Social Media & Text Analysis

lecture 4 - Paraphrase Data Sources

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?

what is Paraphrase?

"sentences or phrases that convey approximately the same meaning using different words" — (Bhagat & Hovy, 2012)

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 ?

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

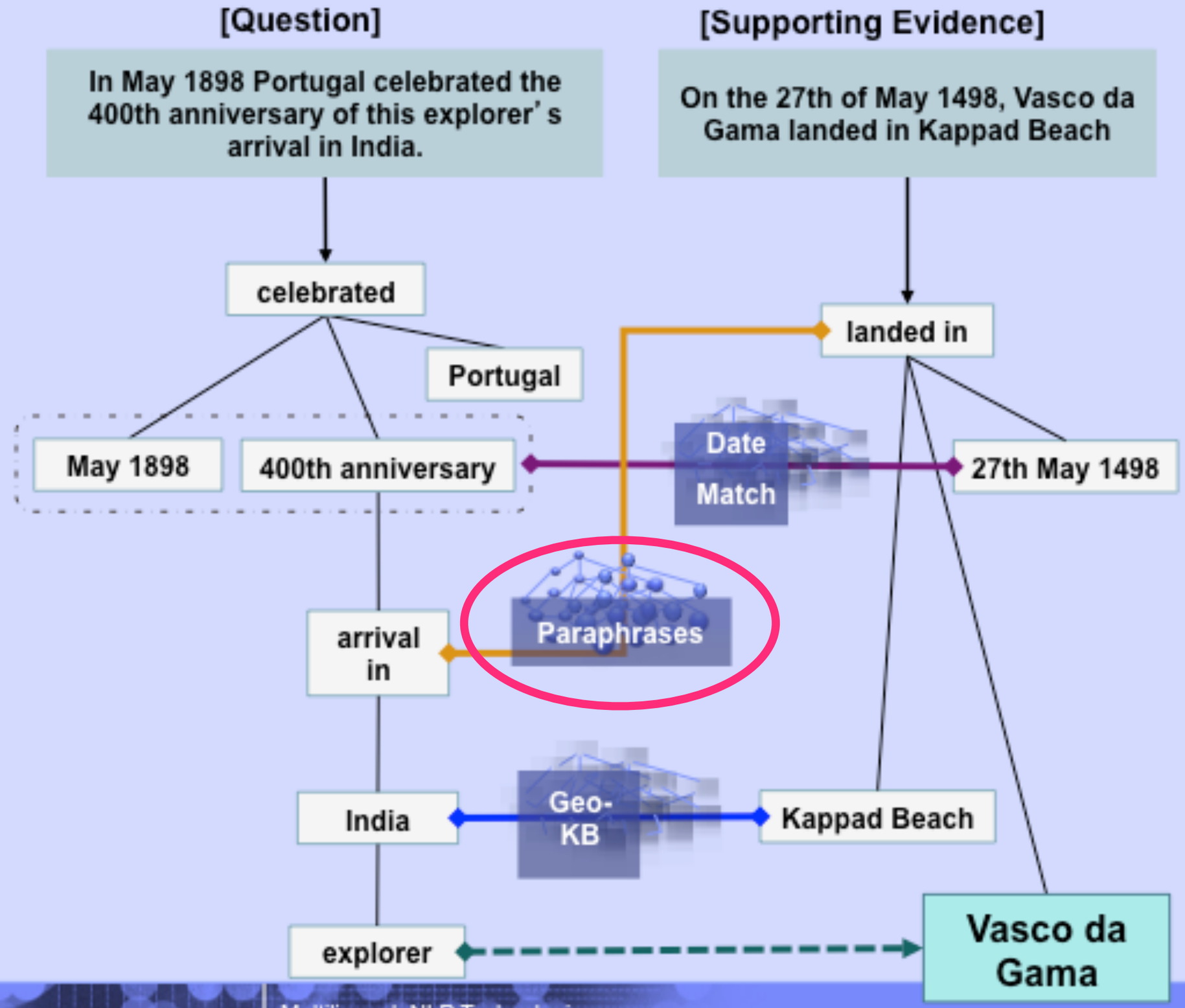
Who is the CEO stepping down from Boeing?

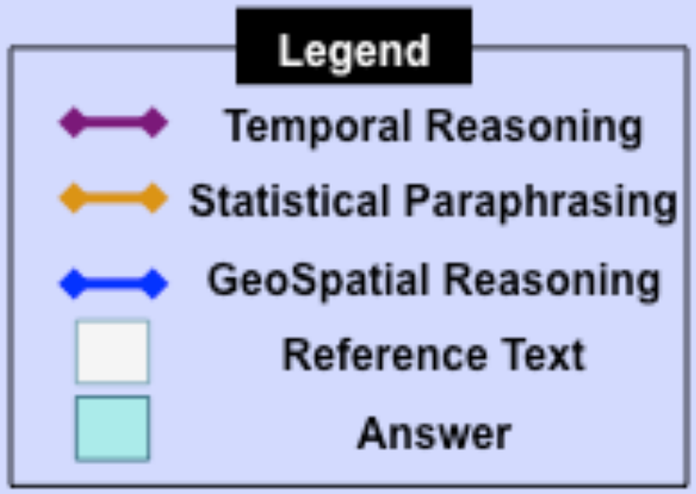
match

… 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





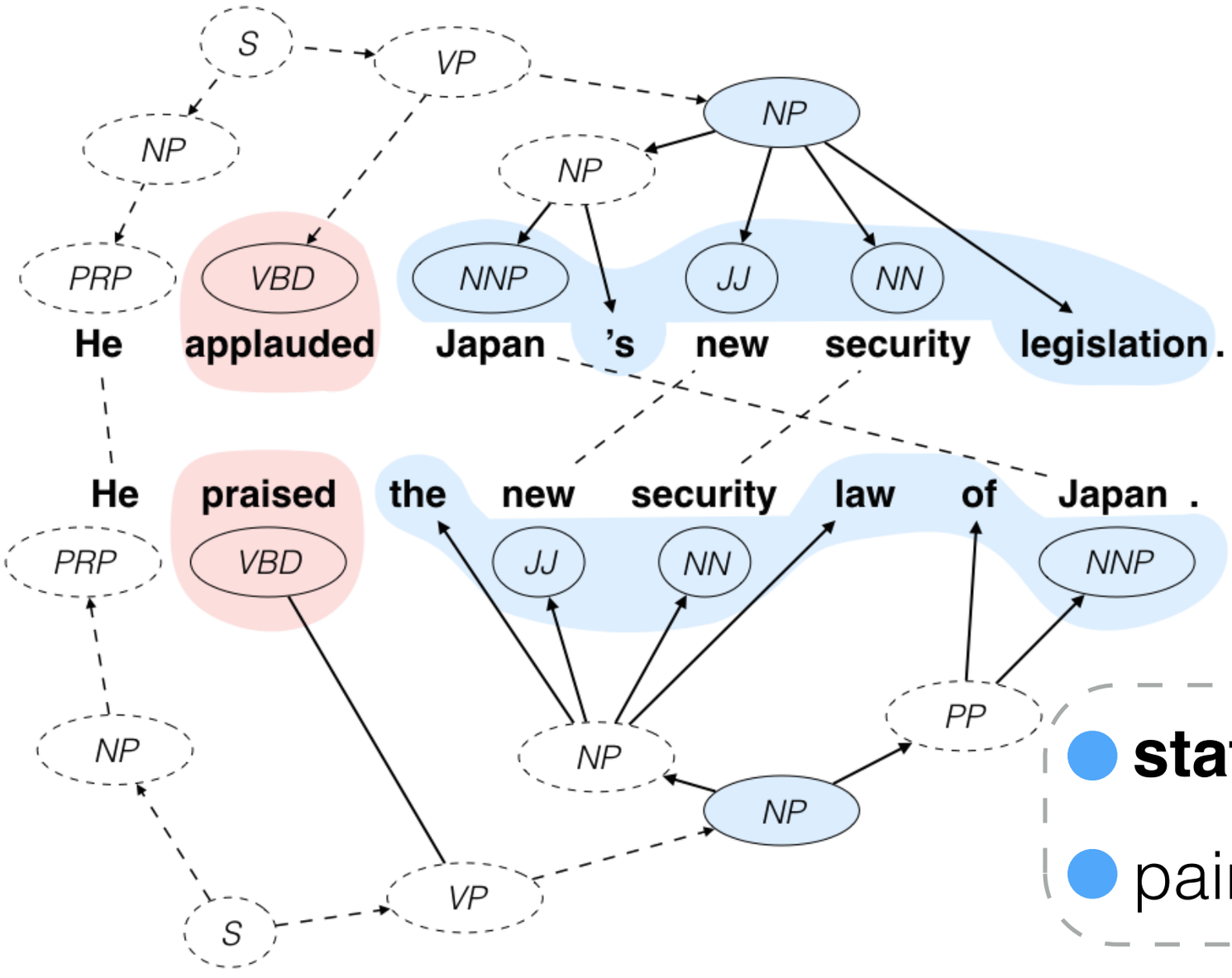
Stronger evidence can be much harder to find and score…

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

(courtesy: Salim Roukos)

Natural Language Generation

e.g. Text Simplification



Techniques

- **statistical 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!

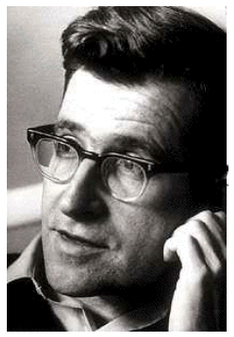
Luke:
Father, please! Help me!

↓

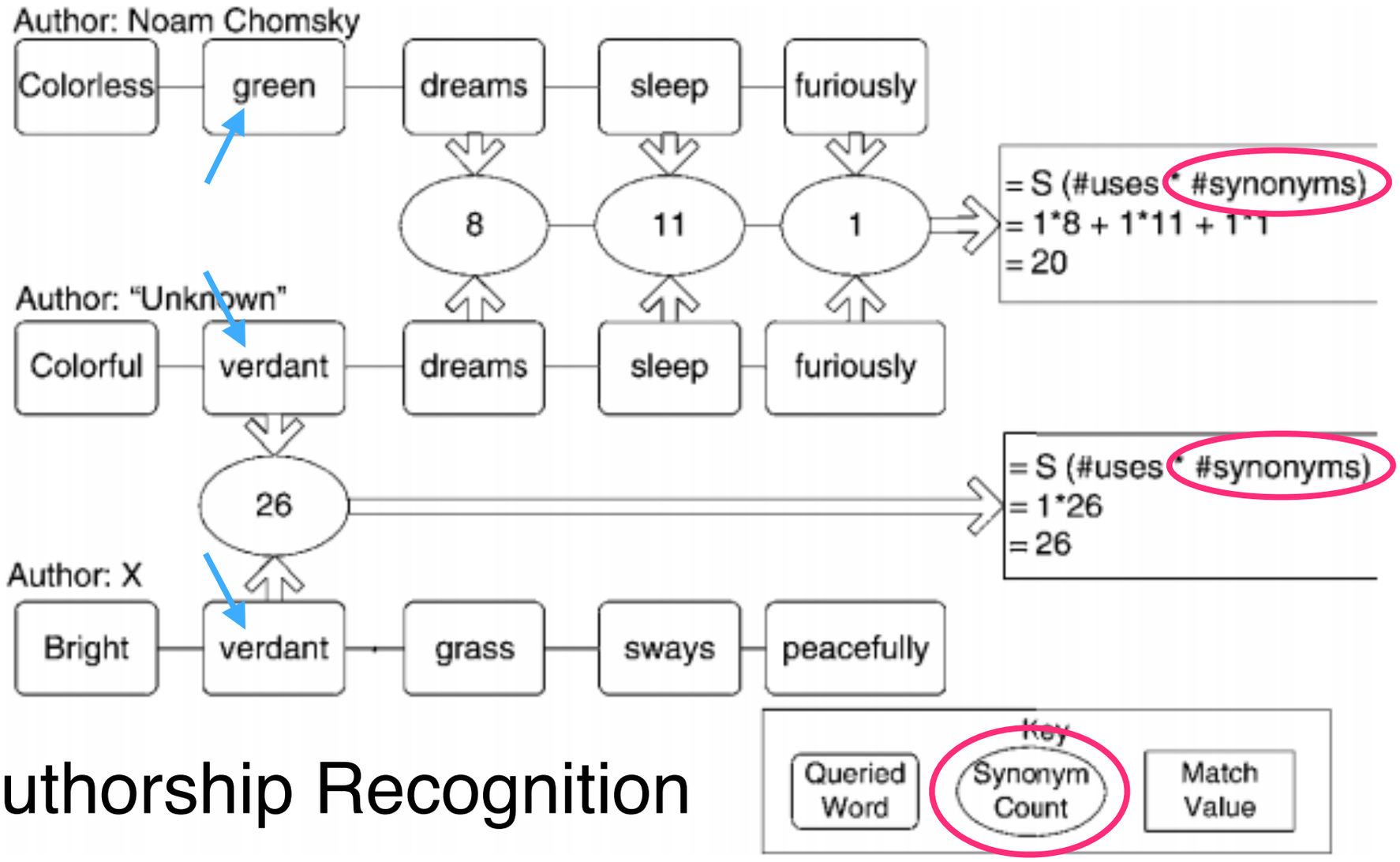
Father, I pray you! Help me!

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



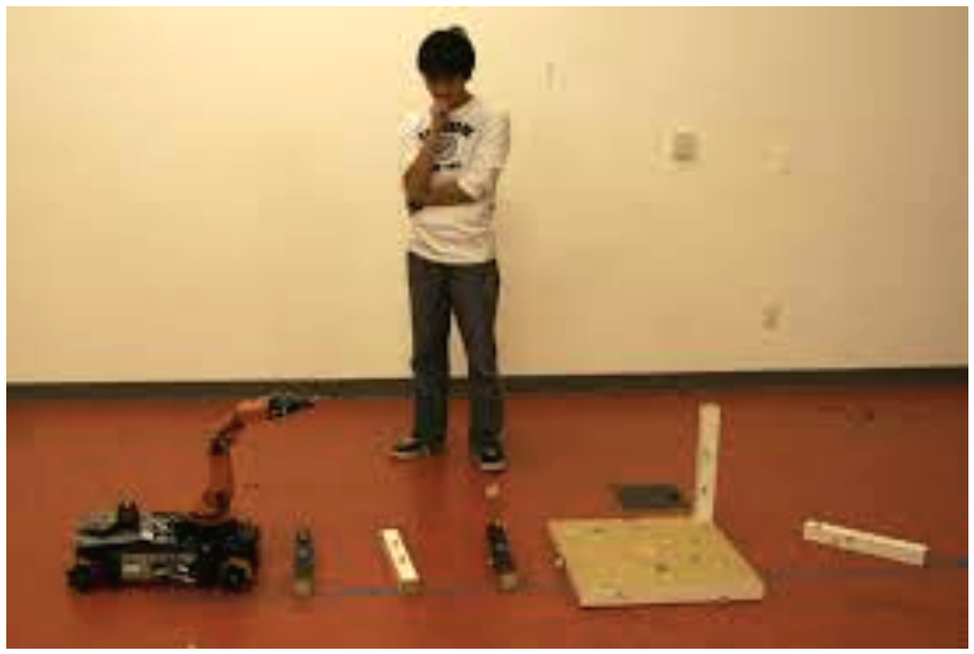
Paraphrases!!



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.

Paraphrases!!

Paraphrases!!

