Penn
UNIVERSITY *of* PENNSYLVANIA

Large-scale Paraphrasing for Natural Language Generation

Chris Callison-Burch

with Ellie Pavlick, Wei Xu, Anne Cocos, Courtney Napoles, Juri Ganitkevitch, Benjamin Van Durme, Xuchen Yao, Peter Clark, Jonny Weese, Matt Post, Tsz Ping Chan, Rui Wang, Trevor Cohn, Mirella Lapata and Colin Bannard

Paraphrases

Differing **textual** expressions of the same **meaning**:

cup	$\leftrightarrow$	mug
the king's speech	$\leftrightarrow$	His Majesty's address
X_1 devours X_2	$\leftrightarrow$	X_2 is eaten by X_1
one JJ instance of NP	$\leftrightarrow$	a JJ case of NP

Paraphrasing in NLP

Recognition or generation of paraphrases plays a part in...

...information extraction, question answering, entailment recognition, summarization, translation, compression, simplification, automatic evaluation of translation or summaries, natural language generation, etc.

Data-Driven Paraphrasing

Monolingual parallel:	English – English
Monolingual comparable:	English ~ English
Plain monolingual:	English
Bilingual parallel:	English – French

Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods.
Nitin Madnani and Bonnie Dorr. 2010. Computational Linguistics, 36(3), pages 341-387.

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**.

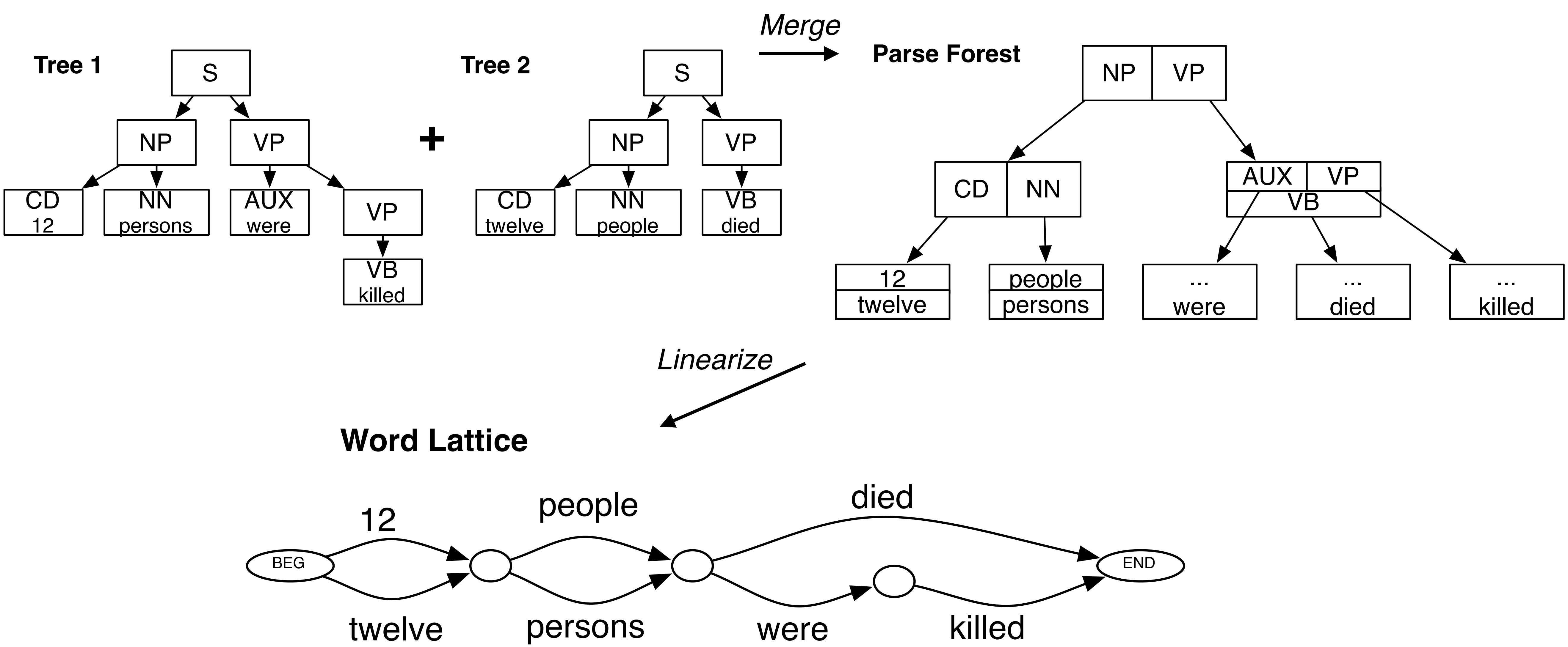
Paraphrasing with parallel monolingual data

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

Paraphrasing with parallel monolingual data



Pang, Knight and Marcu (2003)

Paraphrasing with 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.

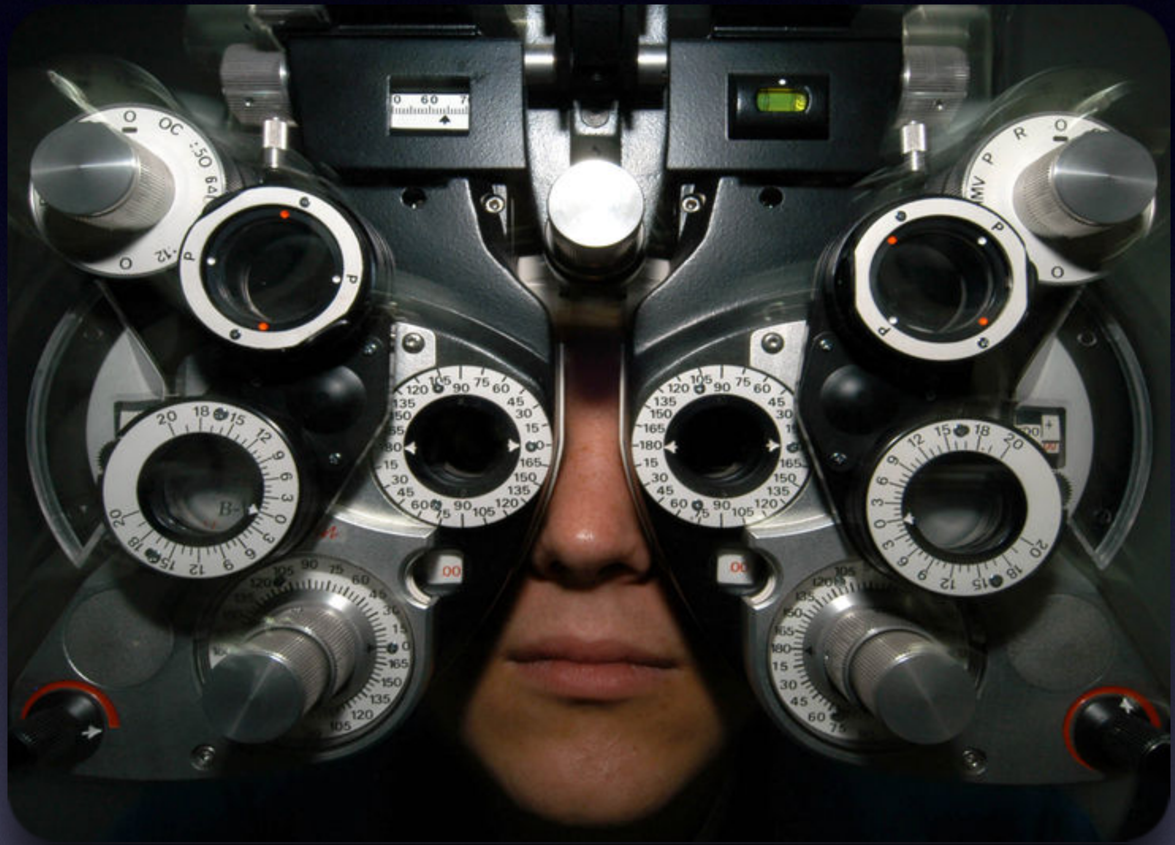
Distributional Hypothesis

If we consider **optometrist** 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 **optometrist** 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)



ACL 1993

DISTRIBUTIONAL CLUSTERING OF ENGLISH WORDS

Fernando Pereira
AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974, USA
pereira@research.att.com

Naftali Tishby
Dept. of Computer Science
Hebrew University
Jerusalem 91904, Israel
tishby@cs.huji.ac.il

Lillian Lee
Dept. of Computer Science
Cornell University
Ithaca, NY 14850, USA
llee@cs.cornell.edu

Abstract

We describe and evaluate experimentally a method for clustering words according to their distribution in particular syntactic contexts. Words are represented by the relative frequency distributions of contexts in which they appear, and relative entropy between those distributions is used as the similarity measure for clustering. Clusters are represented by average context distributions derived from the given words according to their probabilities of cluster membership. In many cases, the clusters can be thought of as encoding coarse sense distinctions. Deterministic annealing is used to find lowest distortion sets of clusters: as the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word coocurrence, and the models evaluated with respect to held-out test data.

INTRODUCTION

Methods for automatically classifying words according to their contexts of use have both scientific and practical interest. The scientific questions arise in connection to distributional views

sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden *senses classes* and associations between the classes themselves. While it may be worth basing such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or *clusters* c with corresponding cluster membership probabilities $p(c|w)$ for each word w. Most other class-based modeling techniques for natural

DIRT

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 ...

Word2vec

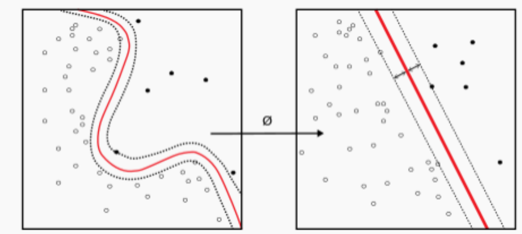
From Wikipedia, the free encyclopedia

Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a high-dimensional space (typically of several hundred dimensions), with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.[1]

Word2vec was created by a team of researchers led by Tomas Mikolov at Google. The algorithm has been subsequently analysed and explained by other researchers[2][3] and a Bayesian version of the algorithm is proposed as well.[4] Embedding vectors created using the Word2vec algorithm have many advantages compared to earlier algorithms like Latent Semantic Analysis.

Contents [hide]

Machine learning and data mining



Problems [show]
Supervised learning (**classification** • **regression**) [show]
Clustering [show]
Dimensionality reduction [show]
Structured prediction [show]
Anomaly detection [show]
Neural nets [show]
Reinforcement Learning [show]
Theory [show]
Machine learning venues [show]
Machine learning portal
V • T • E

Skip grams and CBOW [edit]

Mikolov et al (2013)

My focus:
Paraphrasing & Translation

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

Paraphrasing is translation into the same language.

Inspiration from Statistical Machine Translation

We reuse & adapt:

- Training data + alignment algorithms
- Models + feature functions
- Parameter estimation
- Decoder

Bilingual Data

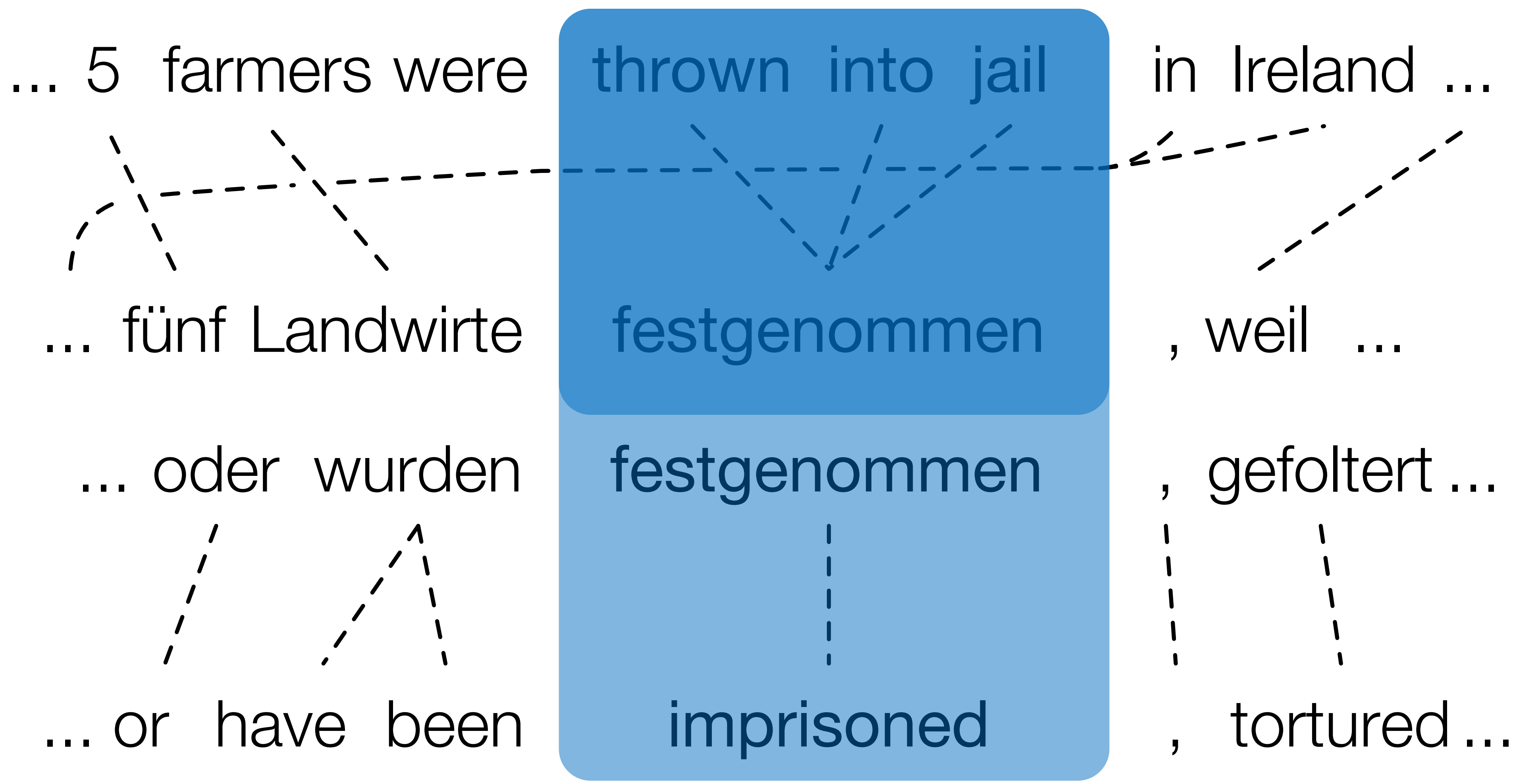
Sentence-aligned parallel corpora in English and any foreign language

