, Anoop Sarkar. Improving Statistical Machine Translation with a Multilingual Paraphrase Database. EMNLP 2015.

Guest Lecture next week

- **Jeniya Tabassum (OSU)**
- Time Expressions in Twitter

TweeTime: A Minimally Supervised Method for Recognizing and Normalizing Time Expressions in Twitter

Jeniya Tabassum, Alan Ritter and Wei Xu
Computer Science and Engineering
Ohio State University
{bintejafar.1,ritter.1492,xu.1265}@osu.edu

Abstract

We describe TweeTIME, a temporal tagger for recognizing and normalizing time expressions in Twitter. Most previous work in social media analysis has to rely on temporal resolvers that are designed for well-edited text, and therefore suffer from reduced performance due to domain mismatch. We present a minimally supervised method that learns from

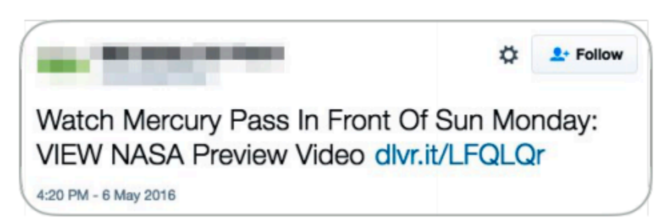


Figure 1: A tweet published on *Friday 5/6/2016* that contains the temporal expression *Monday* referring to the date of the event (*5/9/2016*), which a generic temporal tagger failed to resolve correctly.

socialmedia-class.org