(Tellex et al. 2014)

Other Applications

fundamentally useful for a wide range of applications

semantic similarity *

machine translation *

summarization *

social science *

information extraction *

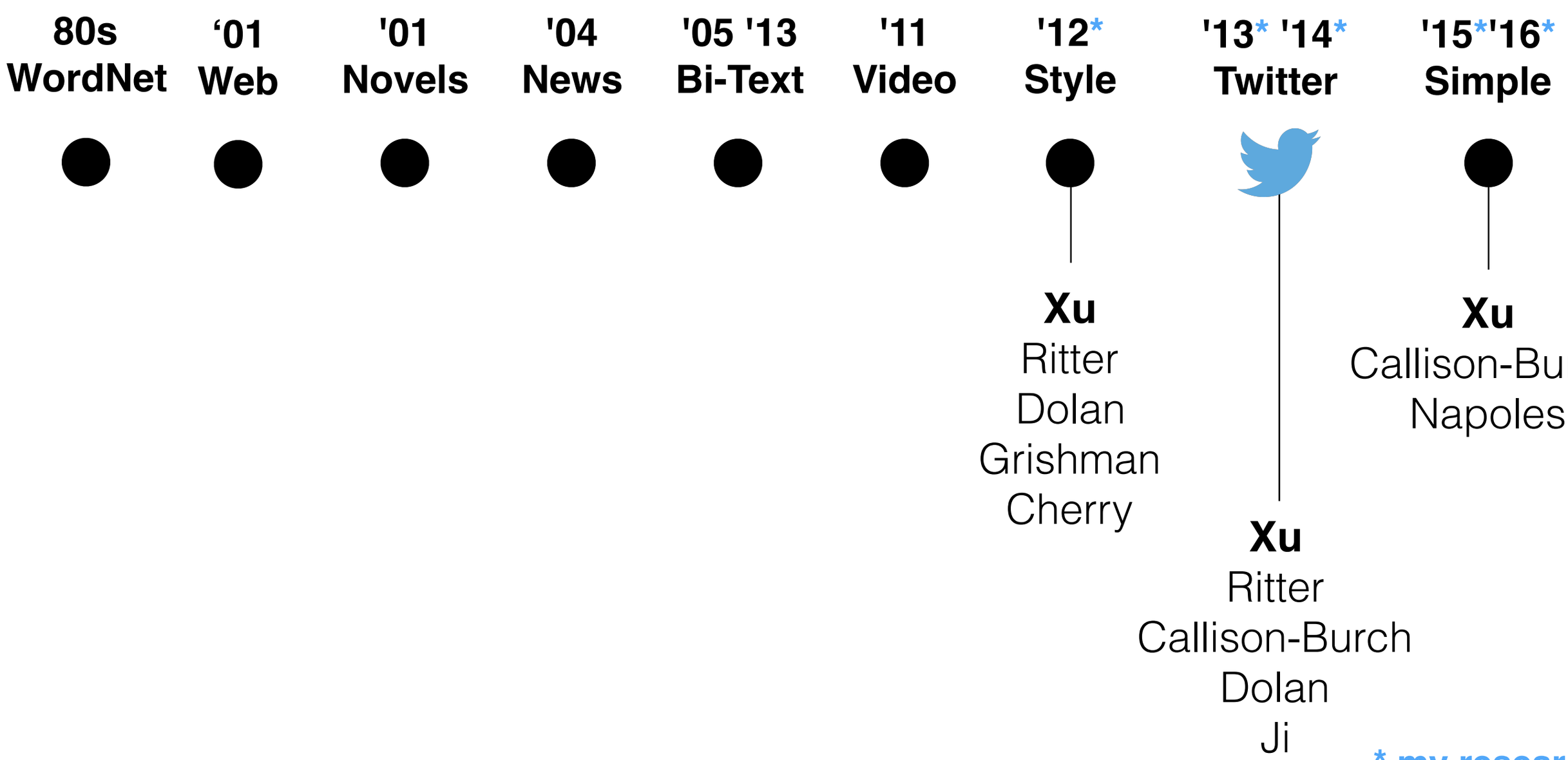
information retrieval *

semantic parsing

…

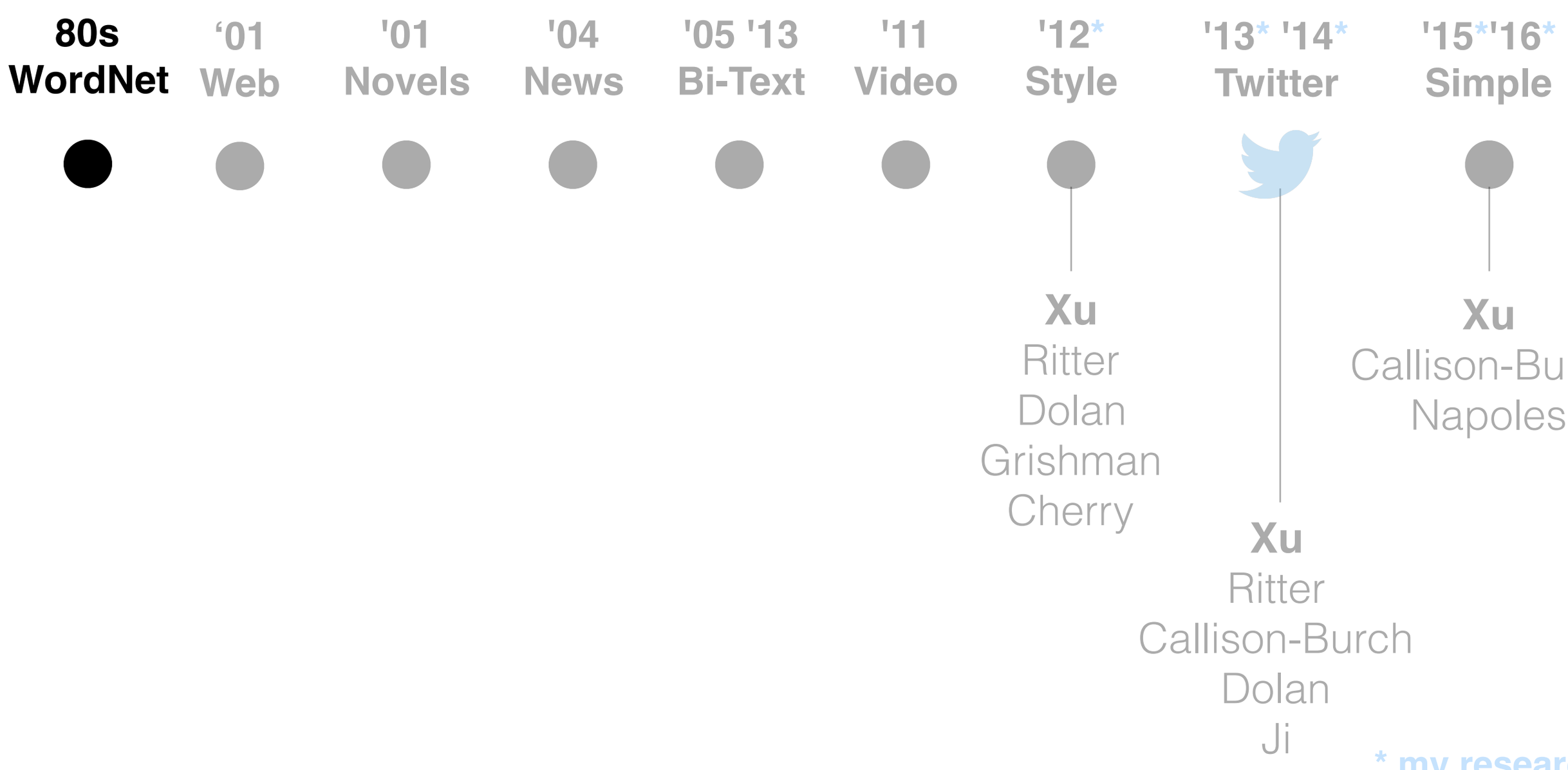
*** my research**

Paraphrase Research



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

Paraphrase Research



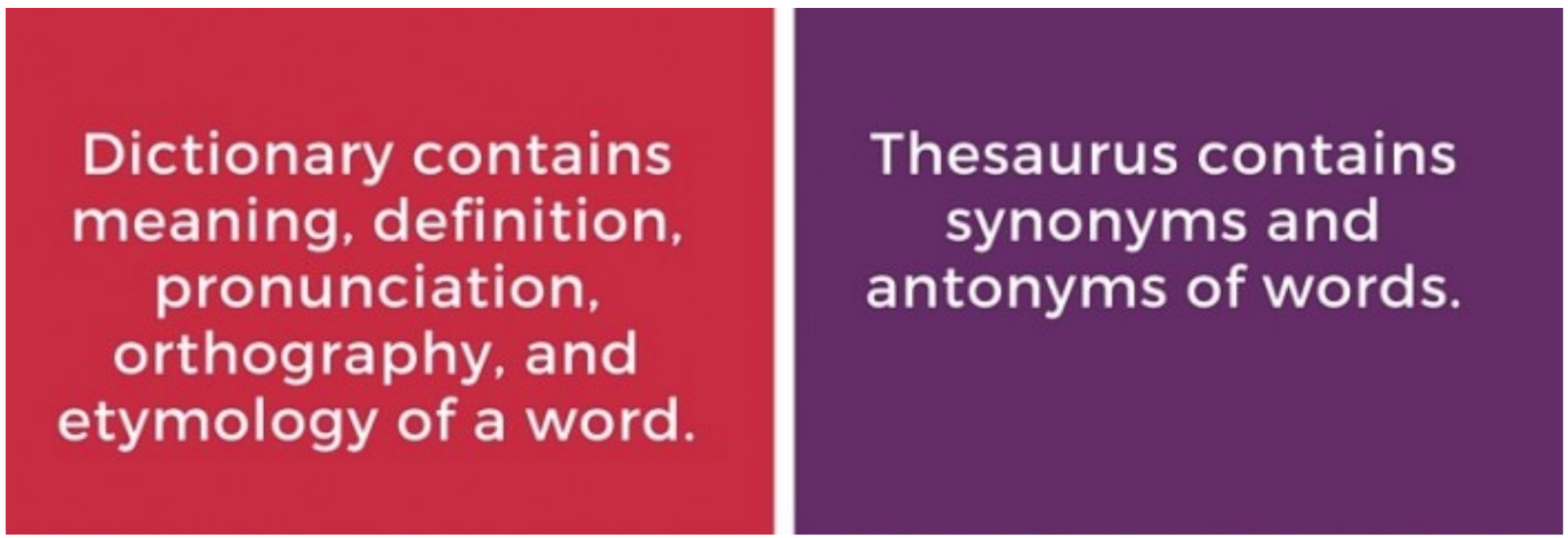
WordNet®

- 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

(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®

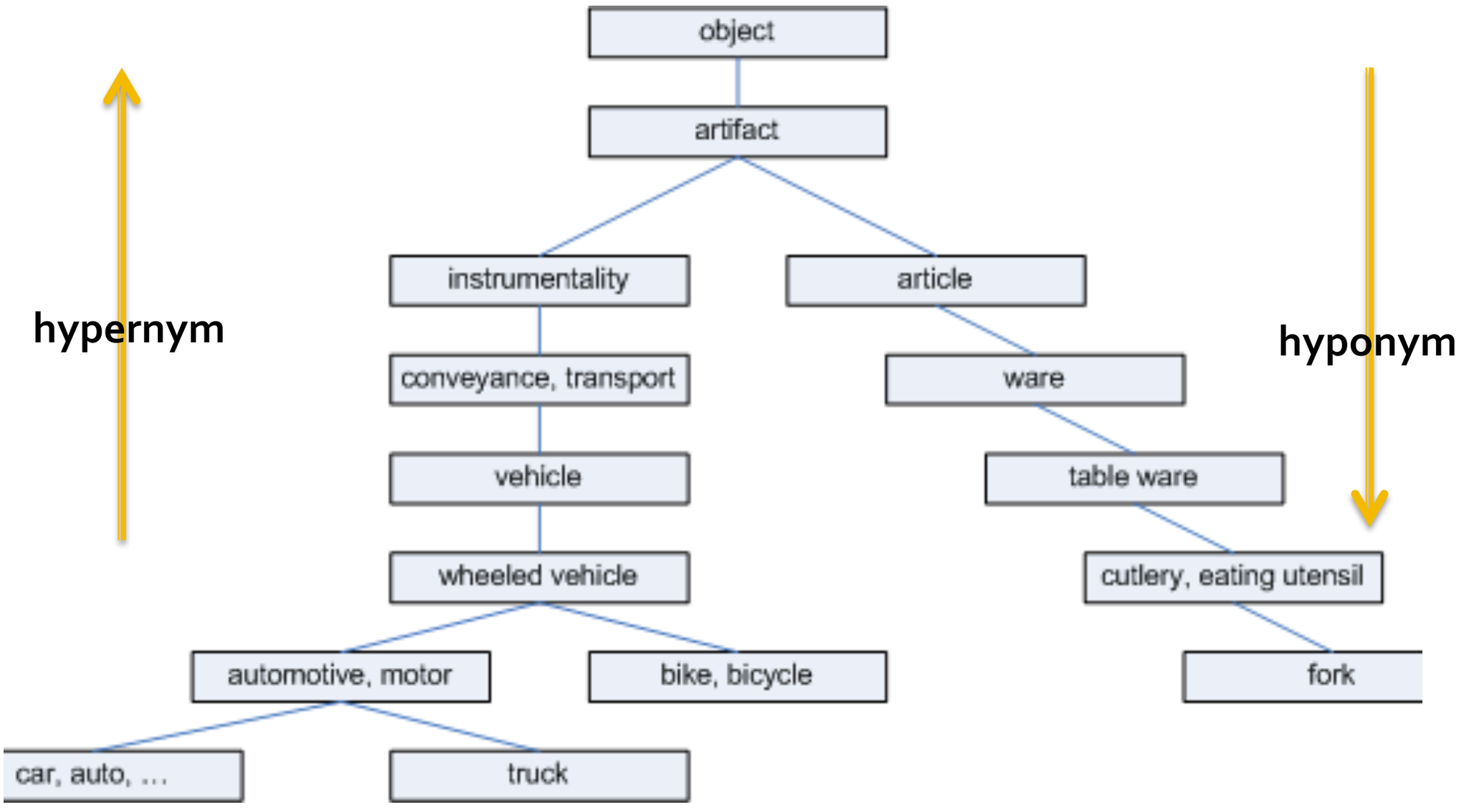
- 4 types of Parts of Speech (POS)
  - Noun, Verb, Adjective, Adverb
- Synset (synonym set)
  - the smallest unit in WordNet
  - represents a specific meaning of a word