Available in large quantities

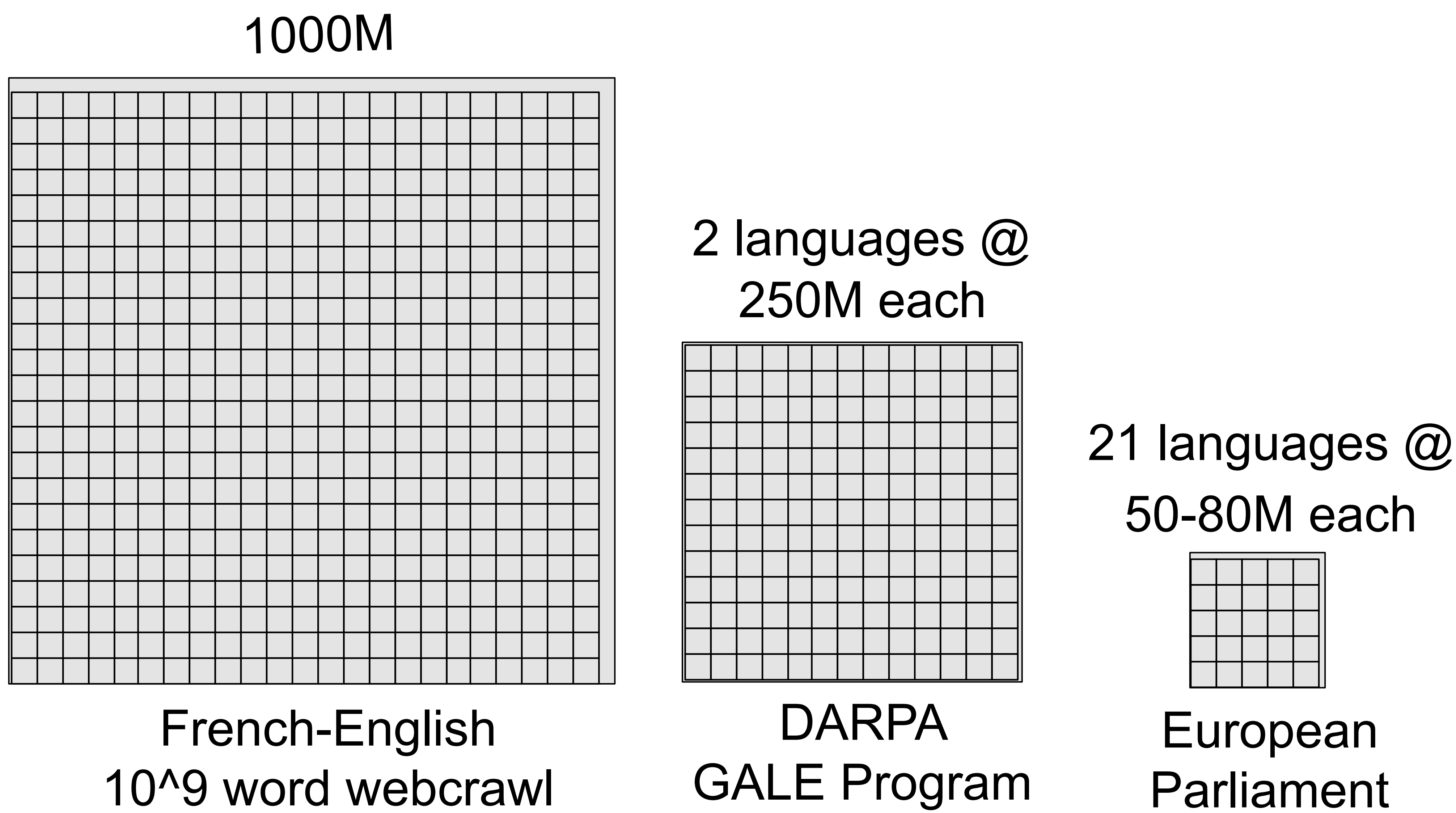
Strong meaning equivalence signal

... but different languages.

Bilingual Pivoting



Large, diverse sets of bilingual training data



Wide range of paraphrases

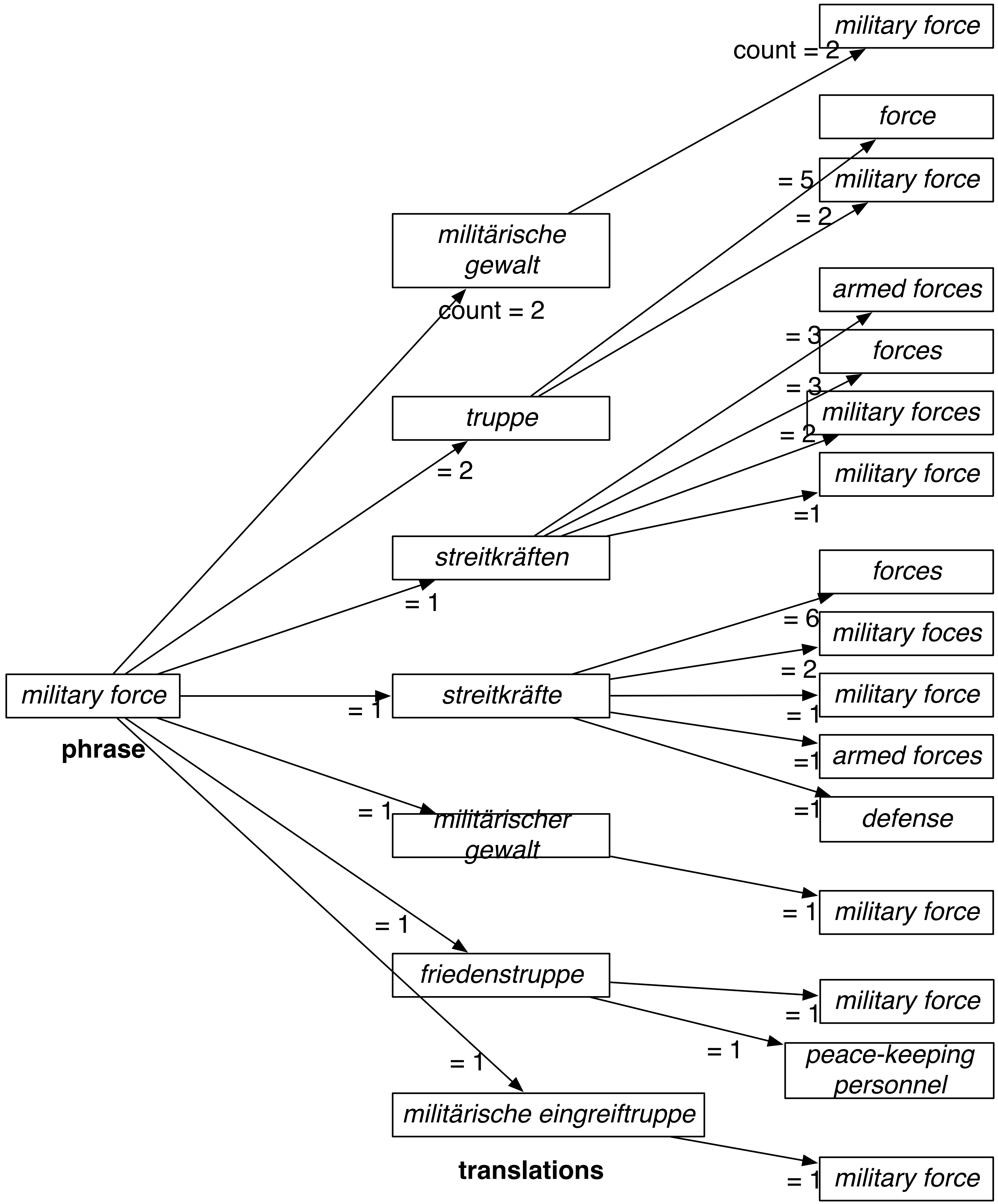
thrown into jail

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	

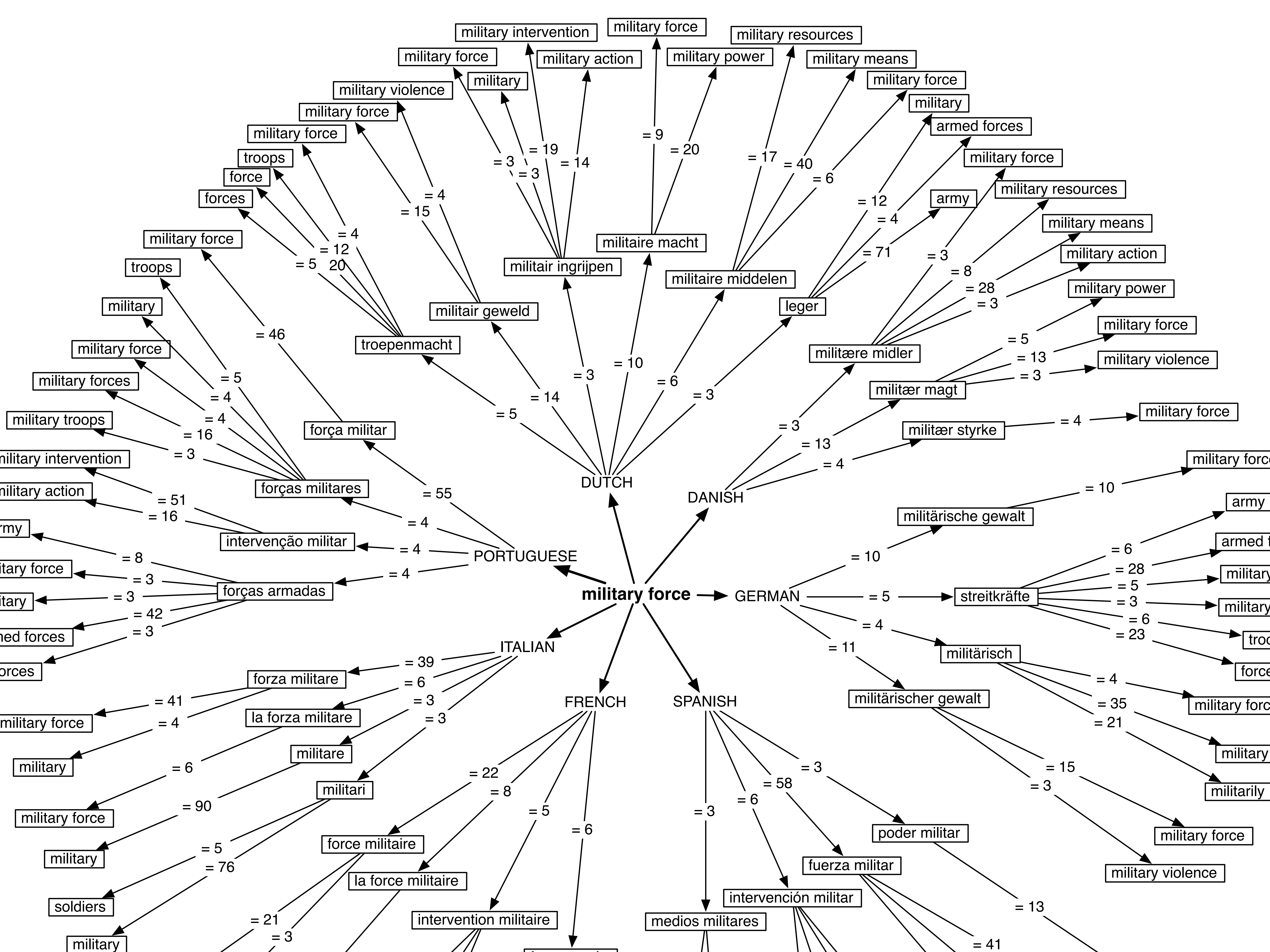
Paraphrase Probability

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

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



military force
count = 2
force
= 5
military force
= 2
militärische gewalt
count = 2
armed forces
= 3
forces
= 3
military forces
= 2
truppe
= 2
military force
=1
streitkräften
forces
= 1
= 6
military foces
= 2
military force
streitkräfte
= 1
= 1
military force
phrase
armed forces
=1
=1
defense
= 1
militärischer gewalt
military force
= 1
= 1
friedenstruppe
military force
= 1
= 1
peace-keeping personnel
= 1
militärische eingreiftruppe
translations
= 1
military force