- 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®

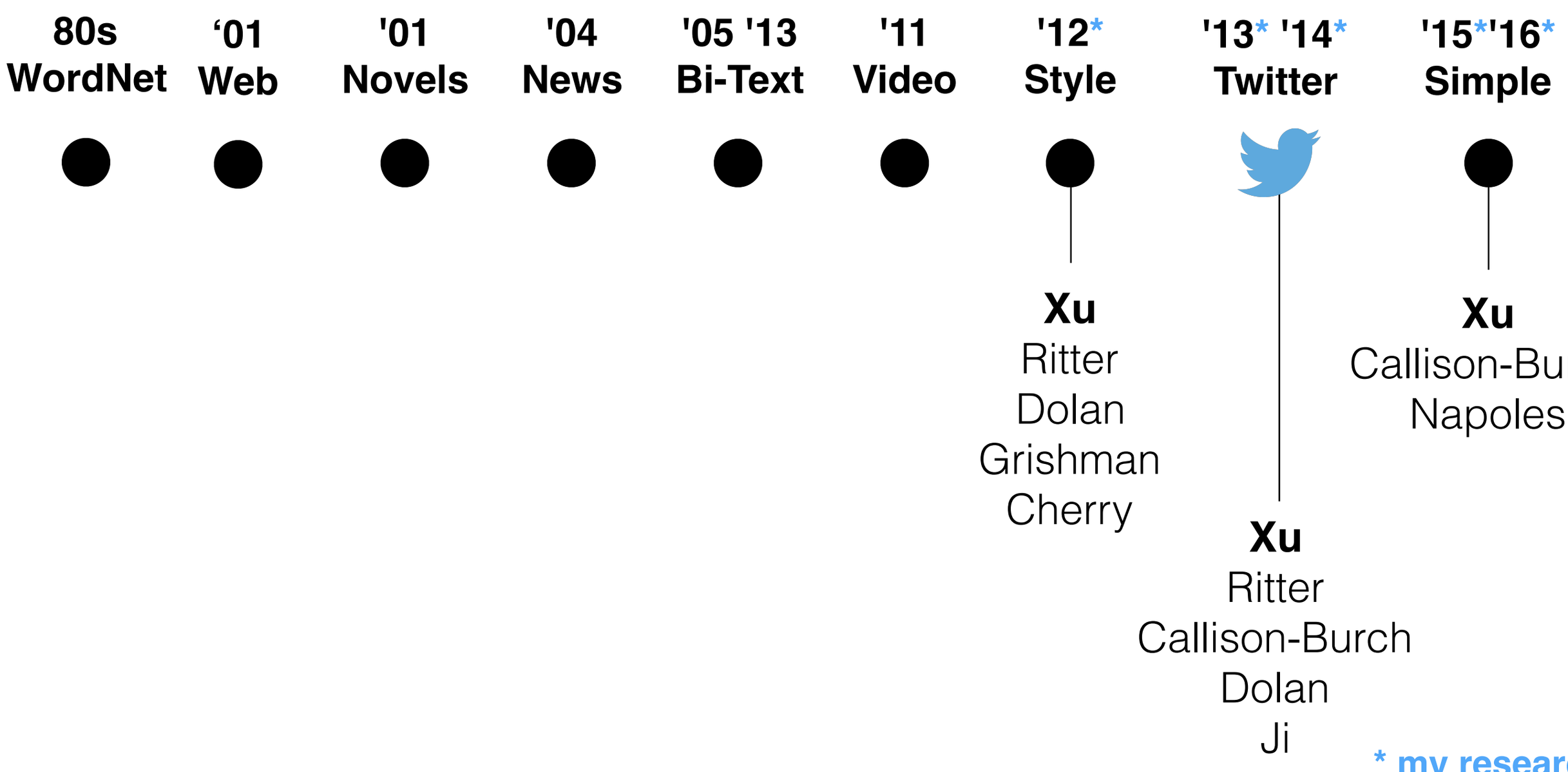

hypernym
hyponym
object
artifact
instrumentality
article
conveyance, transport
ware
vehicle
table ware
wheeled vehicle
cutlery, eating utensil
automotive, motor
bike, bicycle
fork
car, auto, …
truck

WordNet®

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

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

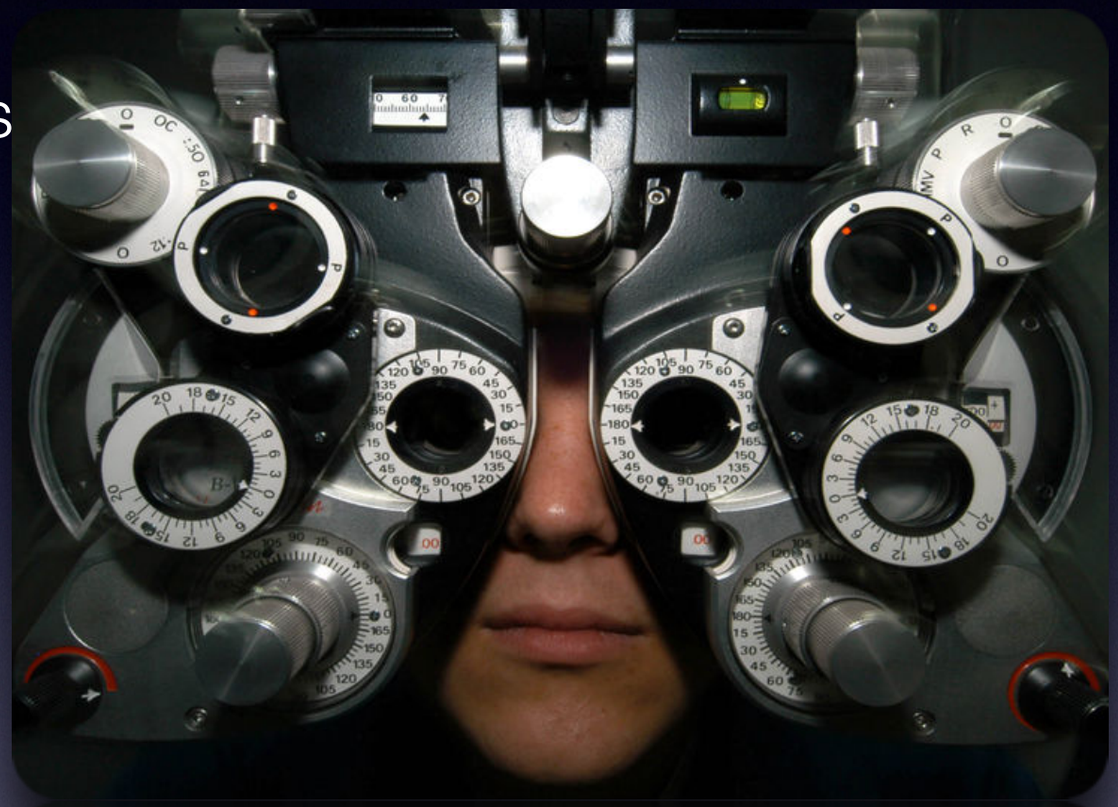
Paraphrase Research



Paraphrase Research

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 Panel (2001) operationalize the Distributional Hypothesis using **dependency relationships** to define **similar environments** (mutual information).

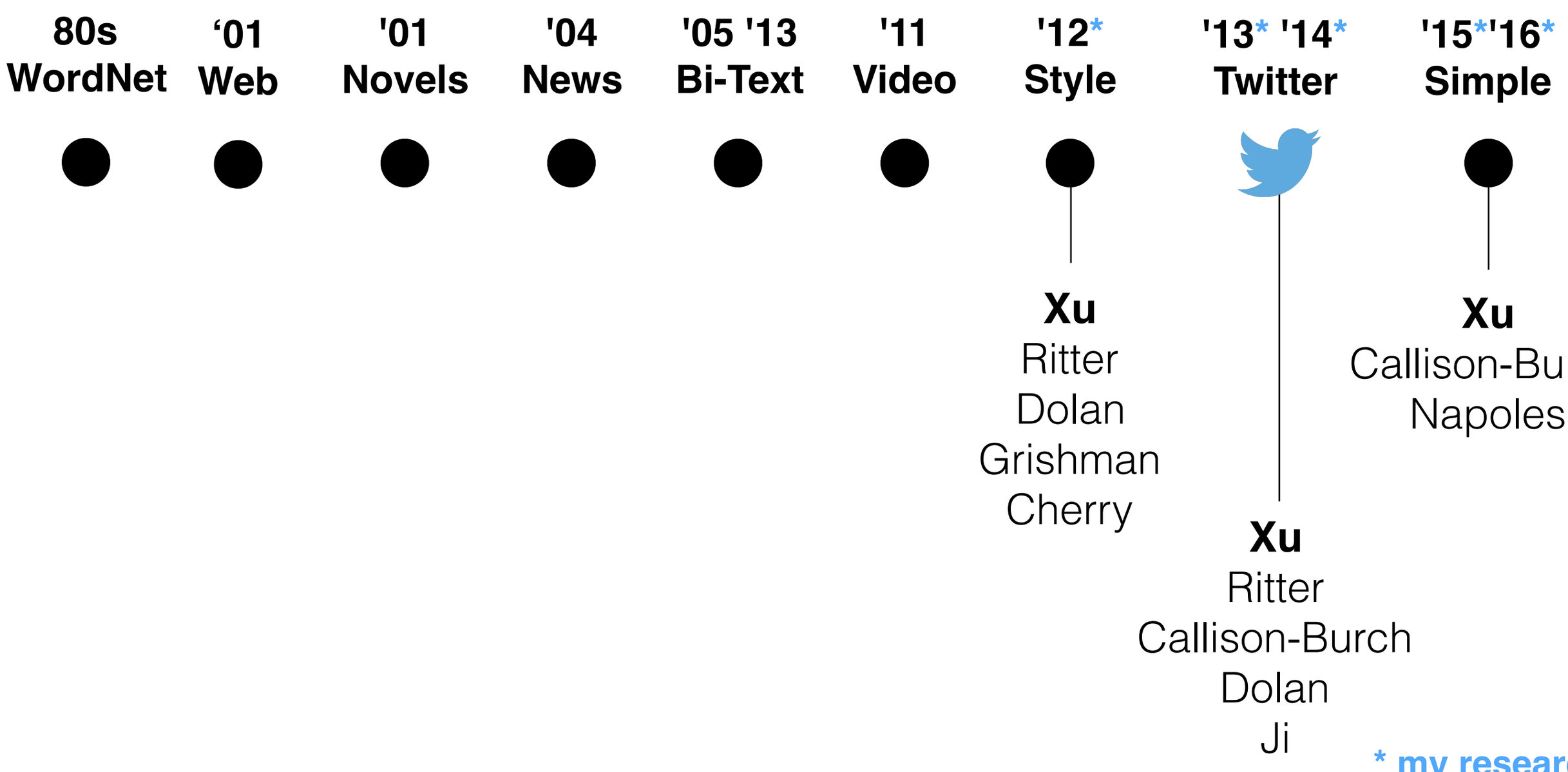
Duty and **responsibility** share a similar set of dependency contexts in large volumes of text:

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

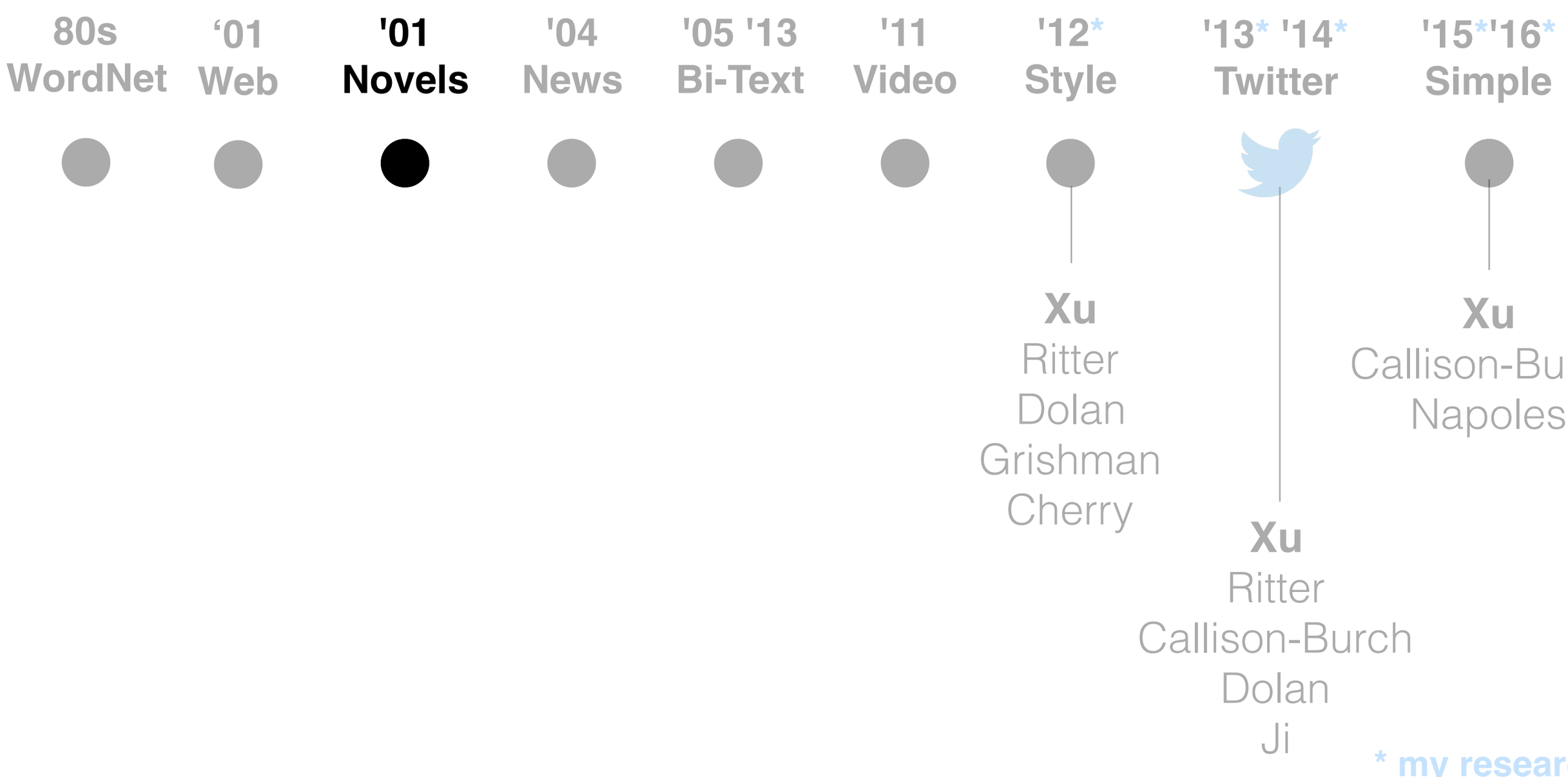
Source: Chris Callison-Burch

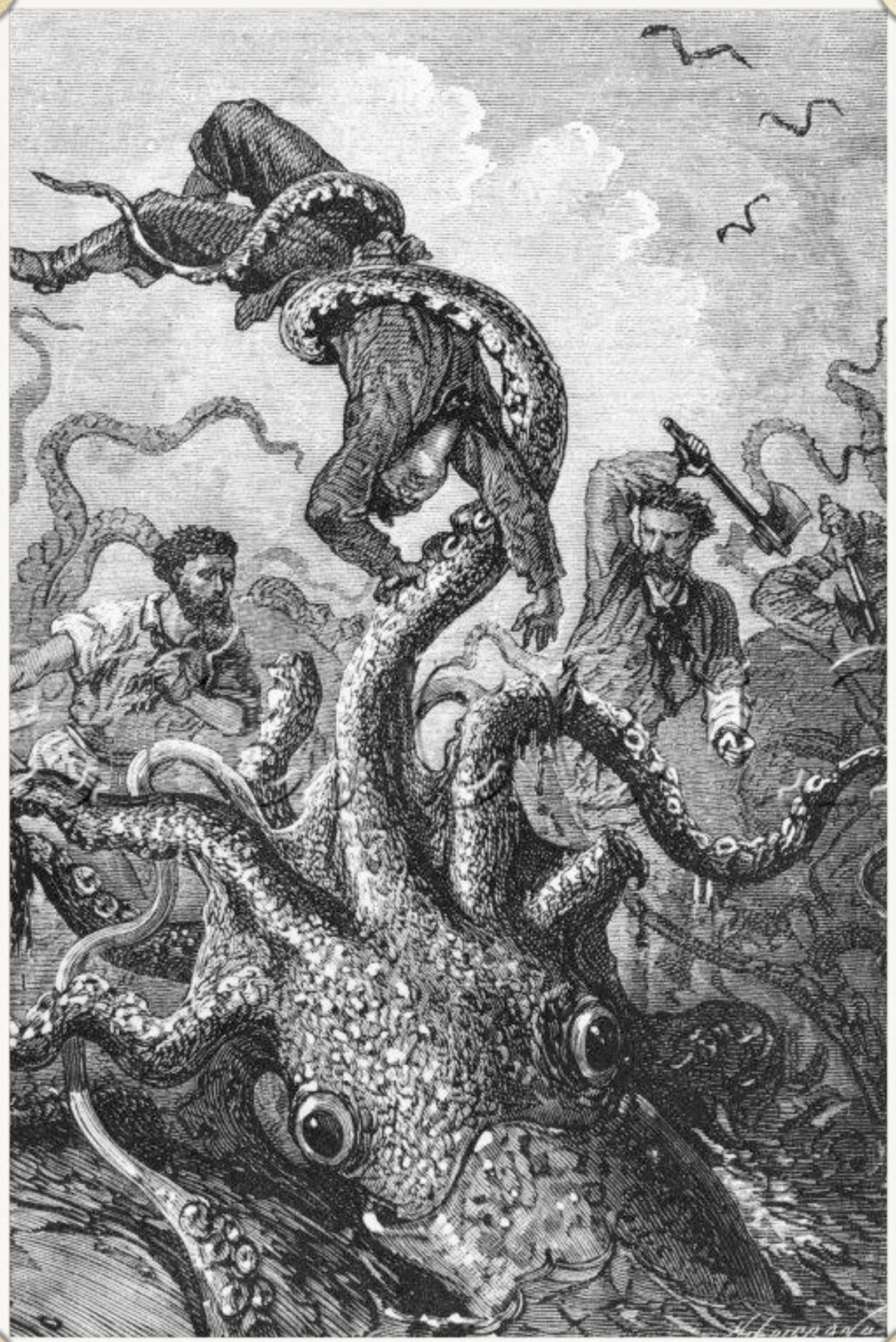
Decking Lin and Patrick Pantel. "DIRT - Discovery of Inference Rules from Text" In KDD (2001)