military force
DUTCH
DANISH
GERMAN
SPANISH
FRENCH
ITALIAN
PORTUGUESE
troepenmacht
militair geweld
militair ingrijpen
militaire macht
militaire middelen
leger
= 5
= 14
= 3
= 10
= 6
= 3
forces
force
troops
military force
military force
military violence
military force
military
military intervention
military action
military force
military power
military resources
military means
military force
military
armed forces
army
= 5
= 12
= 20
= 4
= 15
= 4
= 3
= 3
= 19
= 14
= 9
= 20
= 17
= 40
= 6
= 12
= 4
= 71
militære midler
militær magt
militær styrke
= 3
= 13
= 4
military force
military resources
military means
military action
military power
military force
military violence
military force
= 3
= 8
= 28
= 3
= 5
= 13
= 3
= 4
militärische gewalt
streitkräfte
militärisch
militärischer gewalt
= 10
= 5
= 4
= 11
army
= 10
= 6
= 28
= 5
= 3
= 6
= 23
= 4
= 35
= 21
militarily
= 15
= 3
military force
military violence
poder militar
fuerza militar
intervención militar
medios militares
= 3
= 58
= 6
= 3
= 13
= 41
force militaire
la force militaire
intervention militaire
= 22
= 8
= 5
= 6
= 21
= 3
forza militare
la forza militare
militare
militari
= 39
= 6
= 3
= 3
= 41
= 4
= 6
= 90
= 5
= 76
military force
military
military force
military
soldiers
military
força militar
forças militares
intervenção militar
forças armadas
= 55
= 4
= 4
= 4
= 46
military force
troops
military
military force
military forces
military troops
= 5
= 4
= 4
= 16
= 3
= 51
= 16
= 8
= 3
= 3
= 42
= 3

Syntactic constraints

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~~

~~arrest~~
~~cases~~
~~custody~~
maltreated
~~owners~~
~~protection~~
~~thrown~~

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

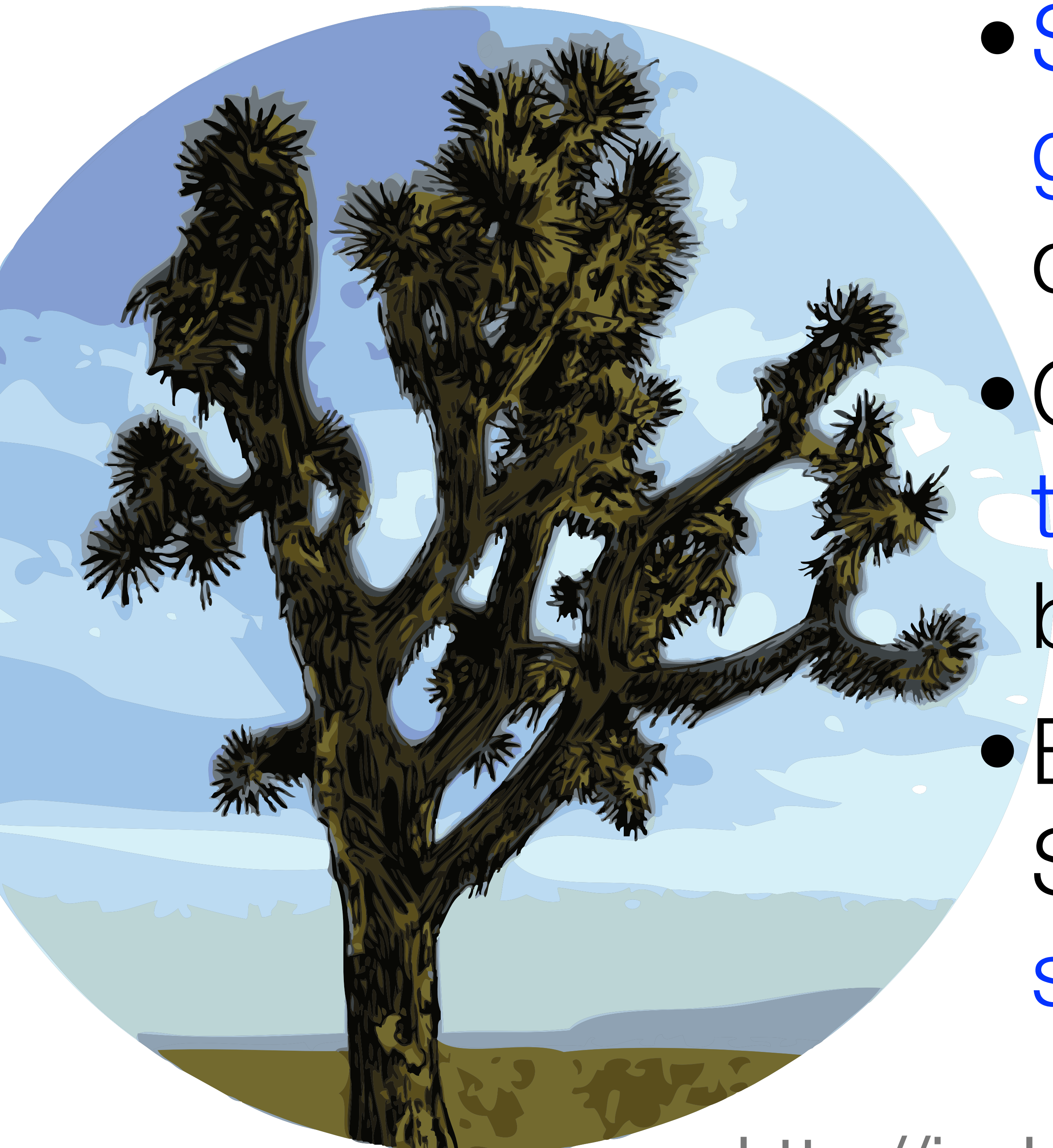
Sentential paraphrases from bitexts?

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?

Syntactic MT in the Joshua Decoder

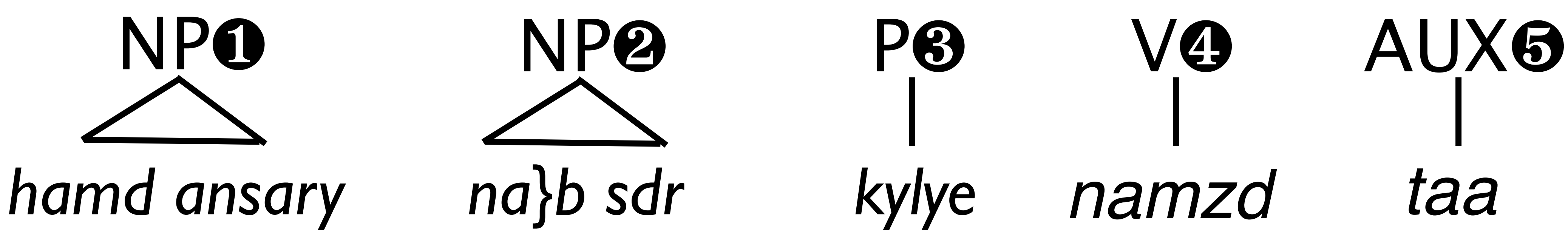


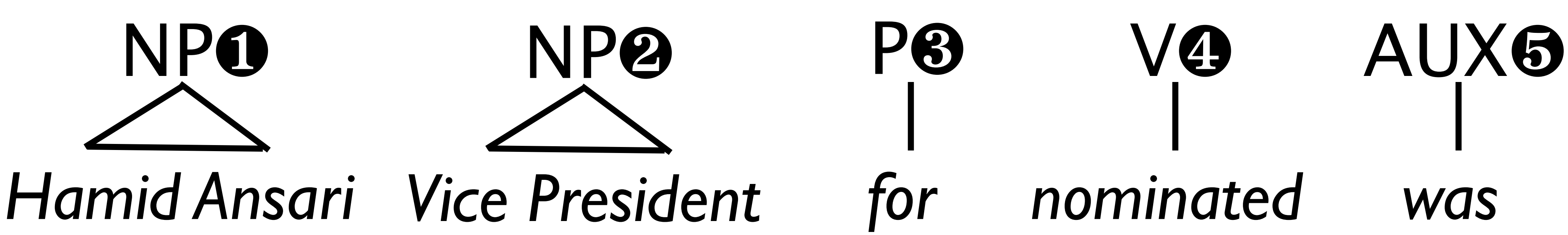
- Synchronous context free grammars generate pairs of corresponding strings
- Can be used to describe translation and re-ordering between languages
- Because Joshua uses SCFGs, it translates sentences by parsing them

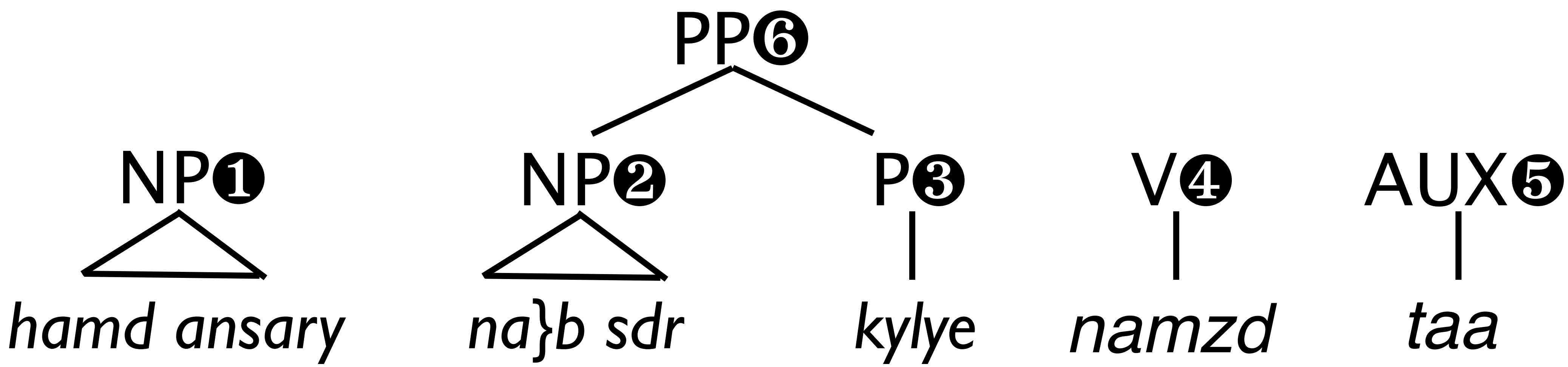
http://joshua-decoder.org

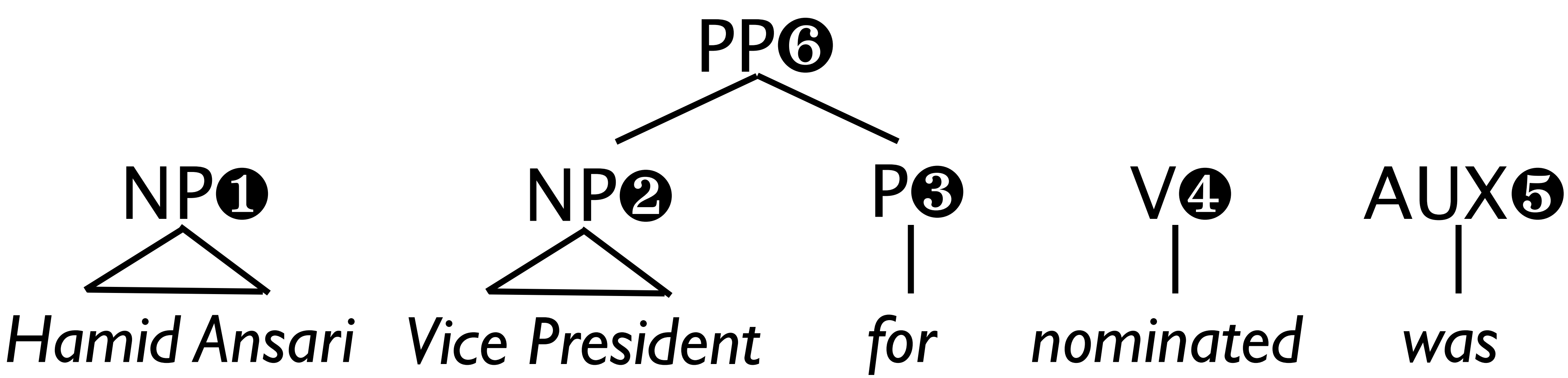
Translation

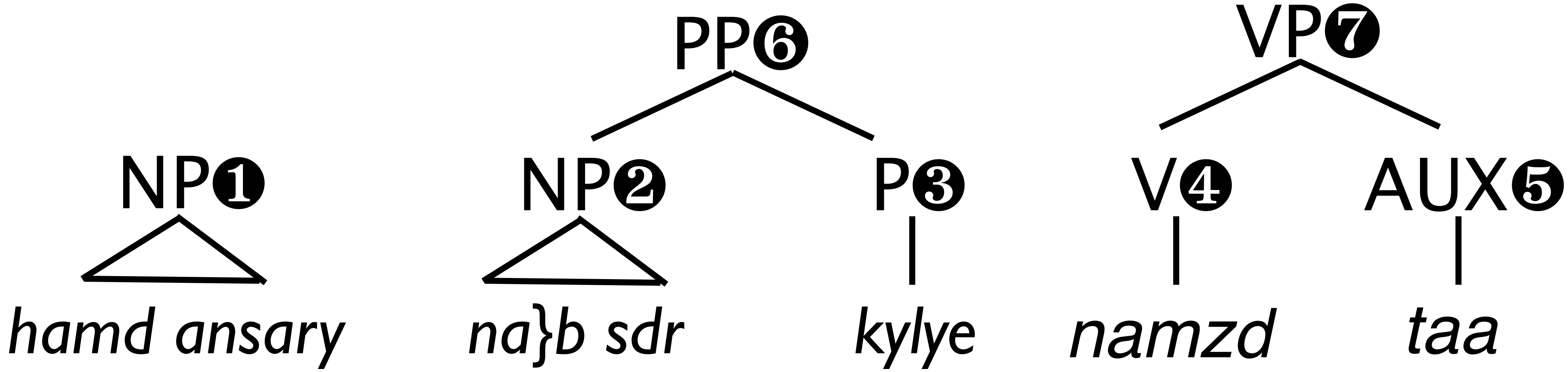
	Urdu	English
S →	NP① VP②	NP① VP②
VP →	PP① VP②	VP② PP①
VP →	V① AUX②	AUX② V①
PP →	NP① P②	P② NP①
NP →	*hamd ansary*	*Hamid Ansari*
NP →	*na}b sdr*	*Vice President*
V →	*namzd*	*nominated*
P →	*kylye*	*for*
AUX →	*taa*	*was*