Paraphrase Research



Paraphrase Research





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:

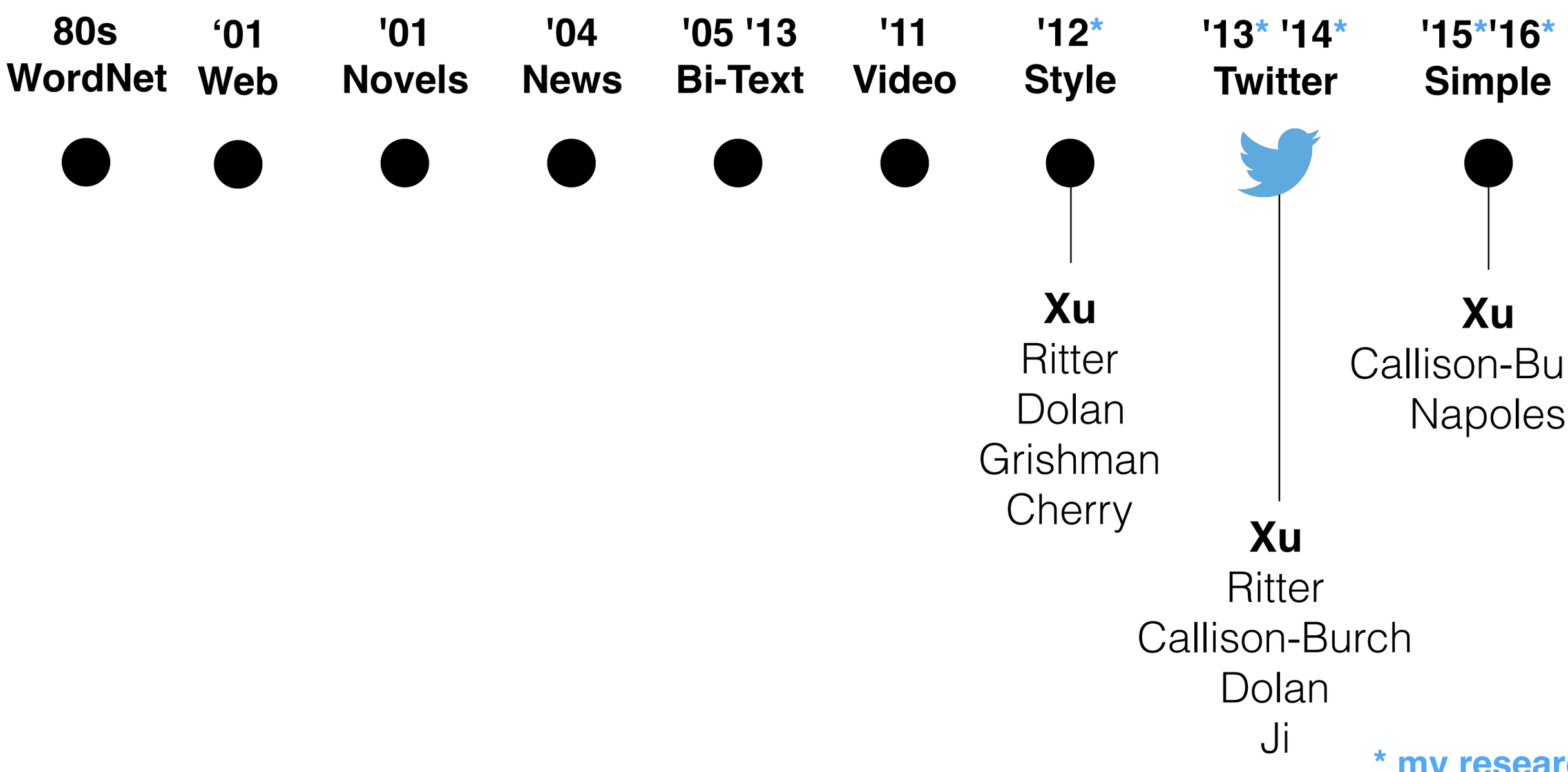
Emma burst into tears **and he tried to** comfort **her,** saying things to make her smile.
Emma cried **and he tried to** console **her,** adorning his words with puns.

burst into tears = cried and comfort = console

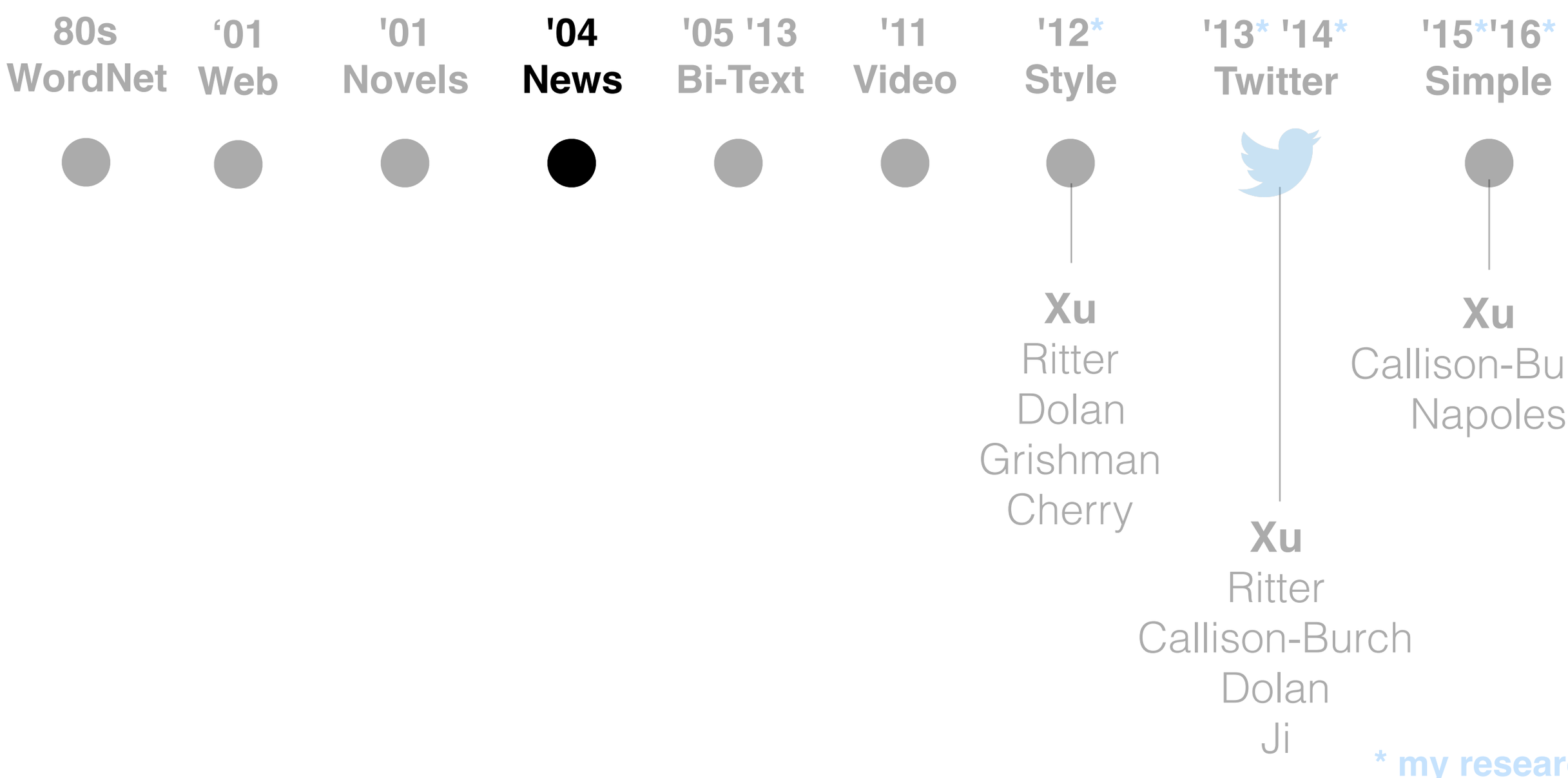
Source: Chris Callison-Burch

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

Paraphrase Research



Paraphrase 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)

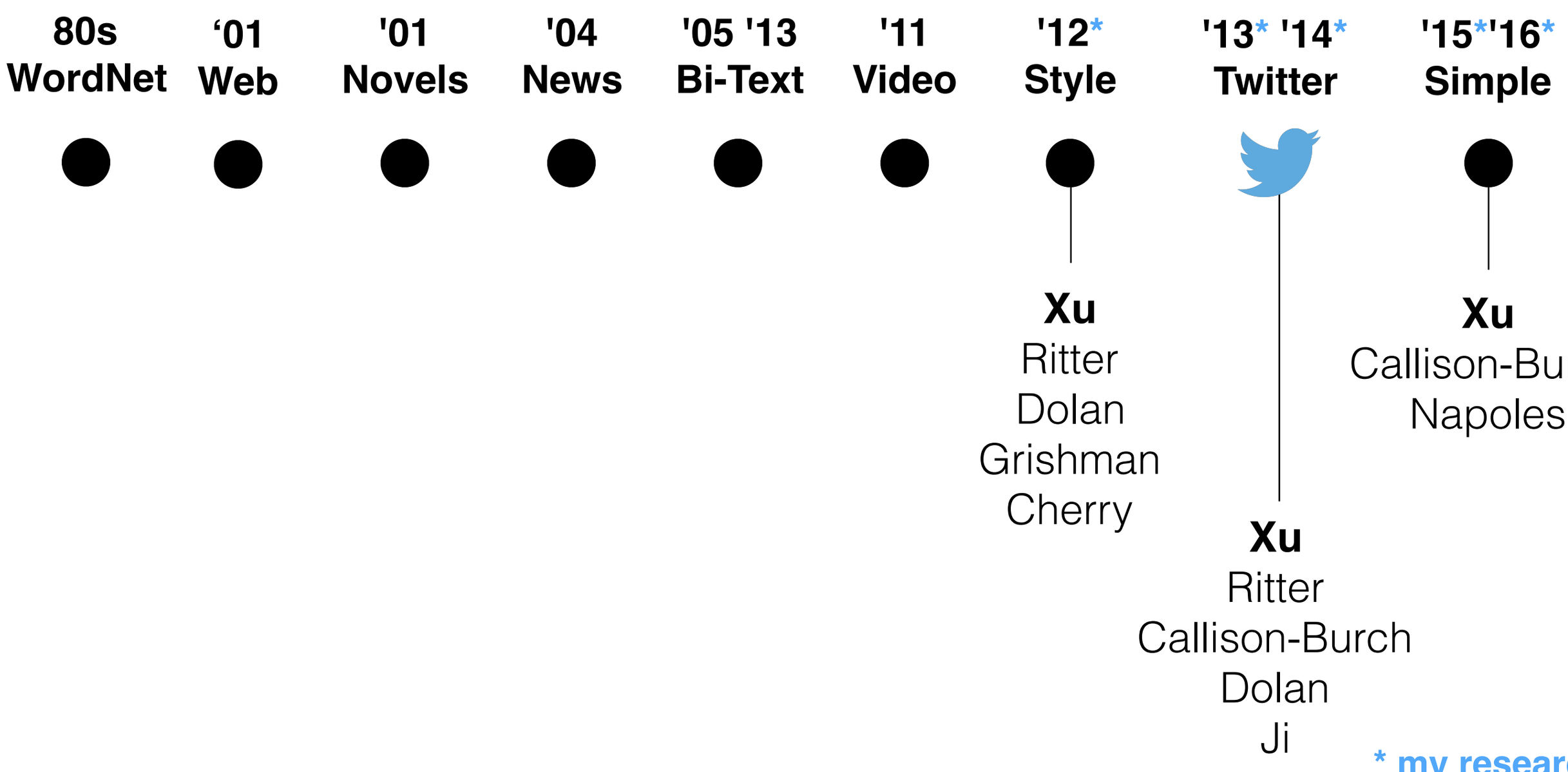
Dolan, Quirk, and Brockett (2004) extract sentential paraphrases from newspaper articles published on the same topic and date:

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



Paraphrase Research

80s WordNet

‘01 Web

'01 Novels

'04 News

'05 '13 Bi-Text

'11 Video

'12* Style

Xu
Ritter
Dolan
Grishman
Cherry

Xu
Ritter
Callison-Burch
Dolan
Ji

'13* '14* Twitter

Xu
Callison-Burch
Napoles

'15*'16* Simple

* 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)

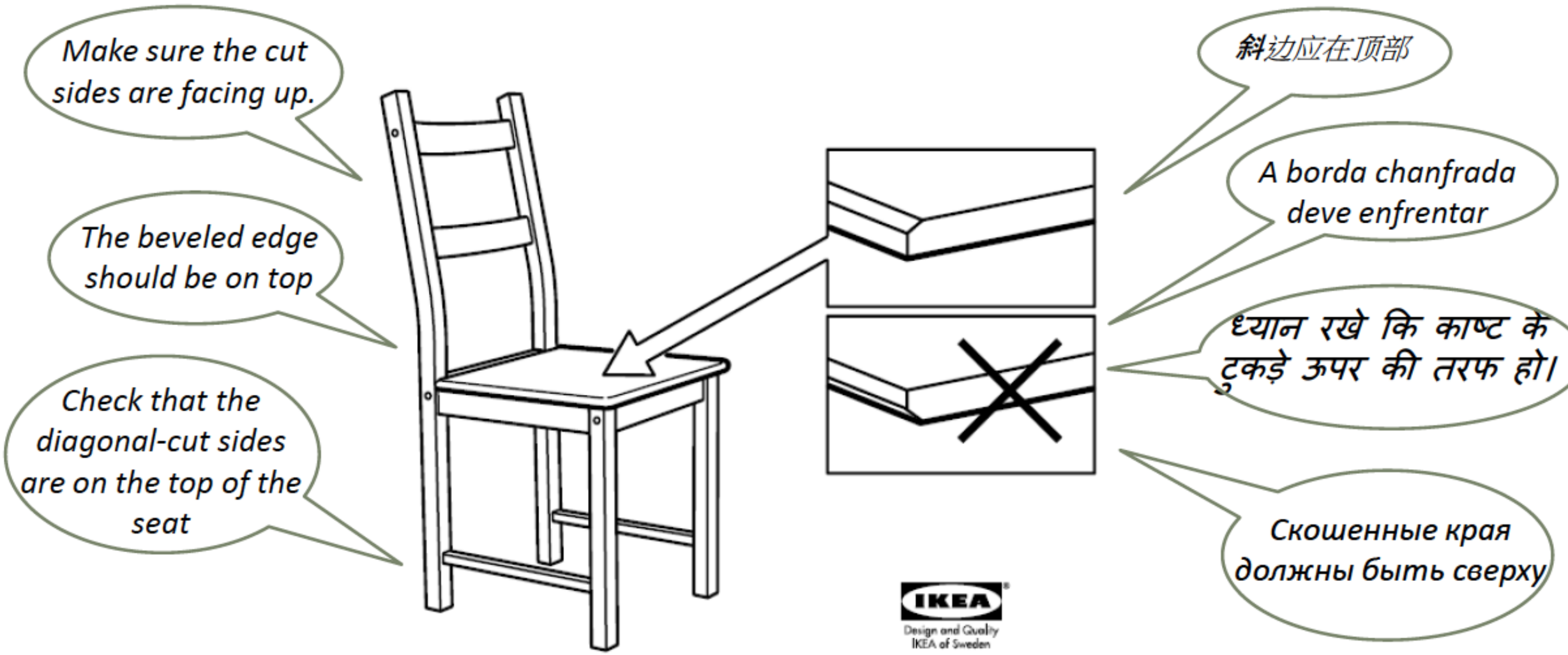
Paraphrasing & Translation

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

Paraphrasing is translation into the same language.