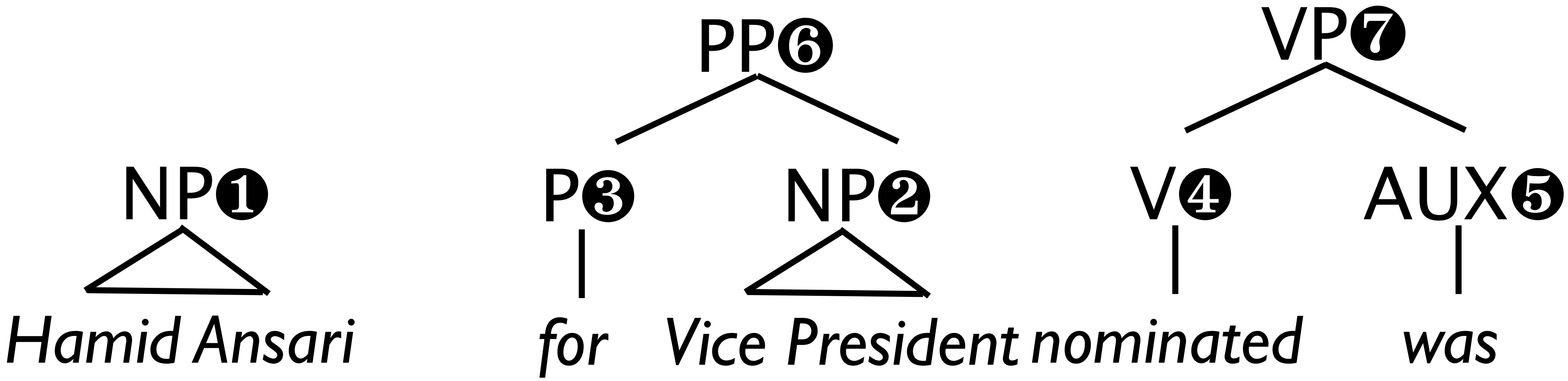

NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

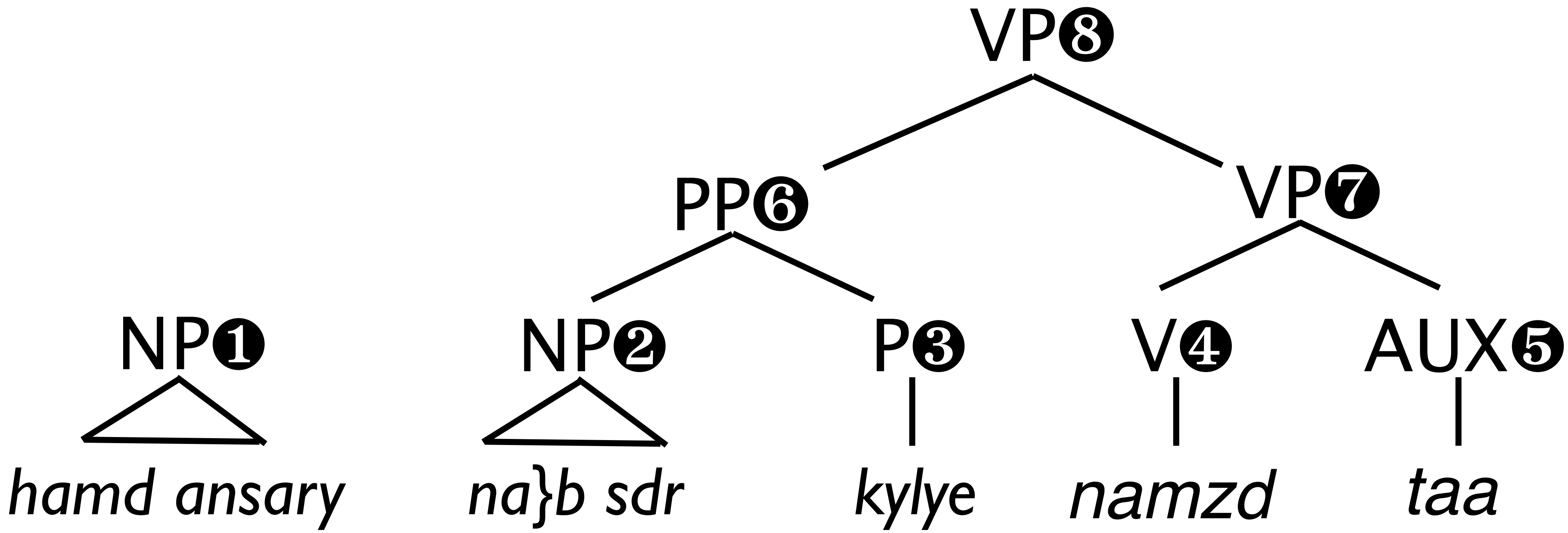

NP❶
NP❷
P❸
V❹
AUX❺
Hamid Ansari
Vice President
for
nominated
was


PP❻
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

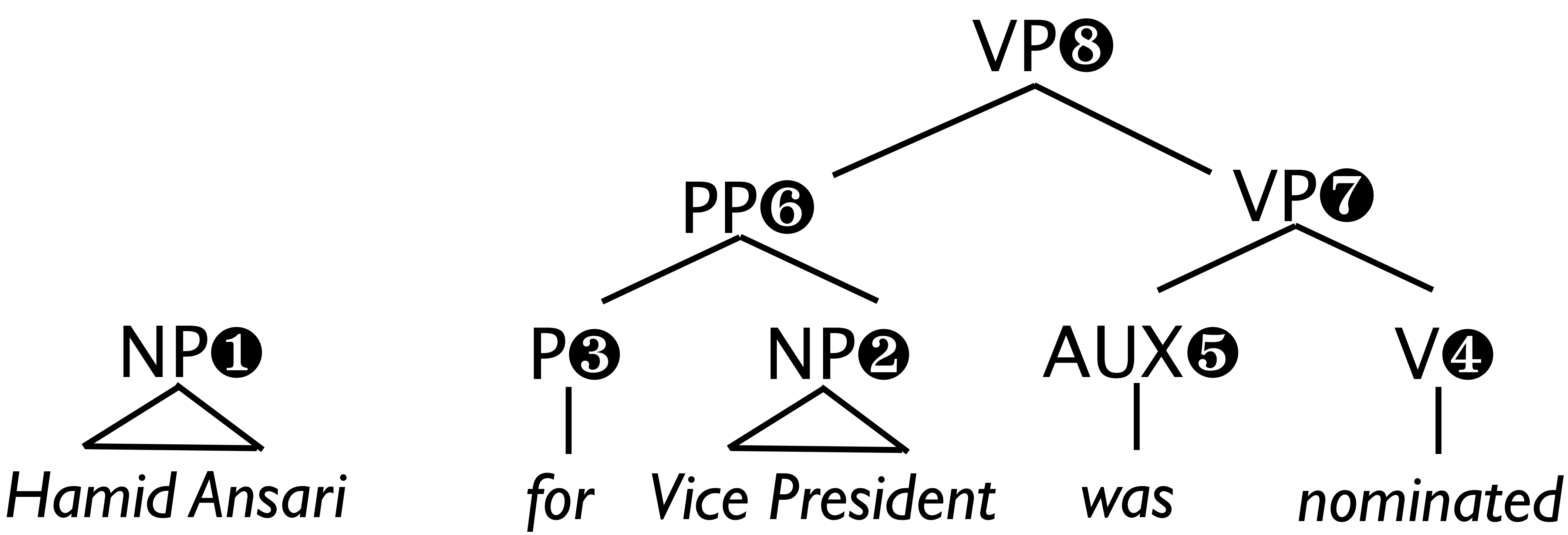

PP❻
NP❶
NP❷
P❸
V❹
AUX❺
Hamid Ansari
Vice President
for
nominated
was


PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

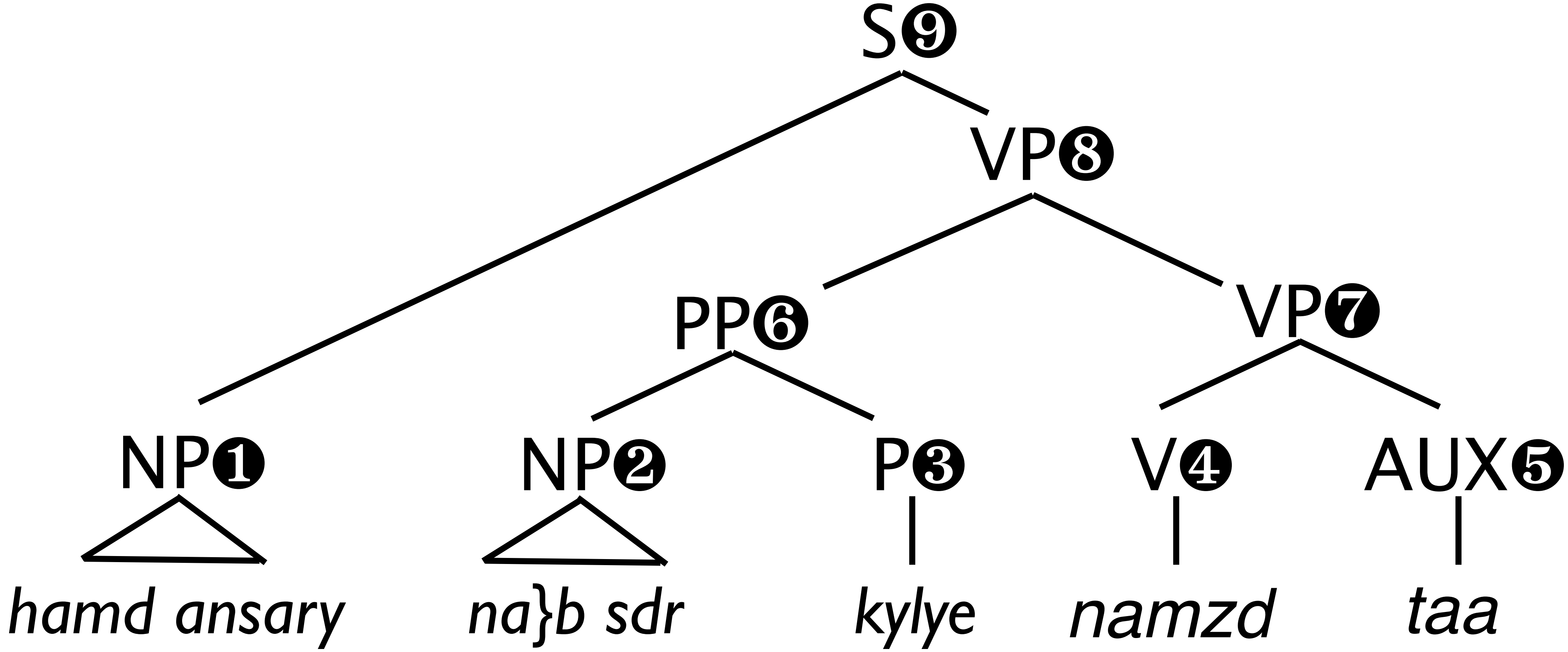

PP❻
VP❼
NP❶
P❸
NP❷
V❹
AUX❺
Hamid Ansari
for
Vice President
nominated
was



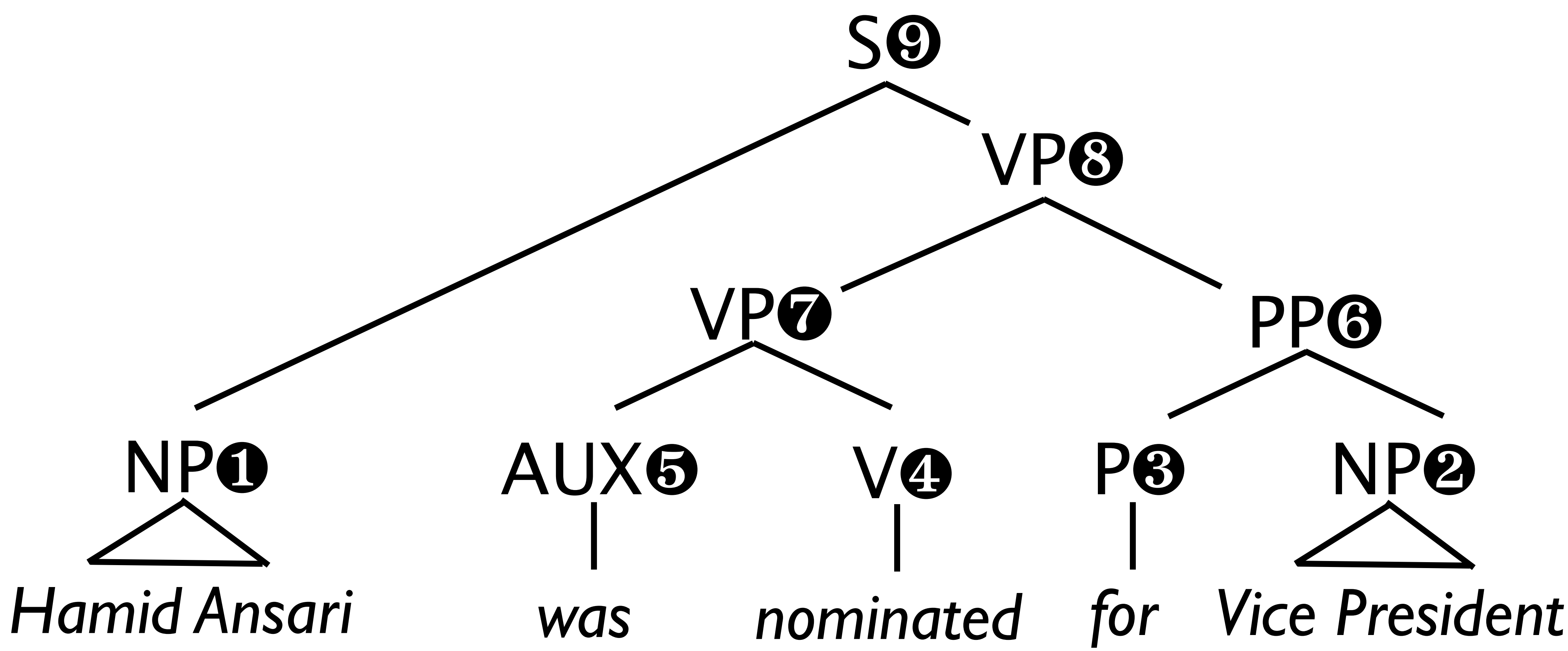
VP❽
PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa



VP❽
PP❻
VP❼
NP❶
P❸
NP❷
AUX❺
V❹
Hamid Ansari
for
Vice President
was
nominated



S❾
VP❽
PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa



S❾
VP❽
VP❼
PP❻
NP❶
AUX❺
V❹
P❸
NP❷
Hamid Ansari
was
nominated
for
Vice President

SCFGs via Pivoting

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

$$NP \rightarrow \boxed{NP\text{ 's }NN} \mid \text{le NN de NP}$$

$$NP \rightarrow \boxed{\text{the NN of NP}} \mid \text{le NN de NP}$$

combine to

$$NP \rightarrow \boxed{NP\text{ 's }NN} \mid \boxed{\text{the NN of NP}}$$

Possessive rule	NP →	the NN of the NNP	the NNP’s NN
	NP →	the NNS$_1$ made by the NNS$_2$	the NNS$_2$’s NNS$_1$
Dative shift	VP →	give NN to NP	give NP the NN
	VP →	provide NP$_1$ to NP$_2$	give NP$_2$ NP$_1$
Adv. | adj. phrase move	S | VP →	ADVP they VBD	they VBD ADVP
	S →	it is ADJP VP	VP is ADJP
Verb particle shift	VP →	VB NP up	VB up NP
Reduced relative clause	SBAR | S	although PRP VBD that	although PRP VBD
	ADJP →	very JJ that S	JJ S
Partitive constructions	NP →	CD of the NN	CD NN
	NP →	all DT\NP	all of the DT\NP
Topicalization	S →	NP, VP.	VP, NP.
Passivization	SBAR →	that NP had VBN	which was VBN by NP
Light verbs	VP →	take action ADVP	to act ADVP
	VP →	to make a decision PP	to decide PP

Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation.
Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. EMNLP 2011.

Text-to-Text Generation

T2T involves generating **meaning-equivalent text** that is *subject to some constraints*:

sentence compression, *shorter*

simplification, *easier to understand*

poetry from prose, *rhyme and meter*

Sentence Compression

Reduce length of a sentence (#chars) while retaining the meaning

Compression ratio: $$\varphi = \frac{length_{compression}}{length_{original}}$$

Paraphrasing as a task and problem is of paramount importance to a multitude of applications in the field of NLP.

Sentence Compression

Reduce length of a sentence (#chars) while retaining the meaning

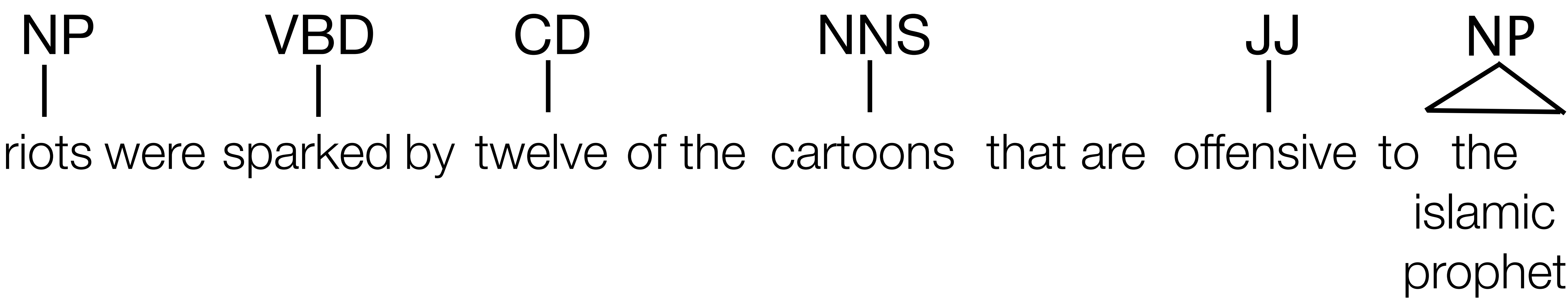
Compression ratio: $\varphi = \dfrac{length_{compression}}{length_{original}}$

is awesome

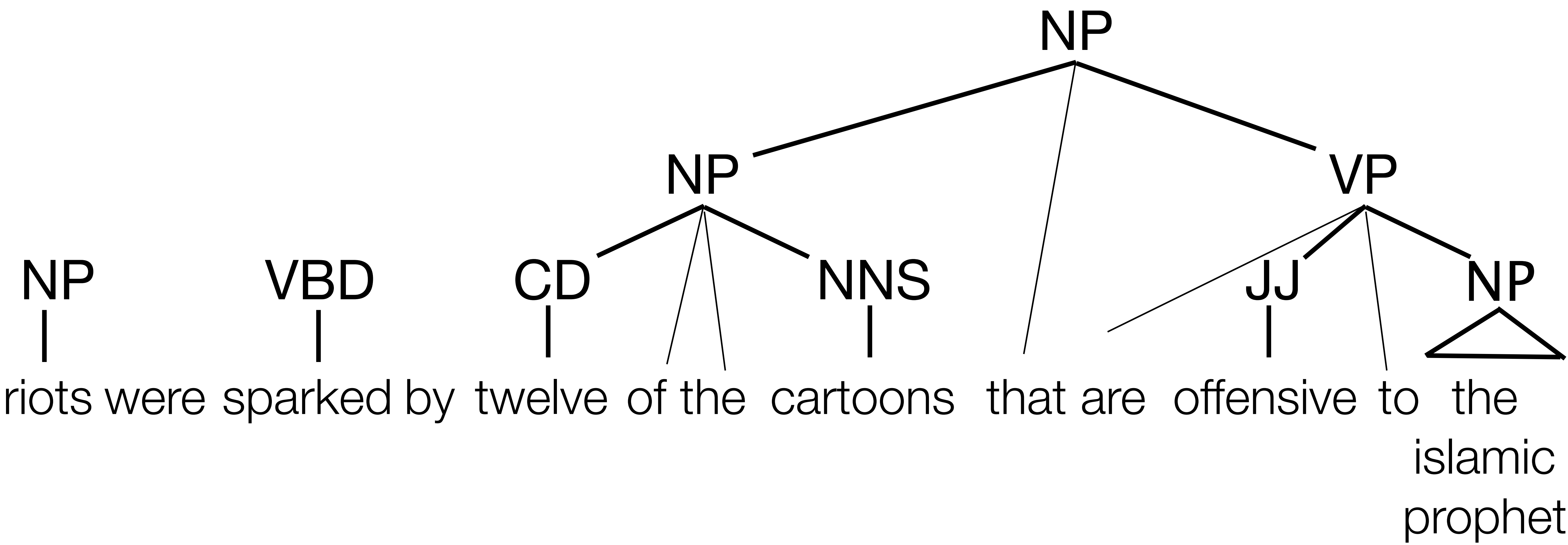
Paraphrasing ~~as a task and problem is of paramount importance to a multitude of applications in the field of NLP~~.