Source: Chris Callison-Burch

Same Meaning, Different Words



Source: Bill Dolan

Bilingual Data

Sentence-aligned parallel corpora in English and any foreign language

Available in large quantities

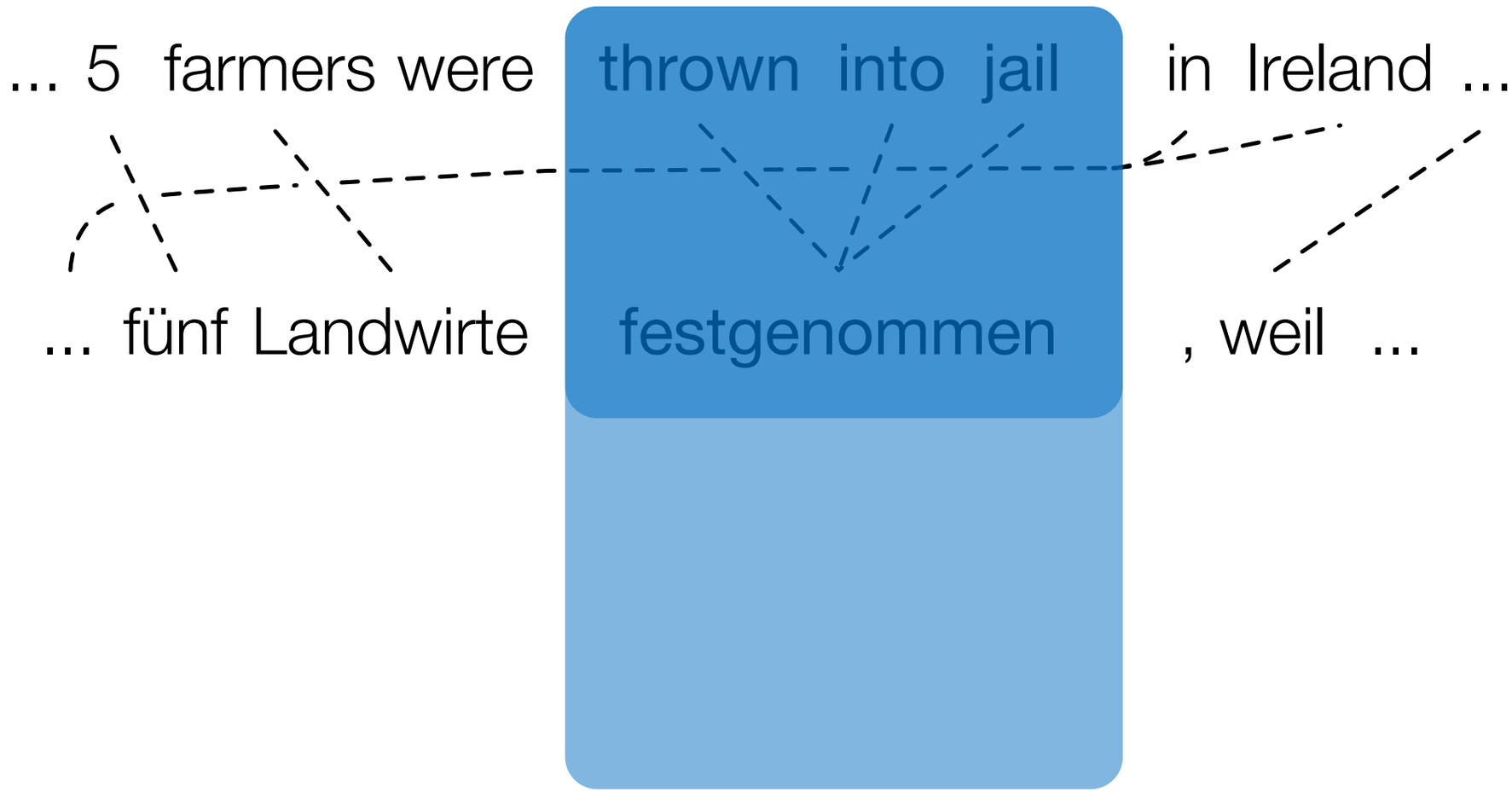
Strong meaning equivalence signal

... but different languages.

Source: Chris Callison-Burch

Bilingual Pivoting

word alignment

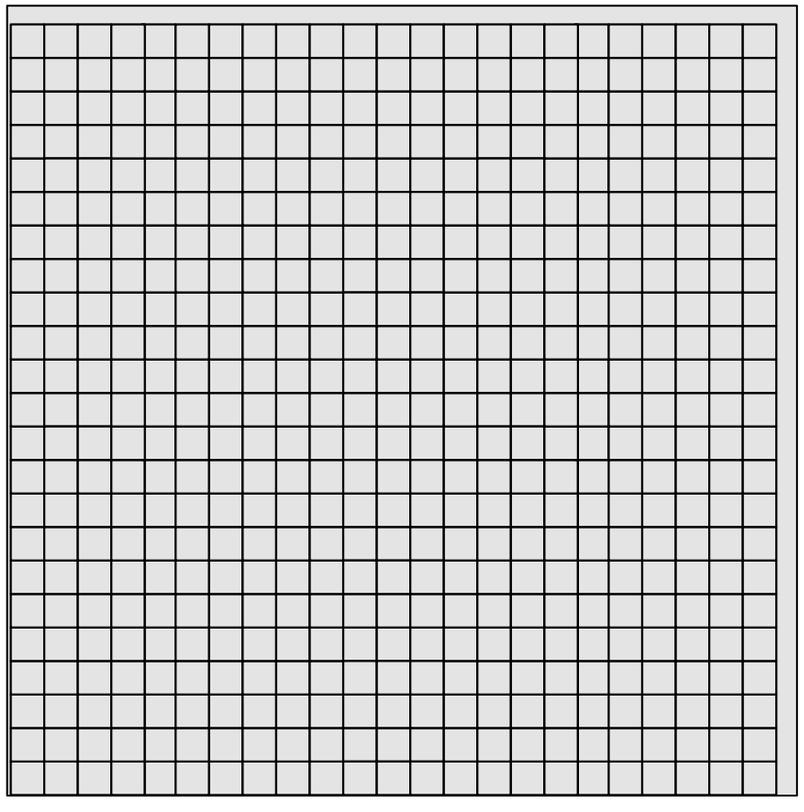


Source: Chris Callison-Burch

Large and diverse

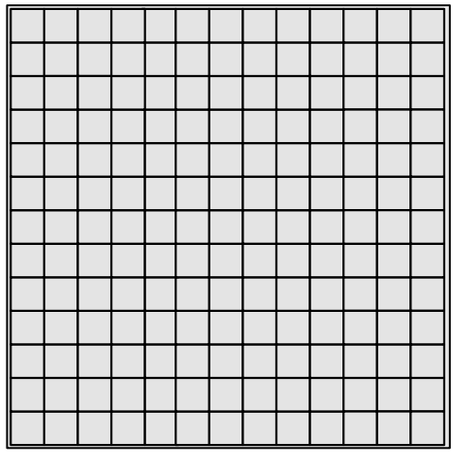
Bilingual Data Sets

1000M



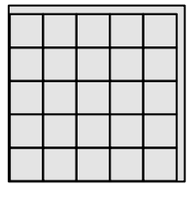
French-English
10^9 word webcrawl

2 languages @
250M each



DARPA
GALE Program

21 languages @
50-80M each



European
Parliament

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

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

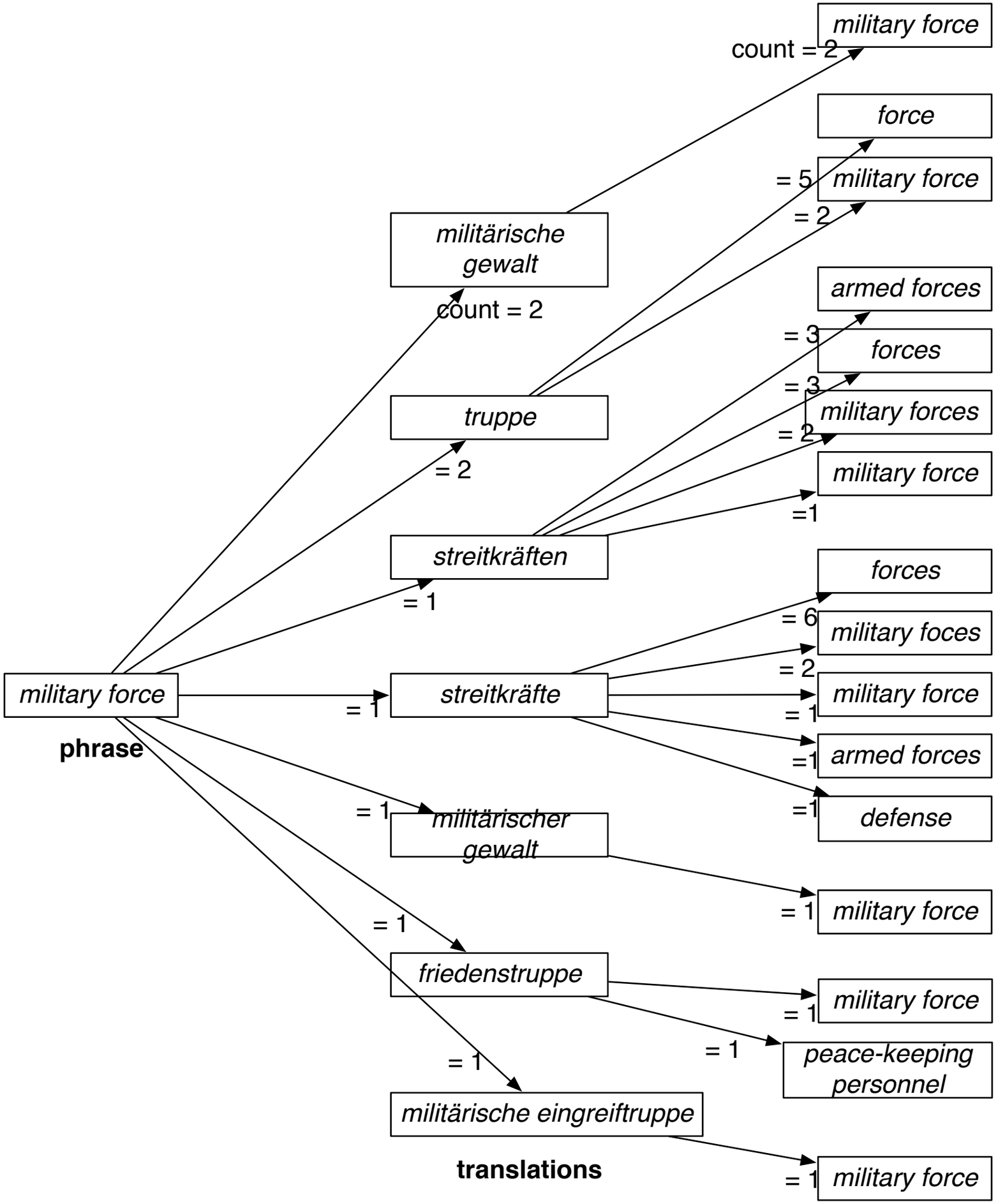
Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. Chris Callison-Burch. EMNLP 2008.

Paraphrase Probability

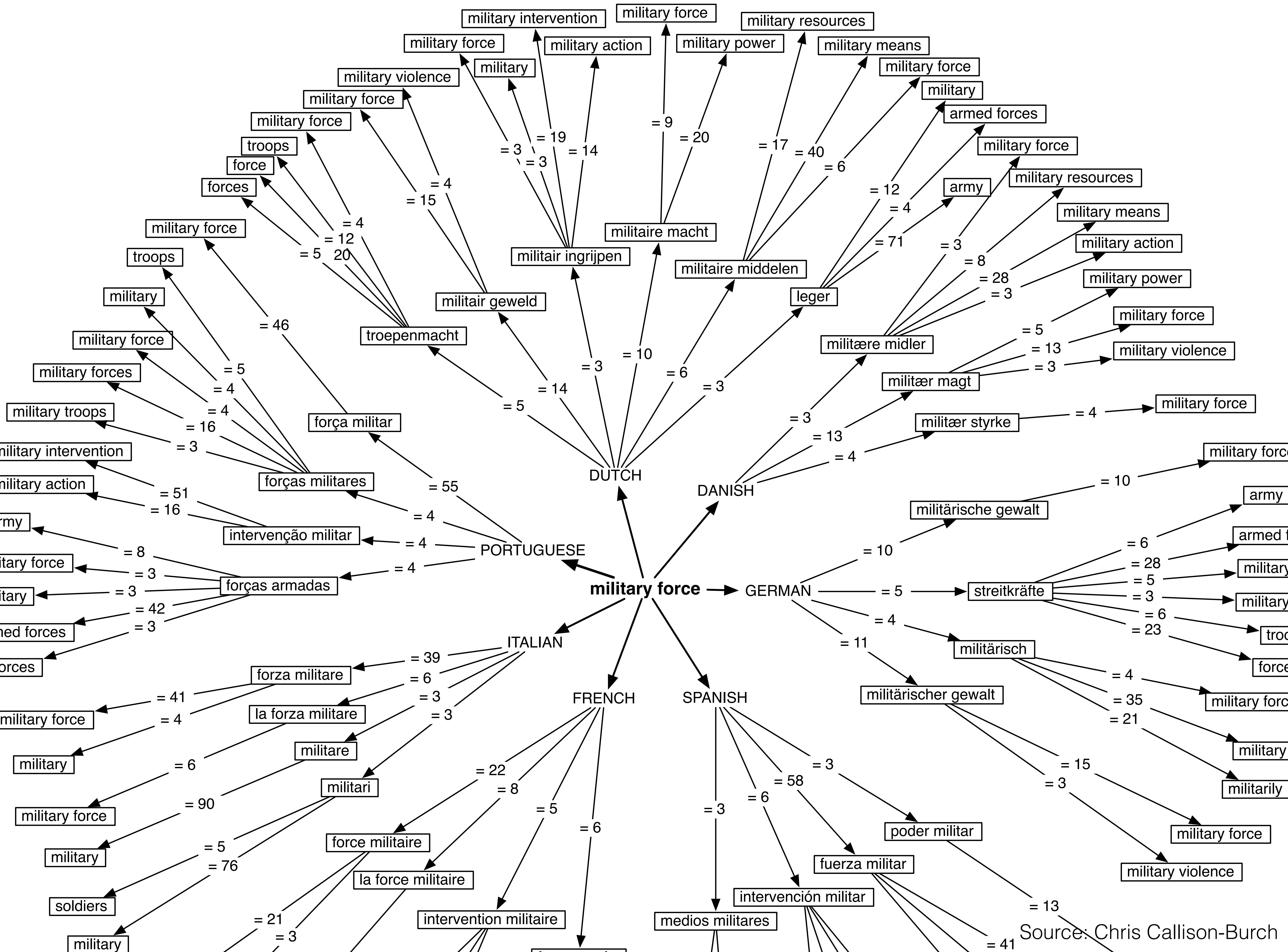
$$
\begin{aligned}
p(e_2|e_1) &= \sum_f p(e_2, f|e_1) \\
&= \sum_f p(e_2|f, e_1)p(f|e_1) \\
&\approx \sum_f p(e_2|f)p(f|e_1)
\end{aligned}
$$

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

How about

Syntactic Paraphrases?