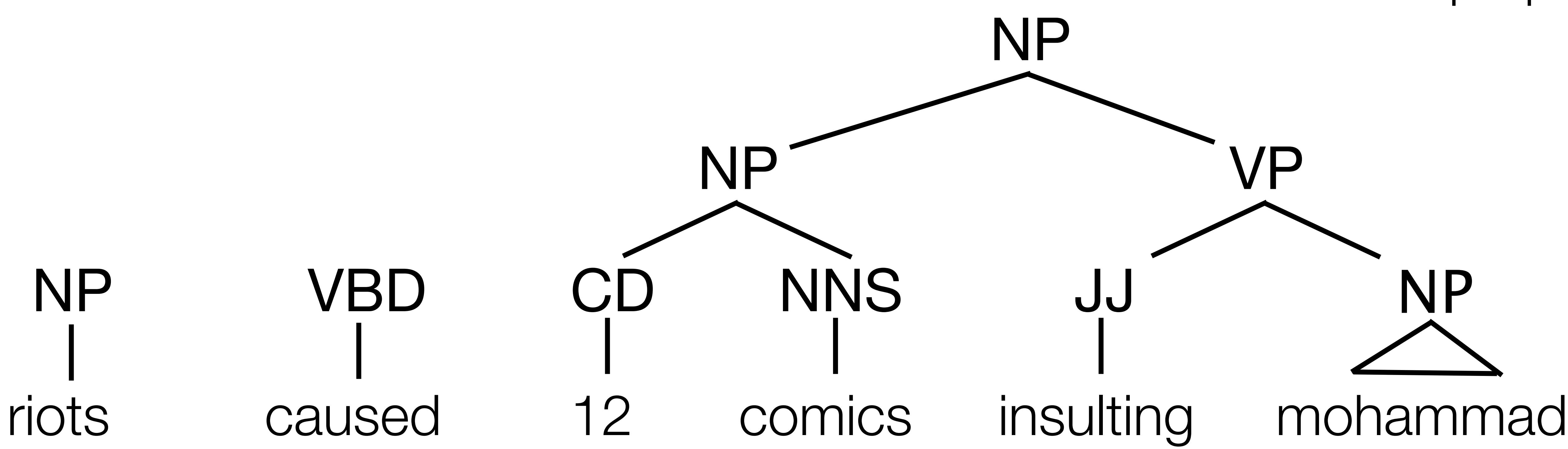
Paraphrase Grammar

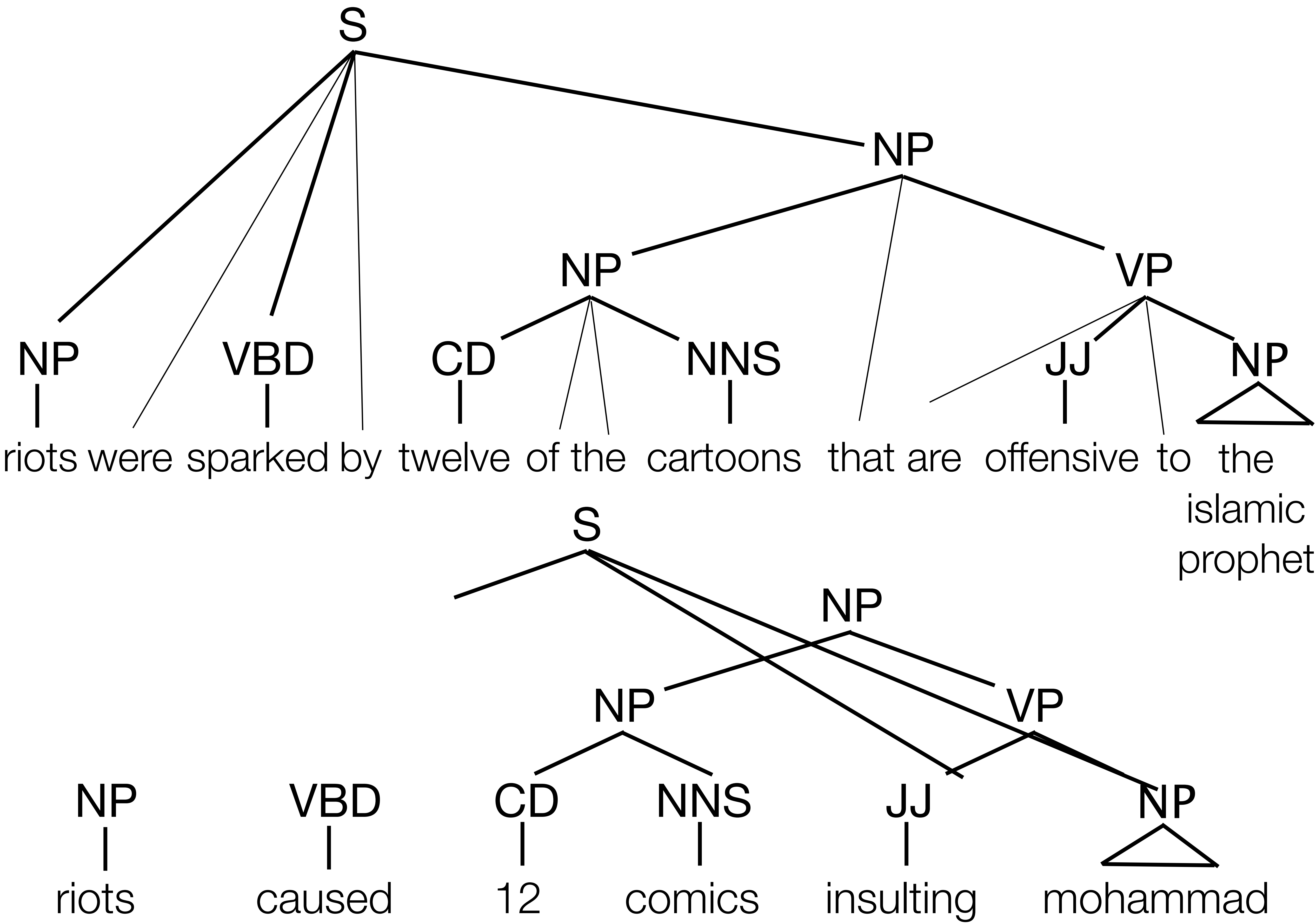
	English	English
S →	NP① were VBD by NP②	NP② VBD NP①
NP→	NP that VP	NP VP
VP→	are JJ to NP	JJ NP
NP →	CD of the NNS	CD NNS
CD →	twelve	12
NNS →	cartoons	comics
JJ →	offensive	insulting
NP →	the islamic prophet	mohammed
VBD →	sparked	caused

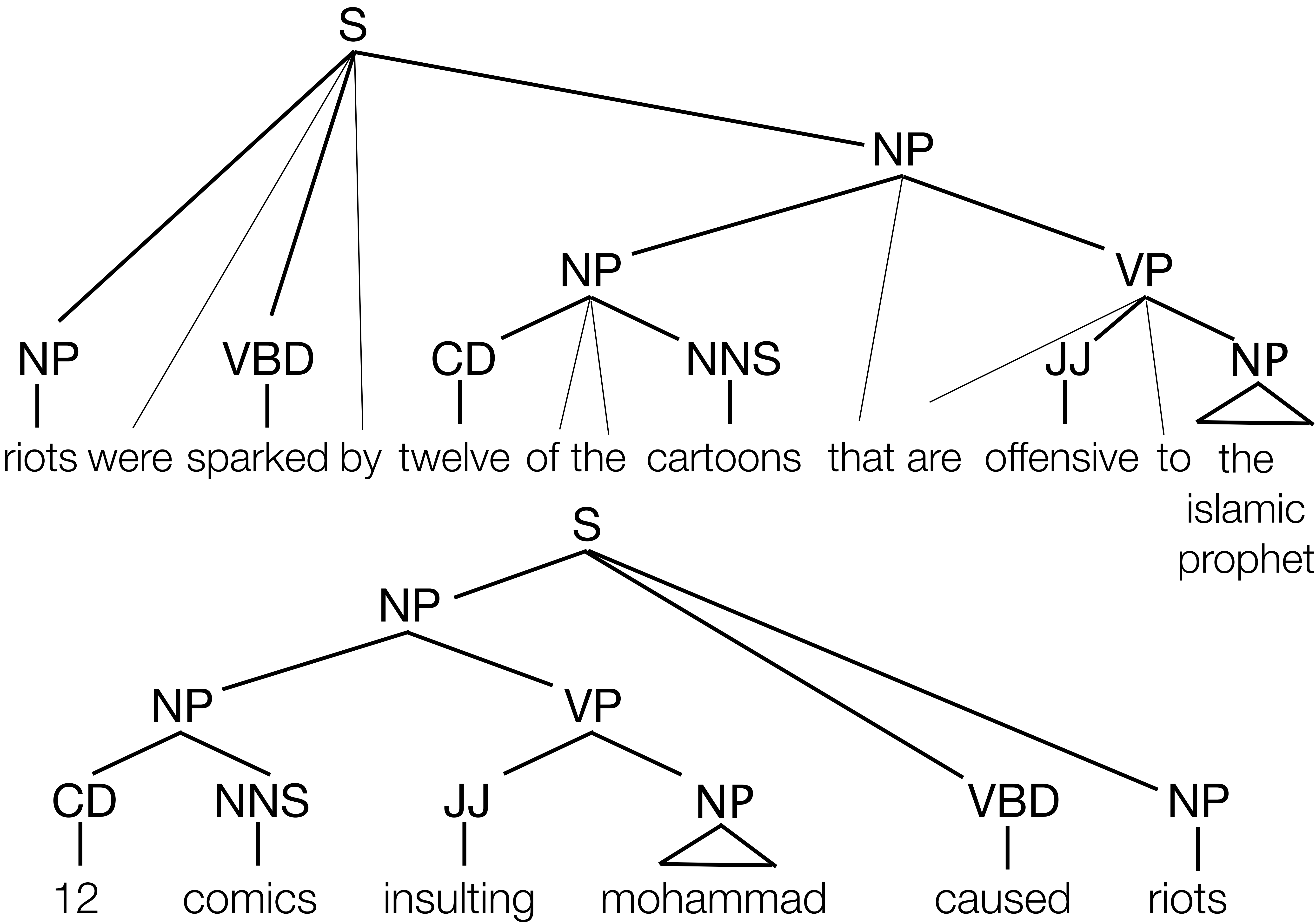

NP VBD CD NNS JJ NP
riots were sparked by twelve of the cartoons that are offensive to the islamic prophet


NP VBD CD NNS JJ NP
riots caused 12 comics insulting mohammad


NP
NP
VP
NP
VBD
CD
NNS
JJ
NP
riots were
sparked by
twelve
of the
cartoons
that are
offensive
to
the islamic prophet


NP
NP
VP
NP
VBD
CD
NNS
JJ
NP
riots
caused
12
comics
insulting
mohammad


S
NP
NP
VP
NP
VBD
CD
NNS
JJ
NP
riots were sparked by twelve of the cartoons that are offensive to the islamic prophet
S
NP
NP
VP
NP
VBD
CD
NNS
JJ
NP
riots
caused
12
comics
insulting
mohammad


S
NP
NP
VBD
CD
NNS
VP
JJ
NP
riots were sparked by twelve of the cartoons that are offensive to the islamic prophet
S
NP
NP
VP
CD
NNS
JJ
NP
VBD
NP
12 comics insulting mohammad caused riots

Text-to-Text Applications

Claim:

Paraphrasing is suitable to tackle sentential text-to-text tasks, and we can re-use SMT machinery for T2T.

However:

Naive application of MT techniques will not work, need to **adapt them**

Task Adaptation

SMT	T2T
Naive application of the MT machinery to the task	Task-specific adaptations

- Development data
- Objective function
- Feature set
- Grammar augmentations

Development Data

SMT	T2T
English reference translations that are used to calculate BLEU for SMT.	Selected pairs of reference translations that significantly differ in length.

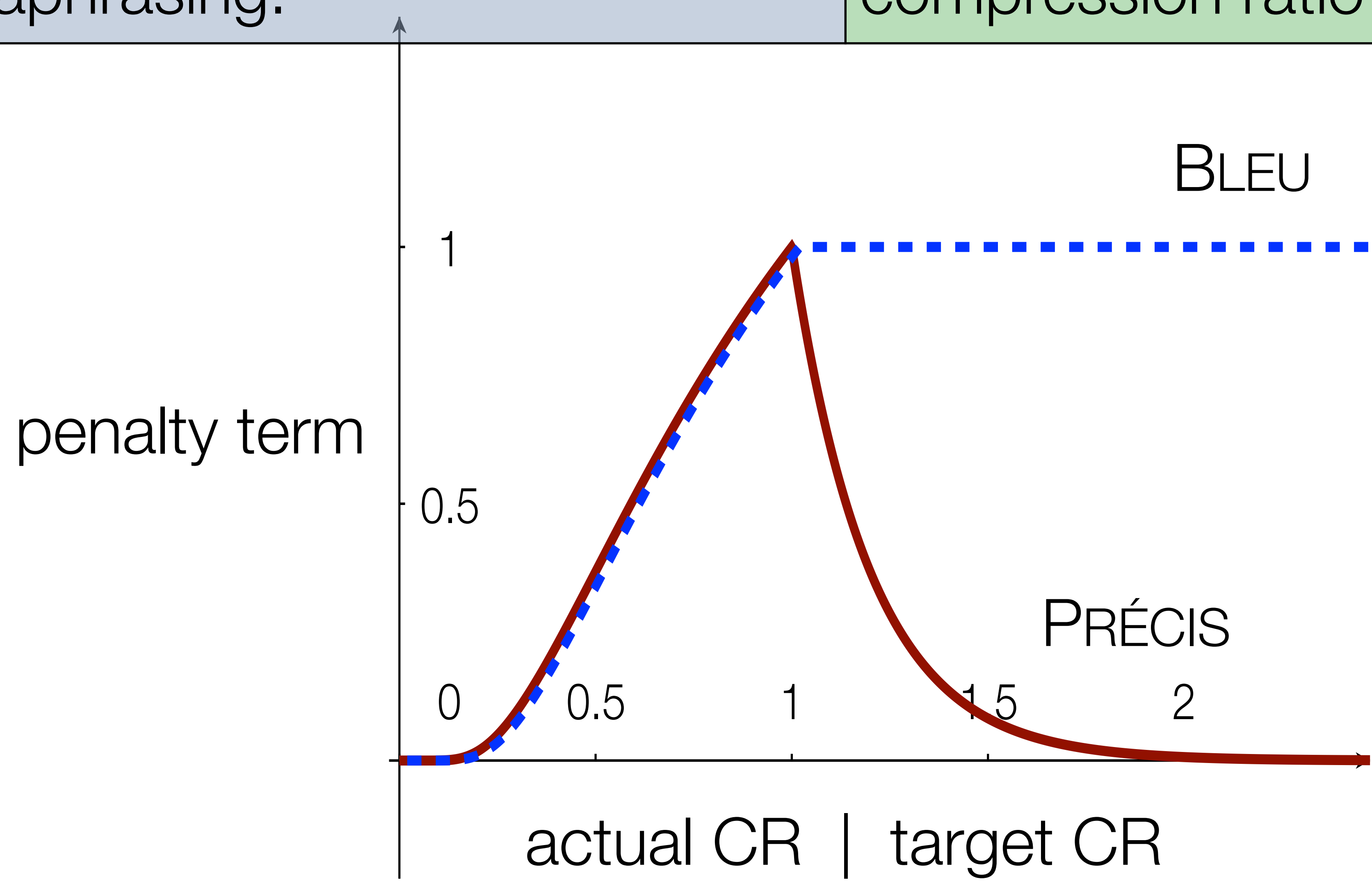
and he said that the project **will cover** the needs of the region **in the long term**. 82

he said the project **includes** all the district's **long-term** needs. 65

compression ratio = 0.79

Objective Function

SMT	T2T
Optimized for English-to-English BLEU score. Causes self-paraphrasing.	Add a "verbosity penalty" to BLEU that allows a target compression ratio to be set.



Features

SMT	T2T
Phrasal and lexical probabilities quantify general paraphrase quality.	Features counting number of source and target words and the difference between them.

$$\text{VP} \rightarrow \text{NP was eaten by NN} \quad | \quad \text{NN ate NP}$$

$$p(e_1|e_2) = 0.1$$

$$c_{e_1} = 14$$
$$c_{e_2} = 5$$
$$c_{diff} = -9$$

$$logCR = \log \frac{c_{e_1}}{c_{e_2}}$$

Augmentations

SMT	T2T
It is not typical for additional task-specific rules to be added in the standard SMT pipeline.	Augment the grammar with deletion rules for specific POS (JJ, RB, DT) allowing for shorter compressions.

$$\text{JJ} \rightarrow \text{superfluous} \mid \varepsilon$$

$$\text{RB} \rightarrow \text{redundantly} \mid \varepsilon$$

$$\text{DT} \rightarrow \text{the} \mid \varepsilon$$

Monolingually-derived Features

SMT	T2T
All features, aside from the LM, are bilingually derived.	Calculate distributional similarity of paraphrase pairs from monolingual data

Orthogonal signal to bilingual pivoting

Even more data available

Incorporated as features in T2T model

Distributional Similarity

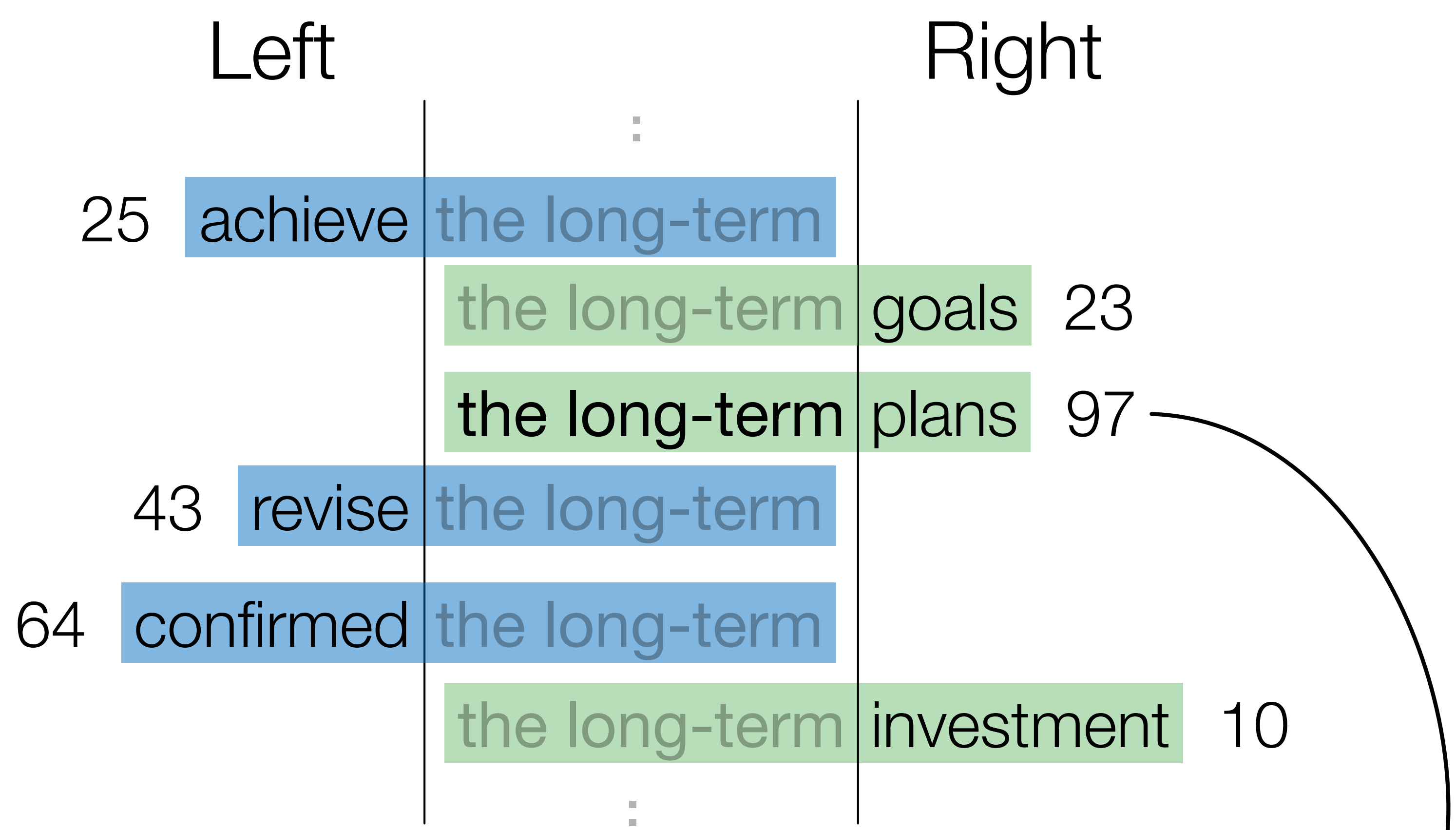
Idea: similar words occur in similar contexts.

Characterize words by their contexts

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

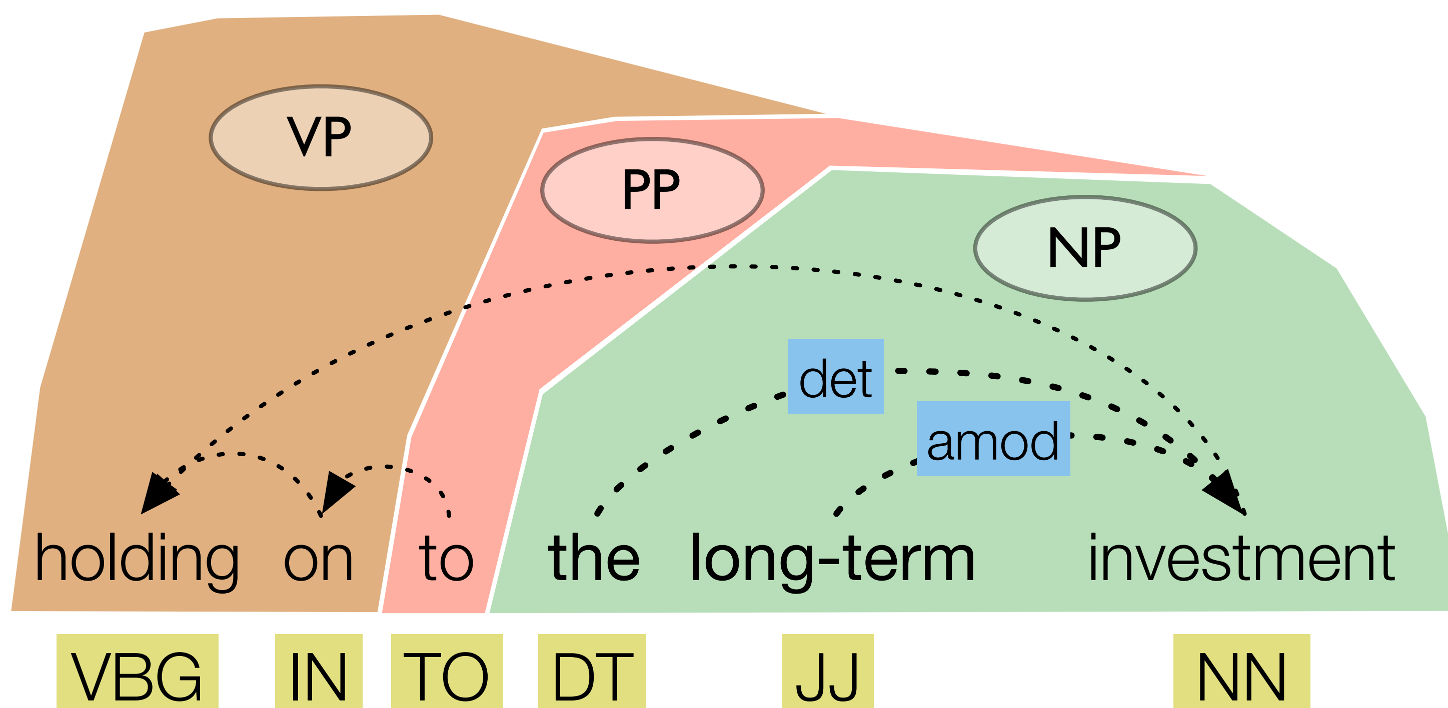
“Are these paraphrases substitutable?”

n-gram Context



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

Syntactic Context



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

Large Monolingual Data Sets

Google n-grams

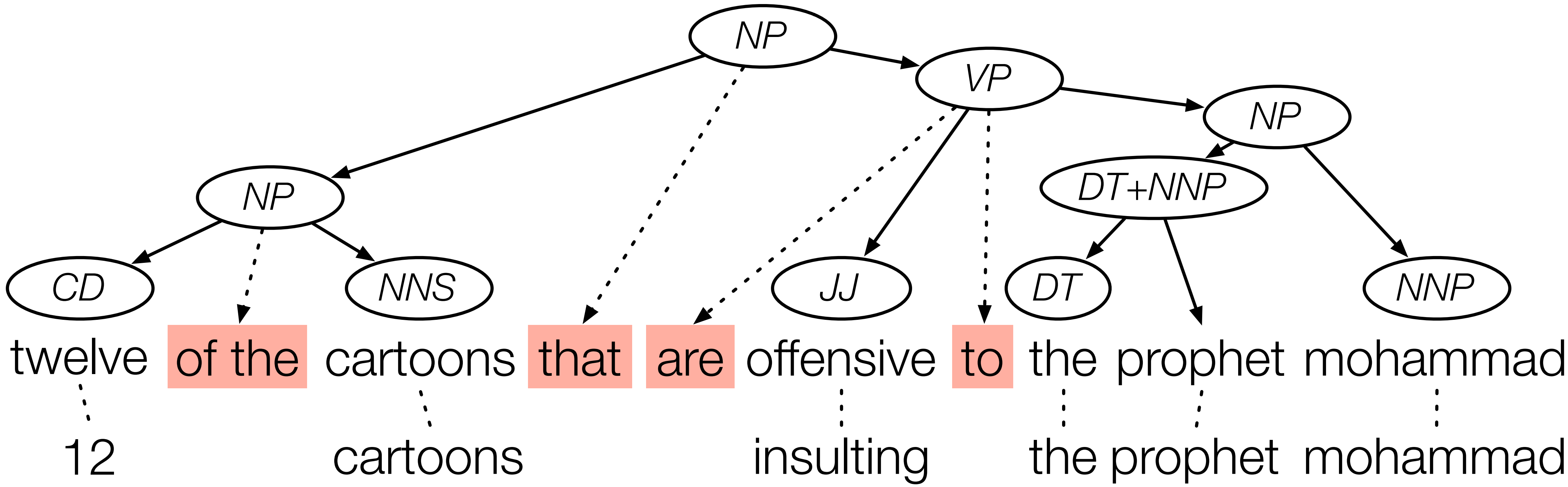
Collection of 1 trillion tokens with counts