Bilingual parallel corpora provide an excellent source of **lexical** and **phrasal** paraphrases.

Sentential/structural paraphrases are more obviously learned from **English-English sentence pairs.**

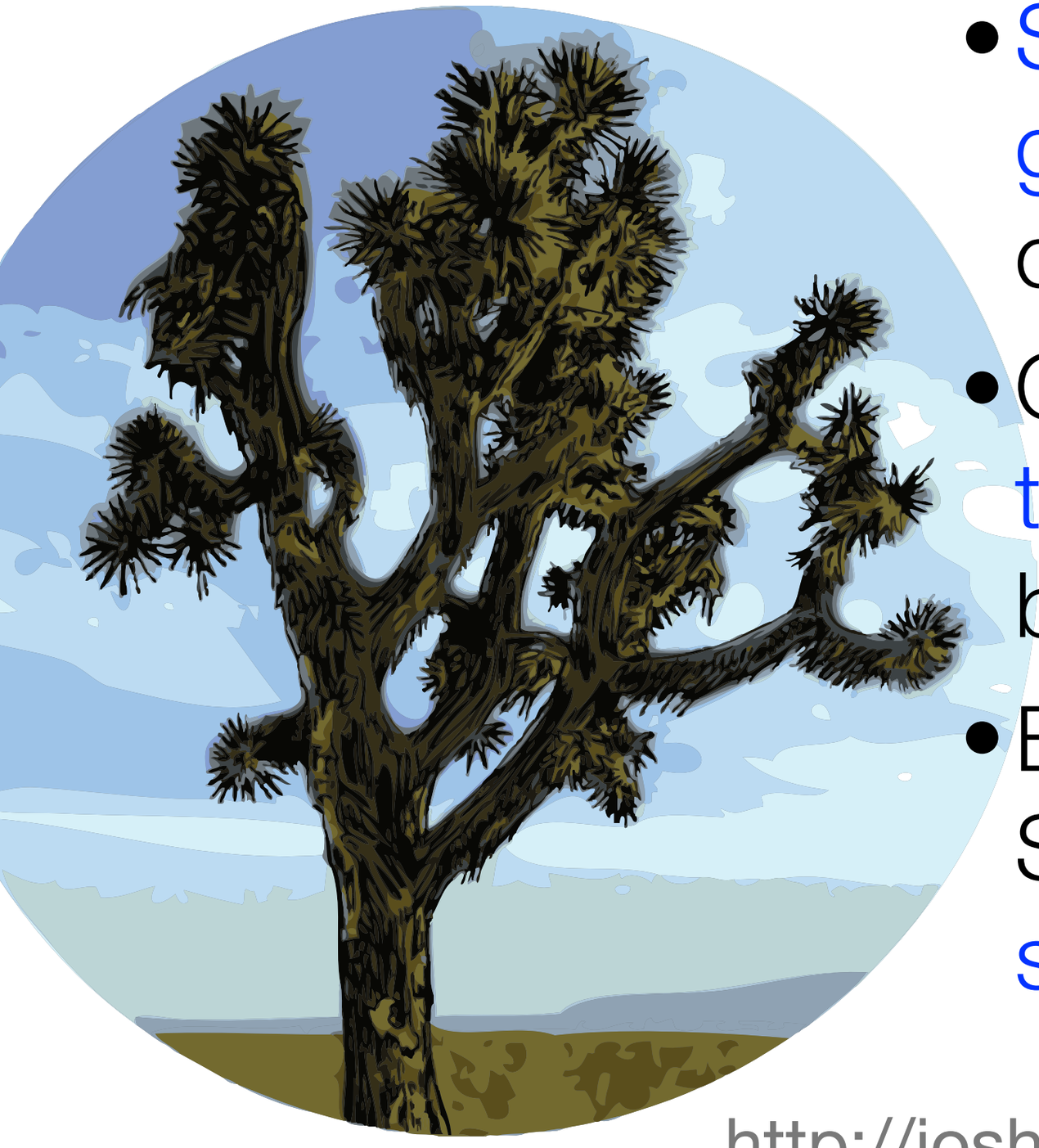
Can we learn structural paraphrases from bitexts?
How should we represent them?

Source: Chris Callison-Burch

Syntactic Machine Translation

	Urdu	**English**
S →	NP① VP②	NP① VP②
VP →	PP① VP②	VP② PP①
VP →	V① AUX②	AUX② V①
PP →	NP① P②	P② NP①
NP →	*hamd ansary*	*Hamid Ansari*
NP →	*na}b sdr*	*Vice President*
V →	*namzd*	*nominated*
P →	*kylye*	*for*
AUX →	*taa*	*was*

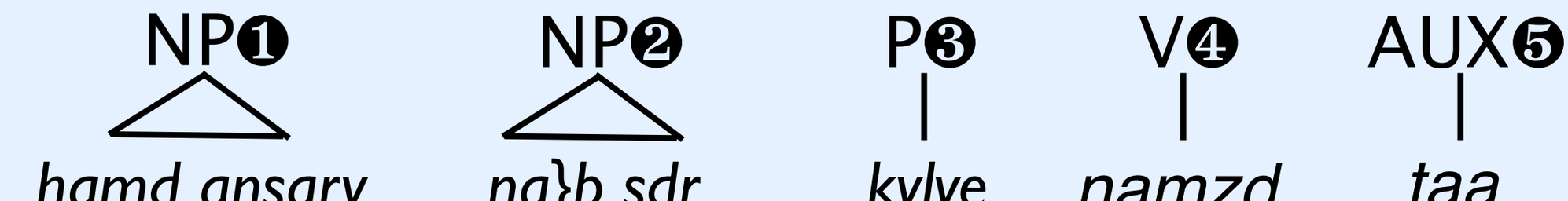
Syntactic MT in the Joshua Decoder



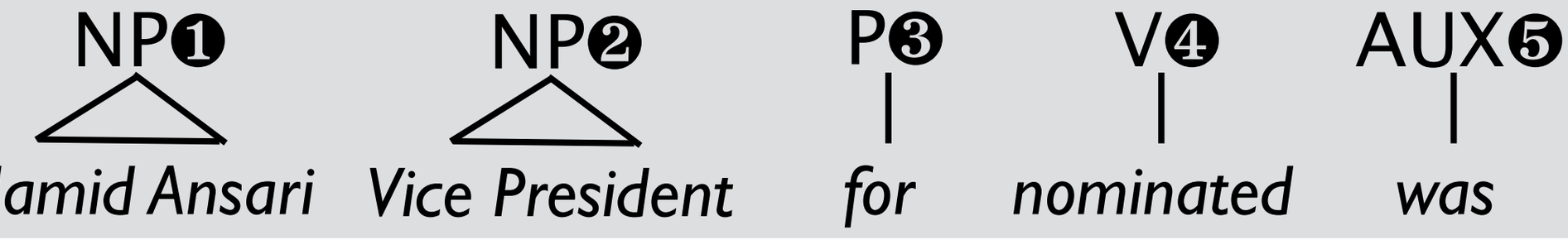
- Synchronous context free grammars generate pairs of corresponding strings
- Can be used to describe translation and re-ordering between languages
- Because Joshua uses SCFGs, it translates sentences by parsing them

http://joshua-decoder.org

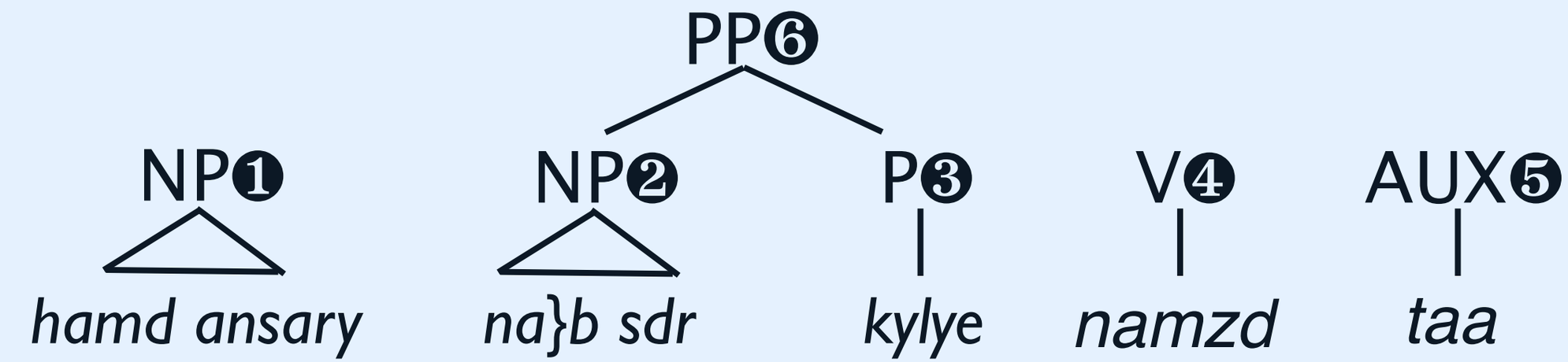
Urdu



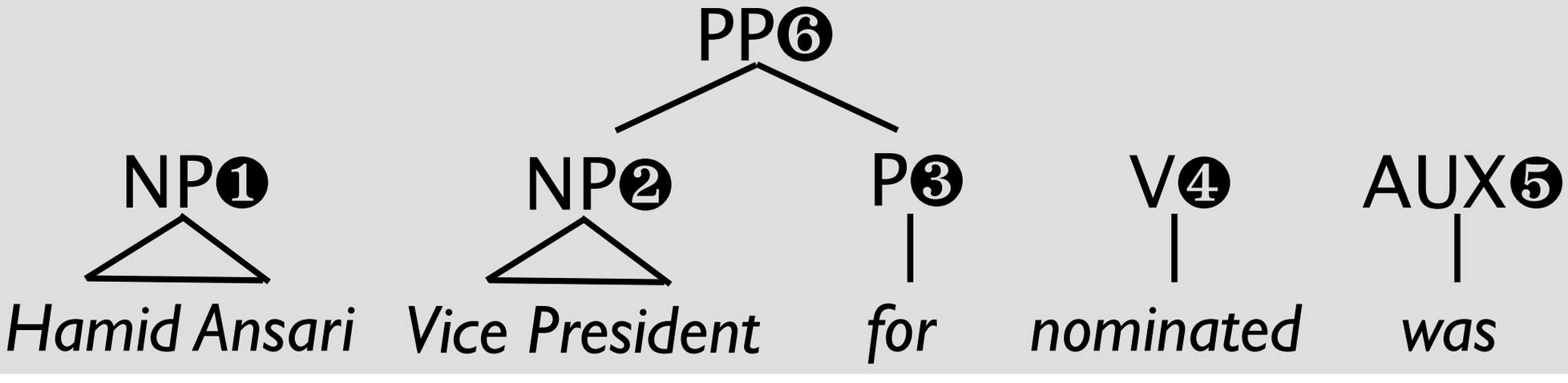
English



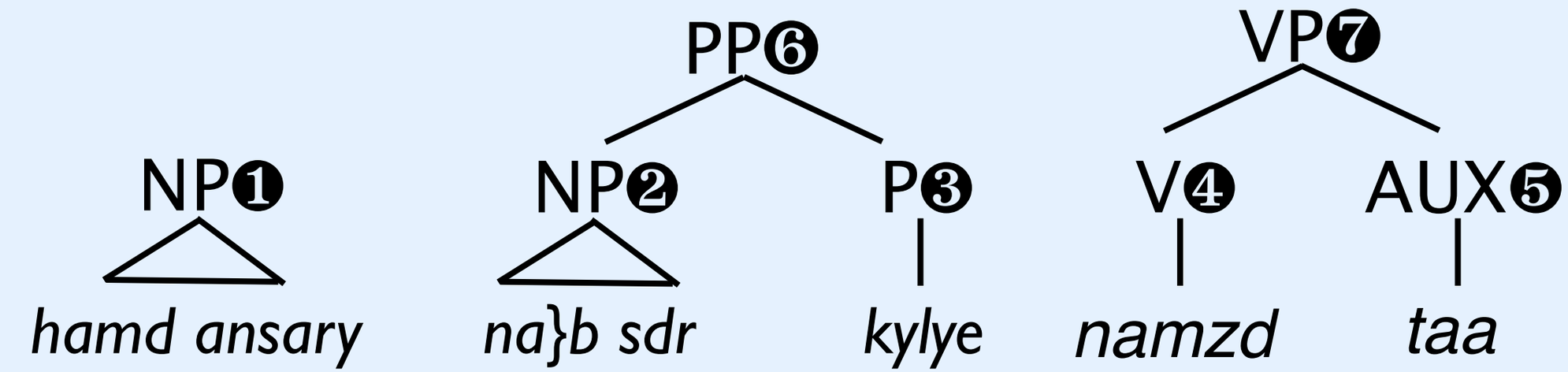
Urdu



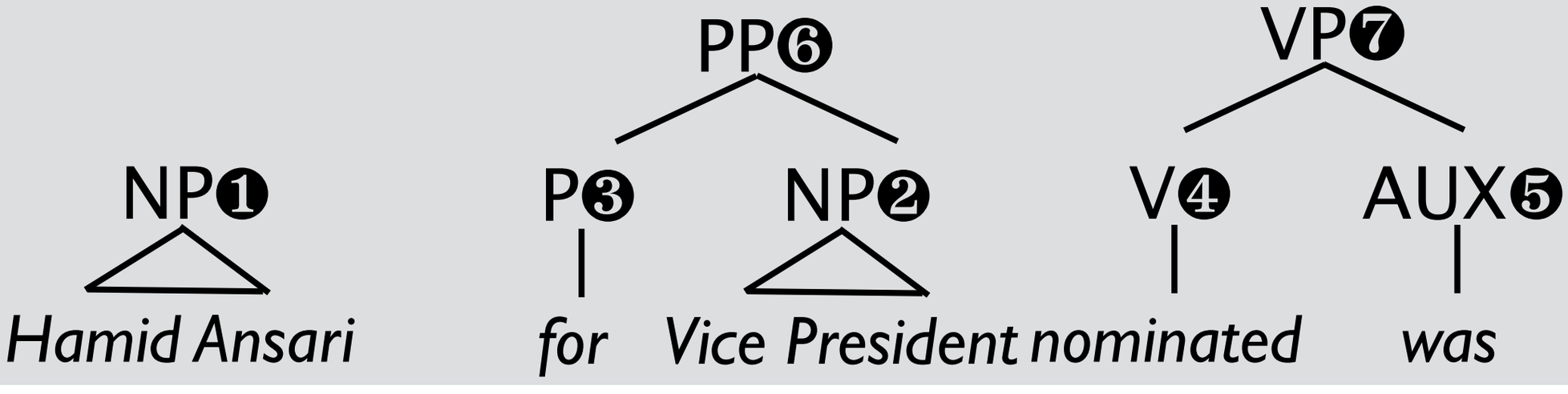
English



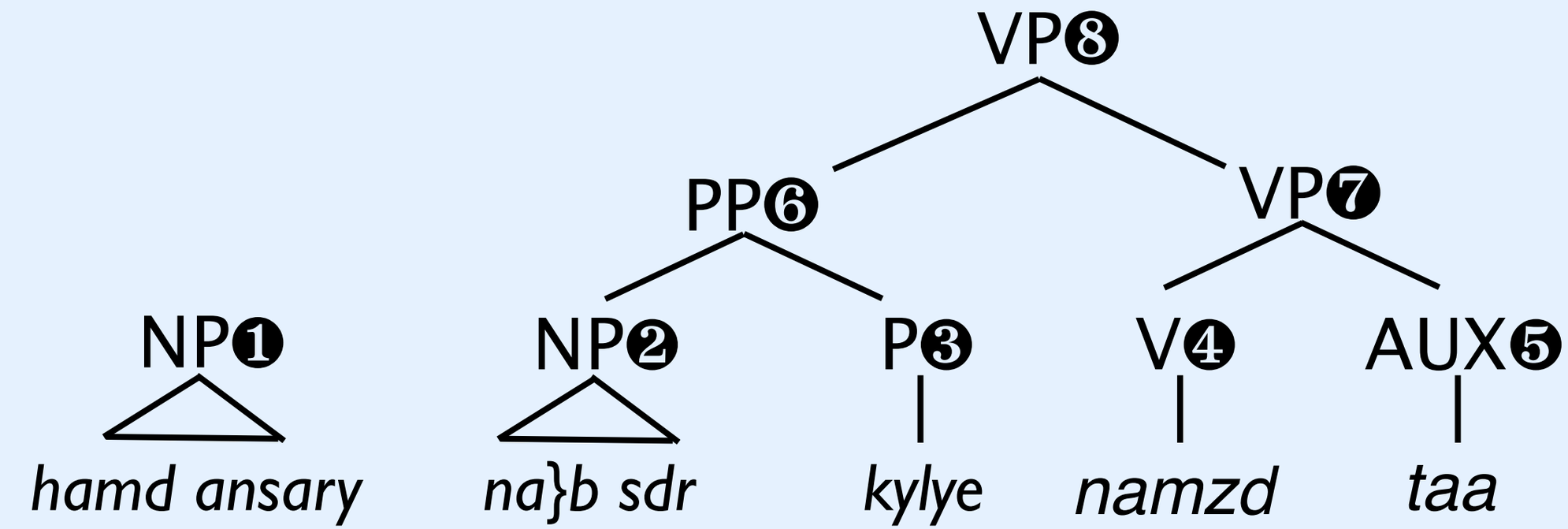
Urdu



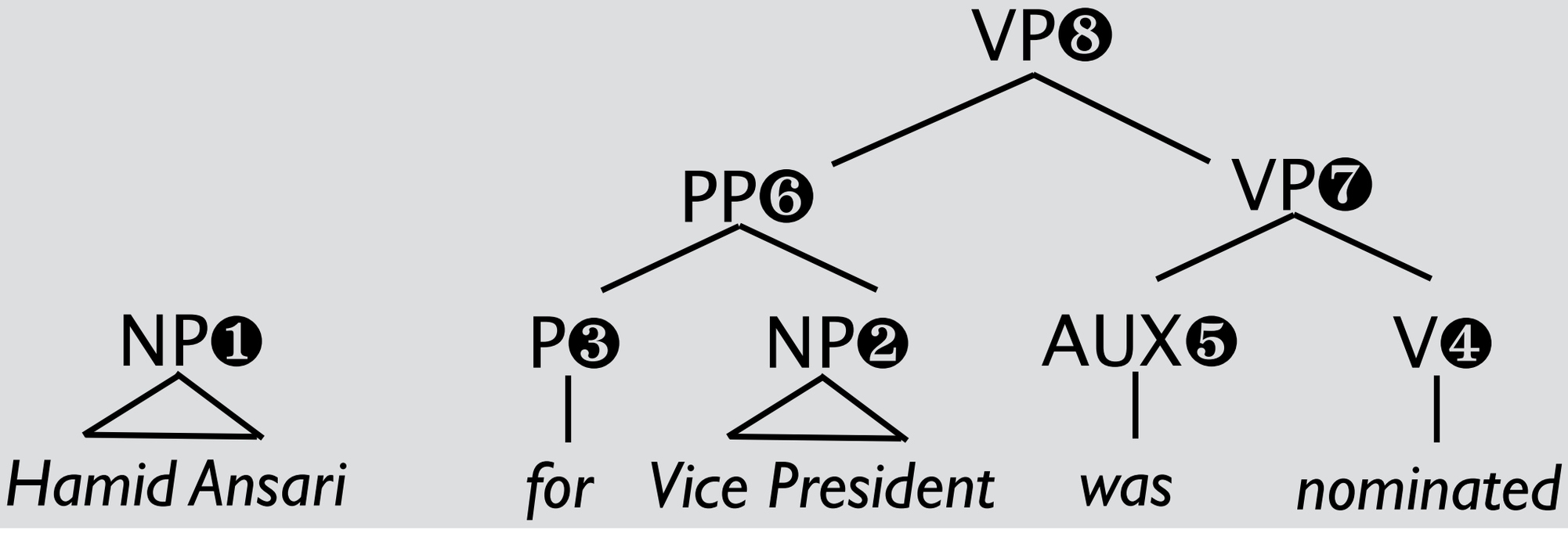
English



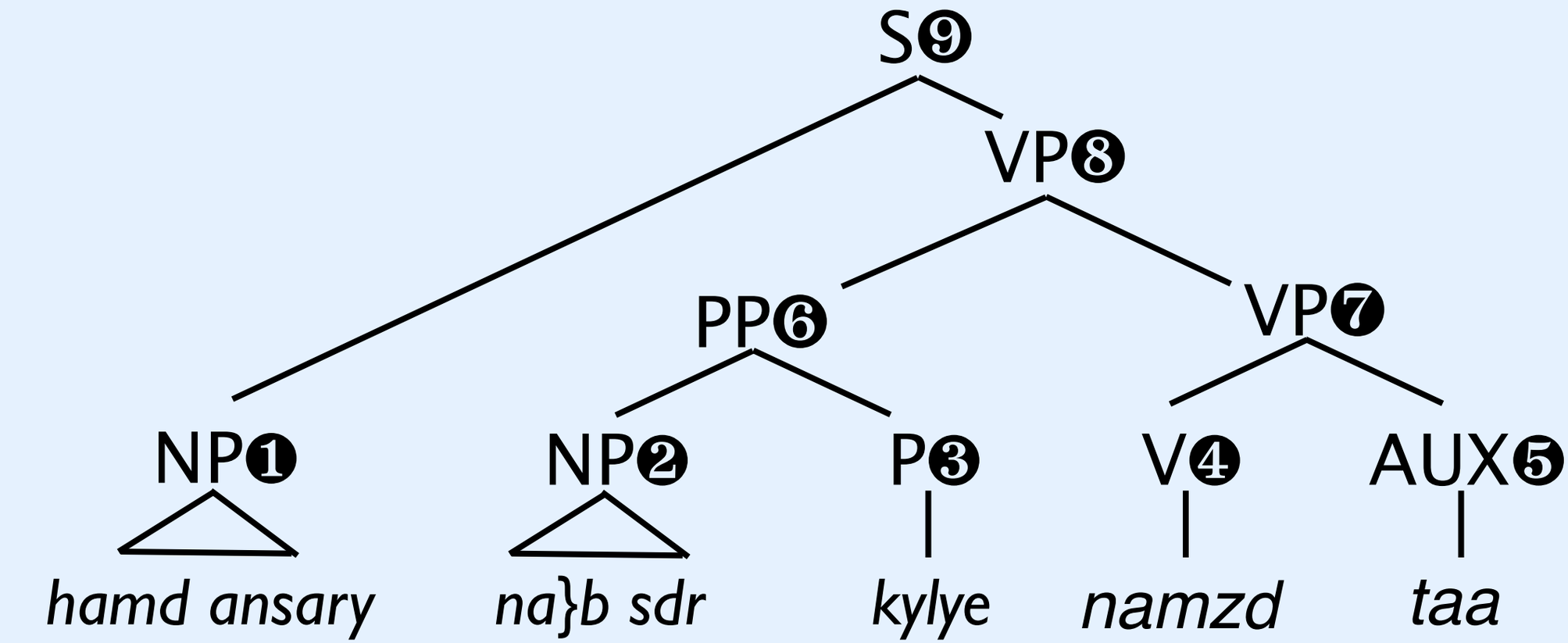
Urdu



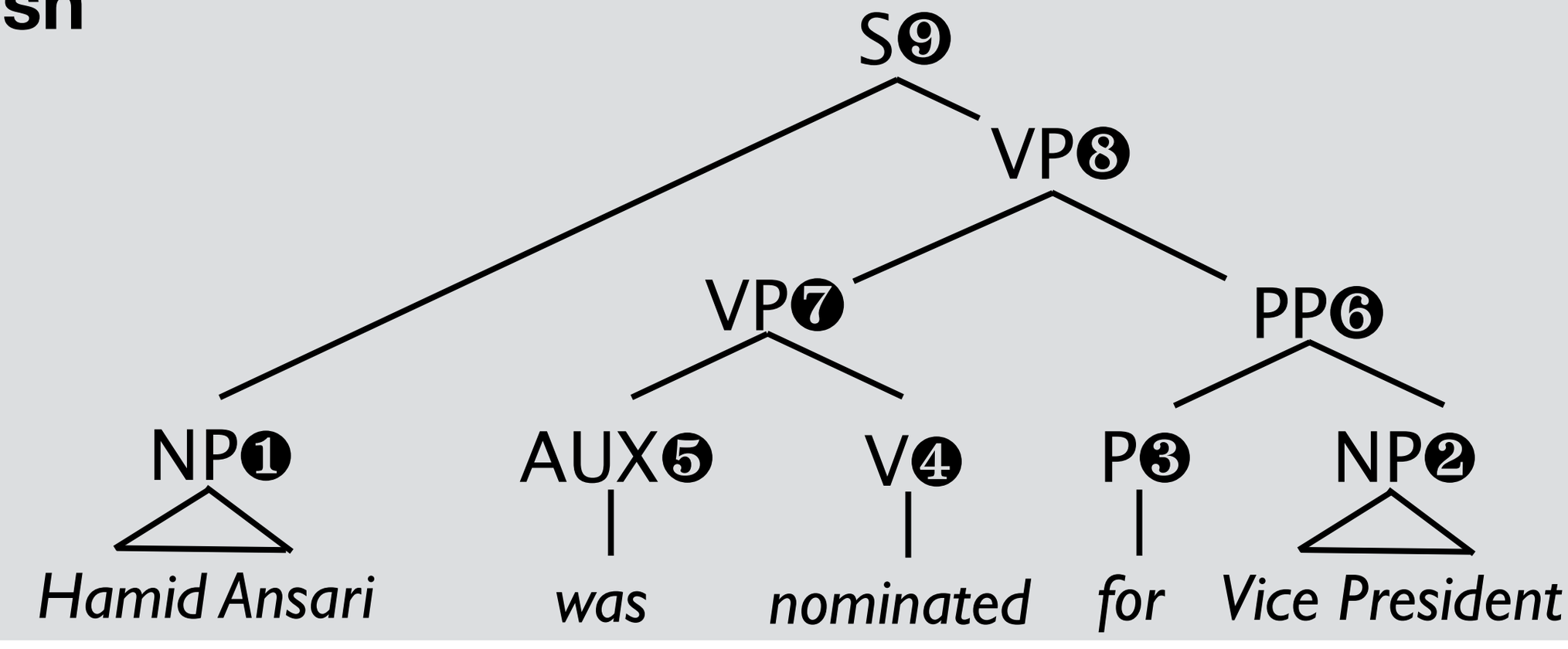
English



Urdu



English



Syntactic Paraphrase via Pivoting

- Adapting our syntactic MT models, we learn structural transformations, like the English possessive rule

NP → NP 's NN | le NN de NP

NP → the NN of NP | le NN de NP

combine to

NP → NP 's NN | the NN of NP

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?"

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.

✓ cup ↔ mug

..sip from a **mug** of cocoa..
..a **mug** of coffee.

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

✓ the king's speech ↔ His Majesty's address

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

? one JJ instance of NP ↔ a JJ case of NP

Source: Chris Callison-Burch

Syntactic Paraphrase Similarity

NN 's NP in the long term

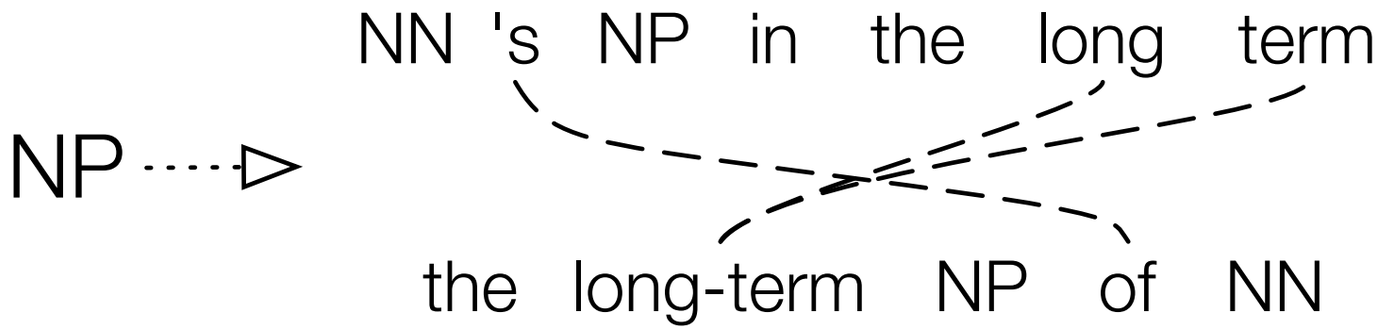
NP ⇢

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.

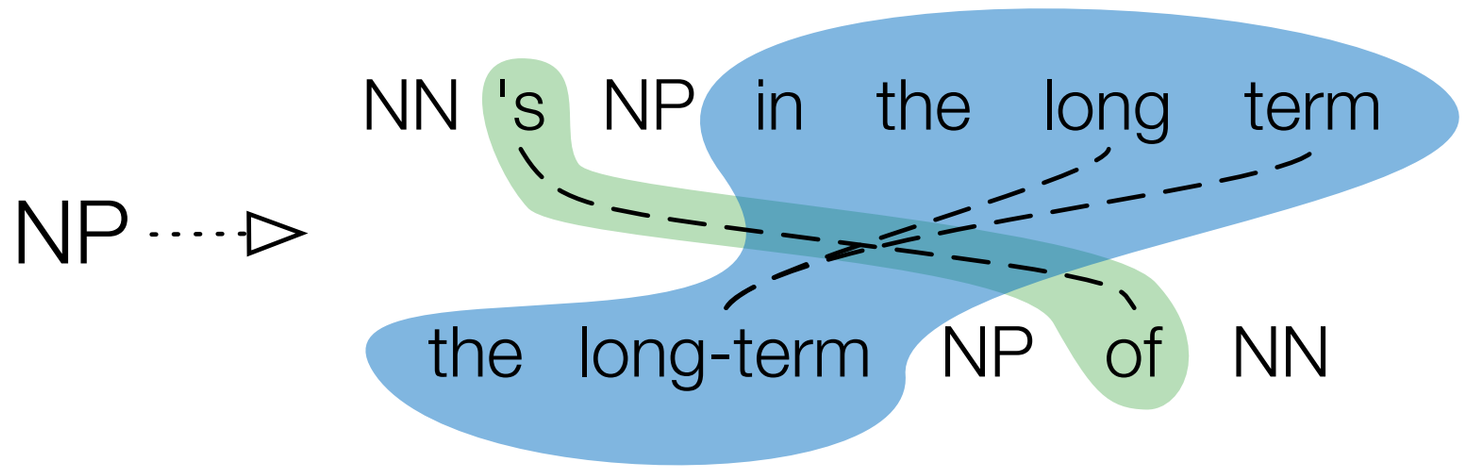
Syntactic Paraphrase Similarity



Source: Chris Callison-Burch

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

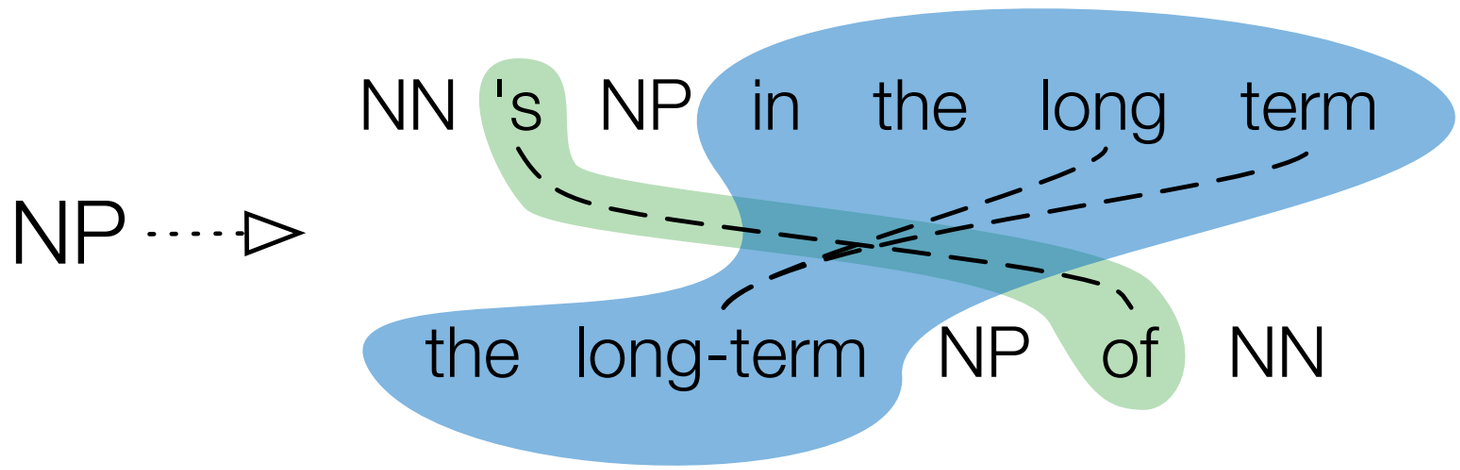
Syntactic Paraphrase Similarity



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

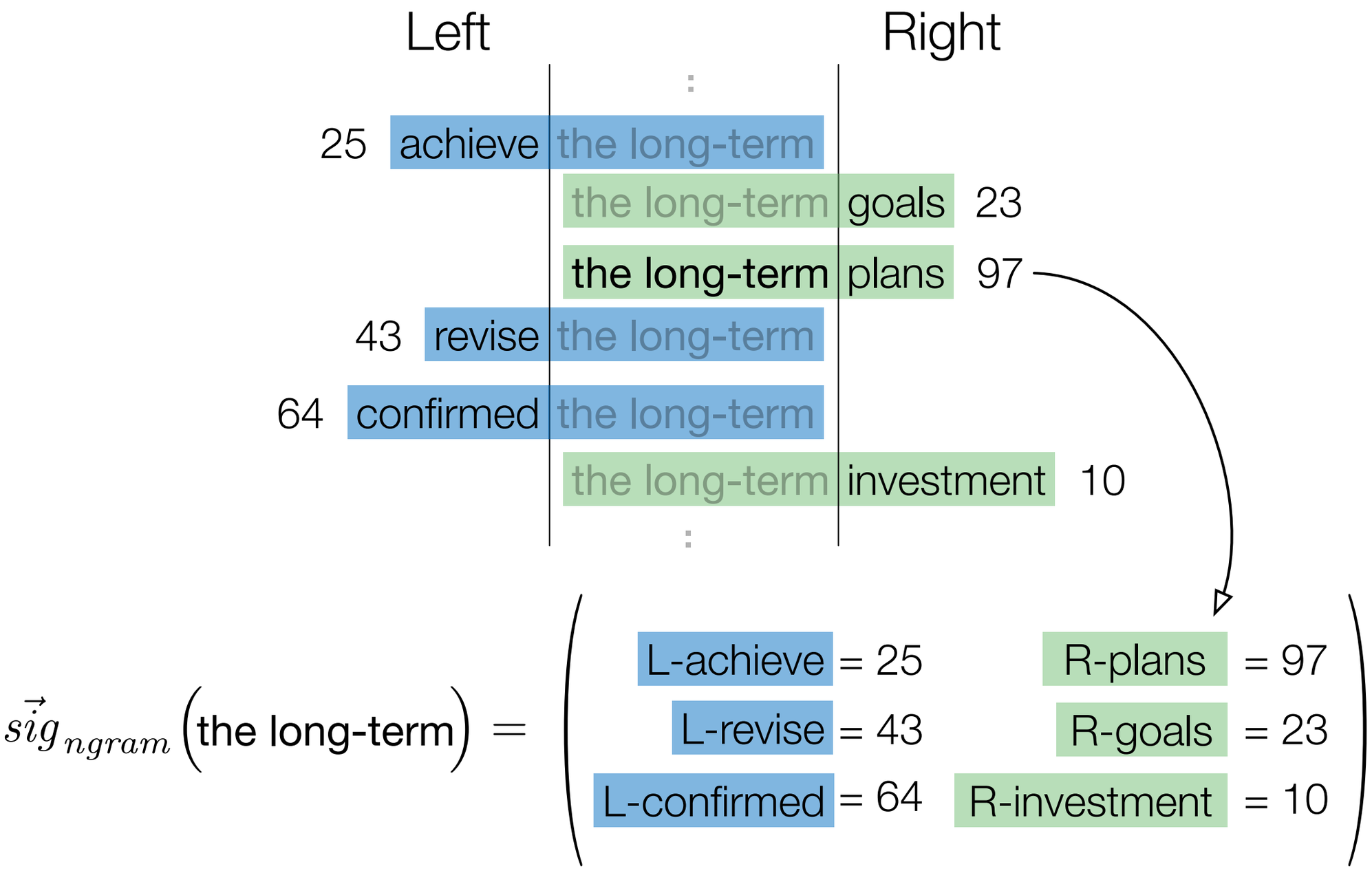


$$sim(\mathbf{r}) = \frac{1}{2}\left(sim\begin{pmatrix}\text{the long-term}\\ \text{in the long term}\end{pmatrix} + sim\begin{pmatrix}\text{'s}\\ \text{of}\end{pmatrix}\right)$$

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

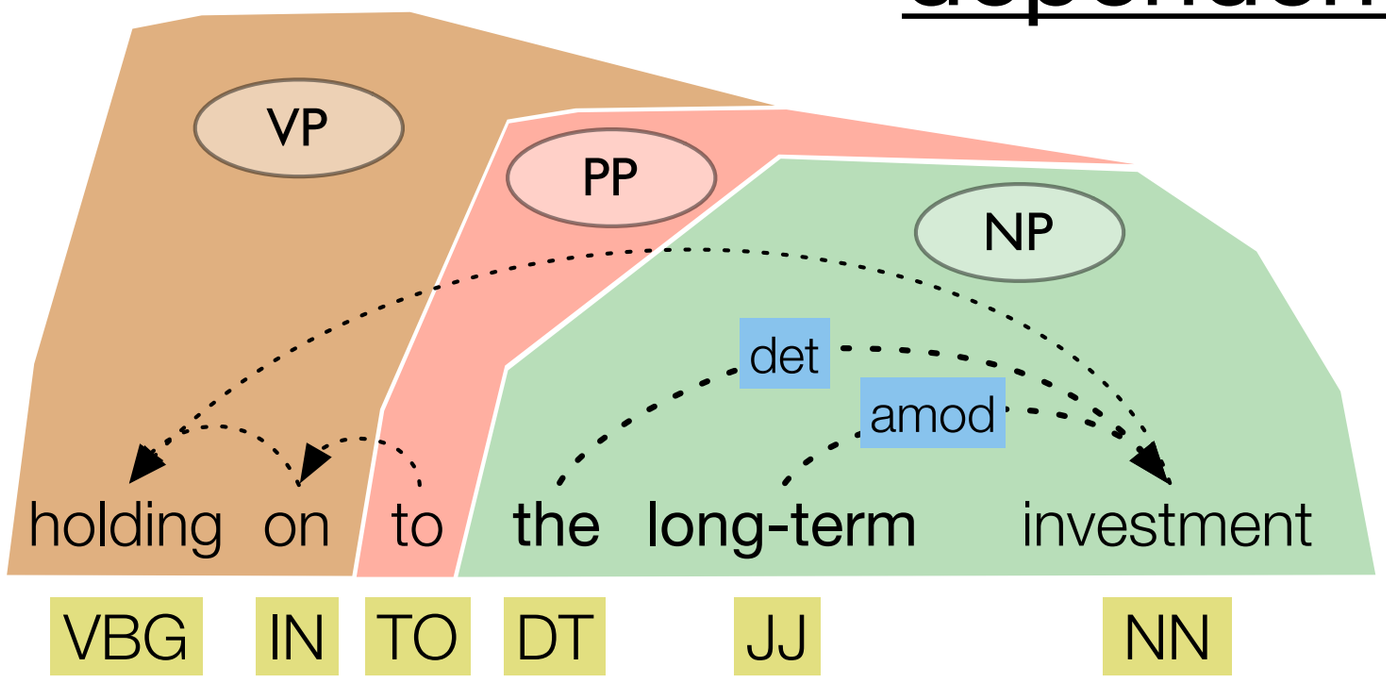


Source: Chris Callison-Burch

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

Syntactic Context

dependency parsing



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

Source: Chris Callison-Burch

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

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. PPDB: The Paraphrase Database. NAACL 2013.

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

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

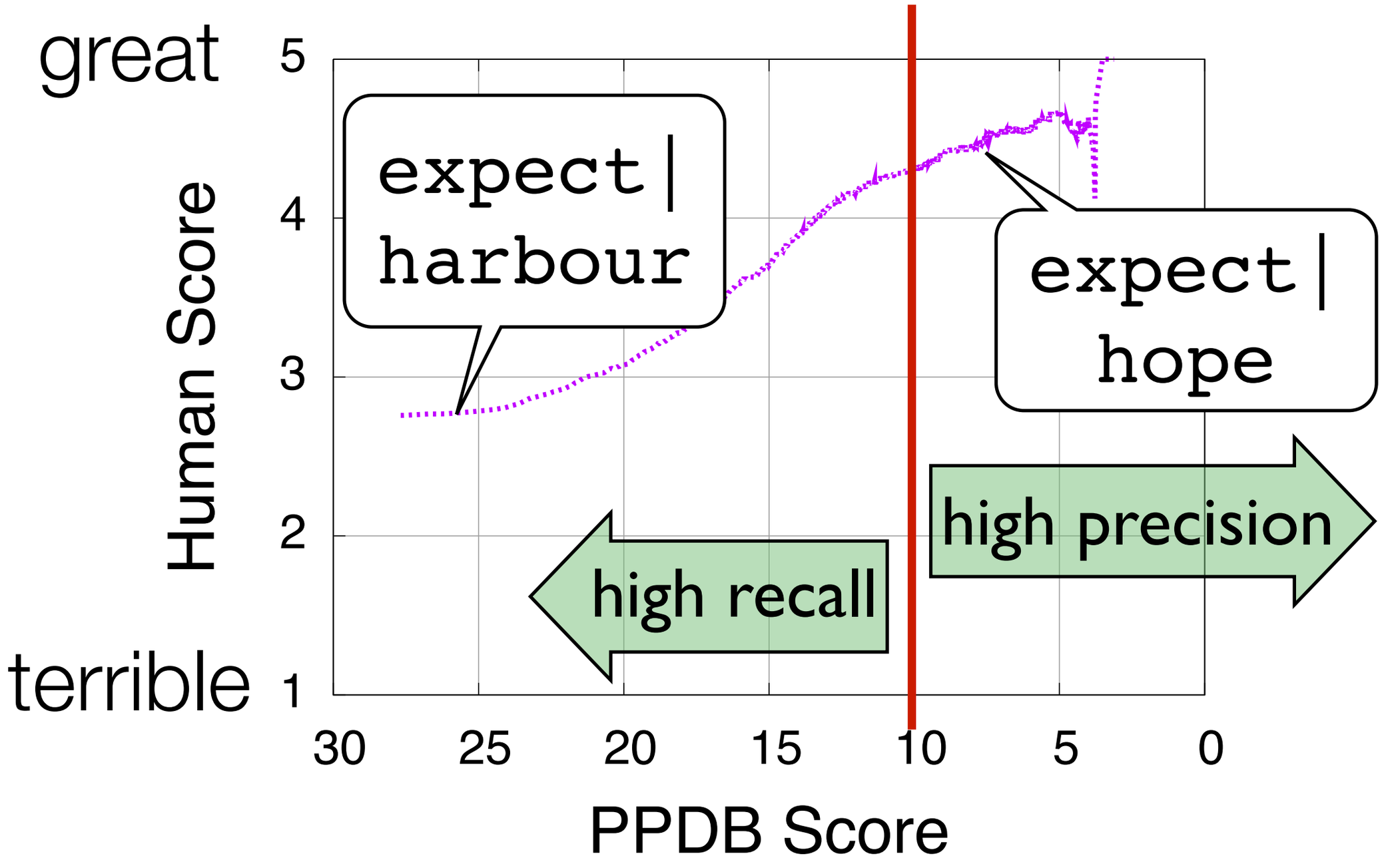
L Size

XL Size

XXL Size

XXXL Size

Do the Scores Work?



Source: Chris Callison-Burch

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

Fun PPDB Examples

munchies ||| hungry

sex am

sheeit osers

abso-fucking-lutely ||| indeed

Source: Chris Callison-Burch

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

Quiz #3

socialmedia-class.org