Based on vast amounts of text

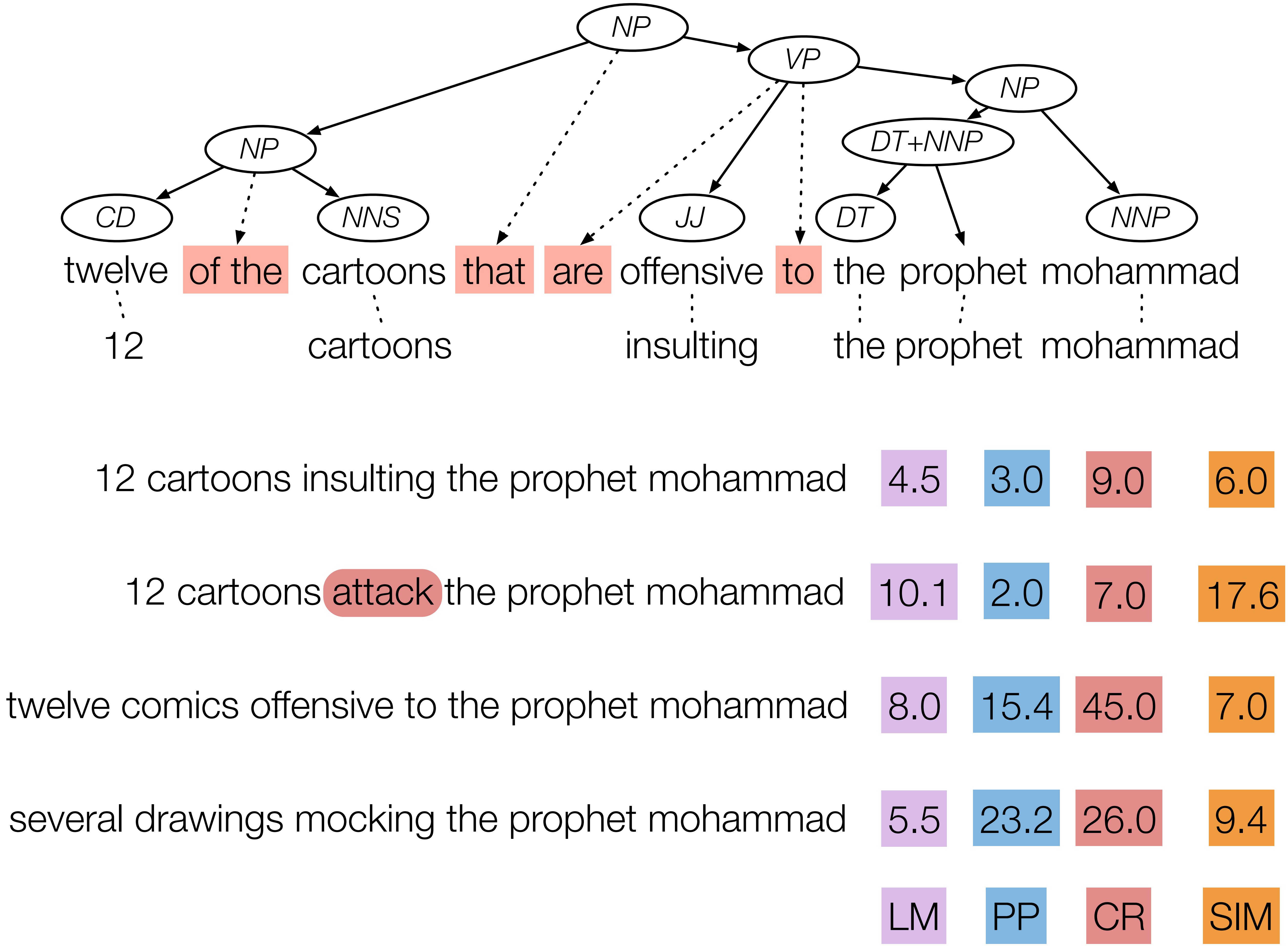
Annotated Gigaword (AKBC-WEKEX '12)

Collection of 4 billion words, parsed and tagged

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



	LM	PP	CR	SIM
12 cartoons insulting the prophet mohammad	4.5	3.0	9.0	6.0
12 cartoons attack the prophet mohammad	10.1	2.0	7.0	17.6
twelve comics offensive to the prophet mohammad	8.0	15.4	45.0	7.0
several drawings mocking the prophet mohammad	5.5	23.2	26.0	9.4


NP
VP
NP
NP
DT+NNP
CD
NNS
JJ
DT
NNP
twelve of the cartoons that are offensive to the prophet mohammad
12
cartoons
insulting
the prophet
mohammad
12 cartoons insulting the prophet mohammad 4.5 3.0 9.0 6.0
12 cartoons attack the prophet mohammad 10.1 2.0 7.0 17.6
twelve comics offensive to the prophet mohammad 8.0 15.4 45.0 7.0
several drawings mocking the prophet mohammad 5.5 23.2 26.0 9.4
LM
PP
CR
SIM

Task-based Evaluation

Evaluated paraphrases in the context of a T2T compression task.

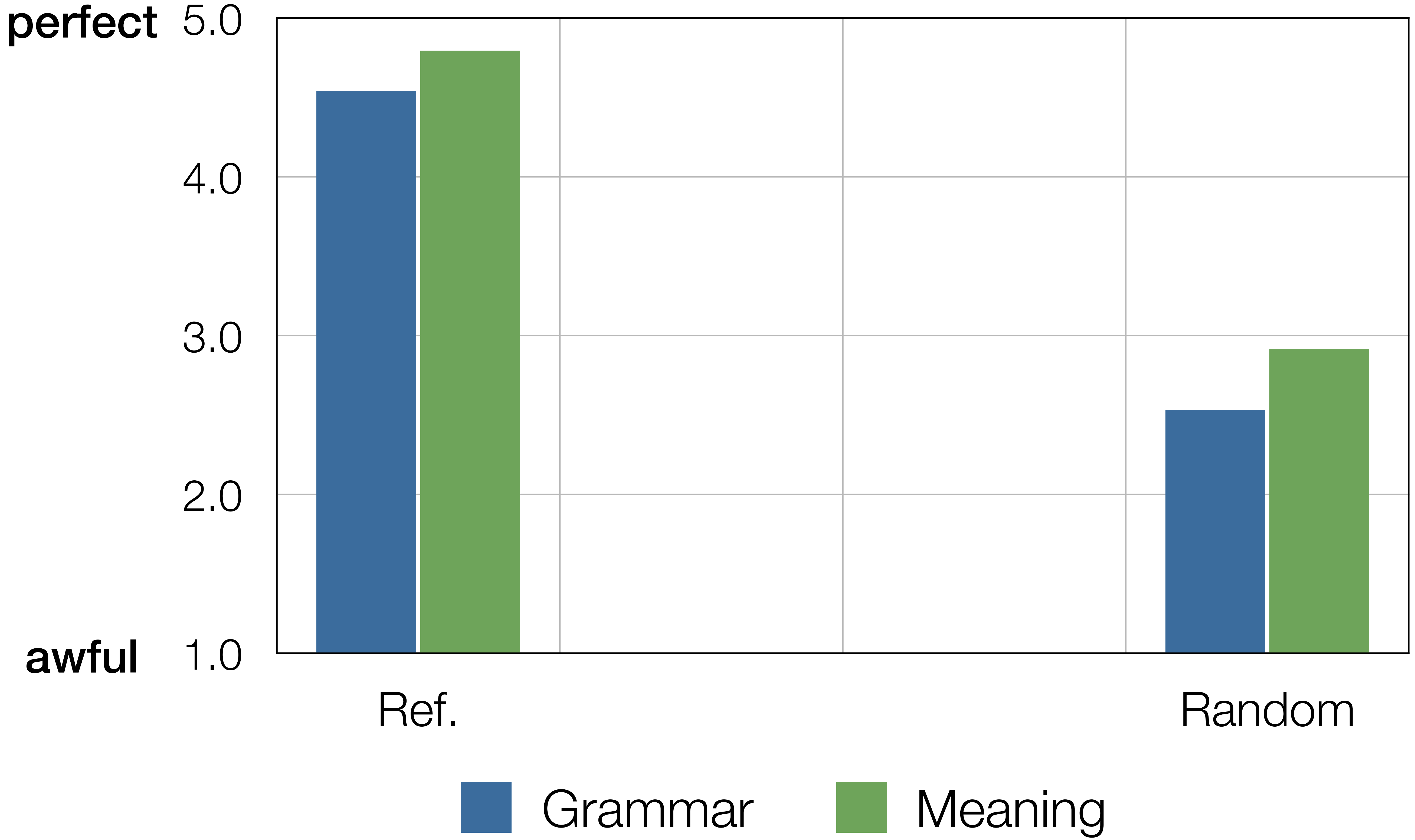
Compared against a state of the art system.

Human assessment (5-point scale):

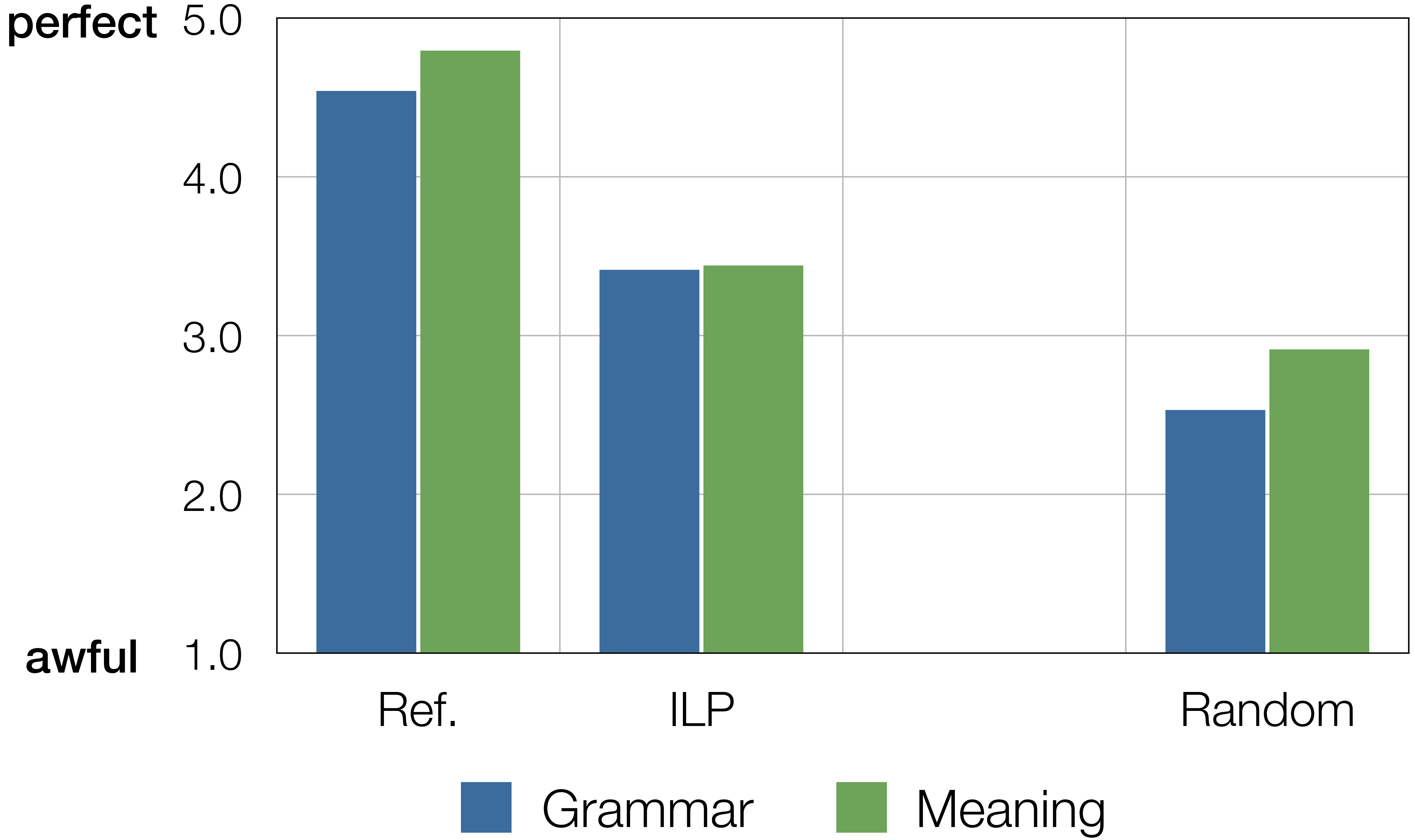
How well do these sentences retain the meaning of original?

How grammatical is the resulting sentence?

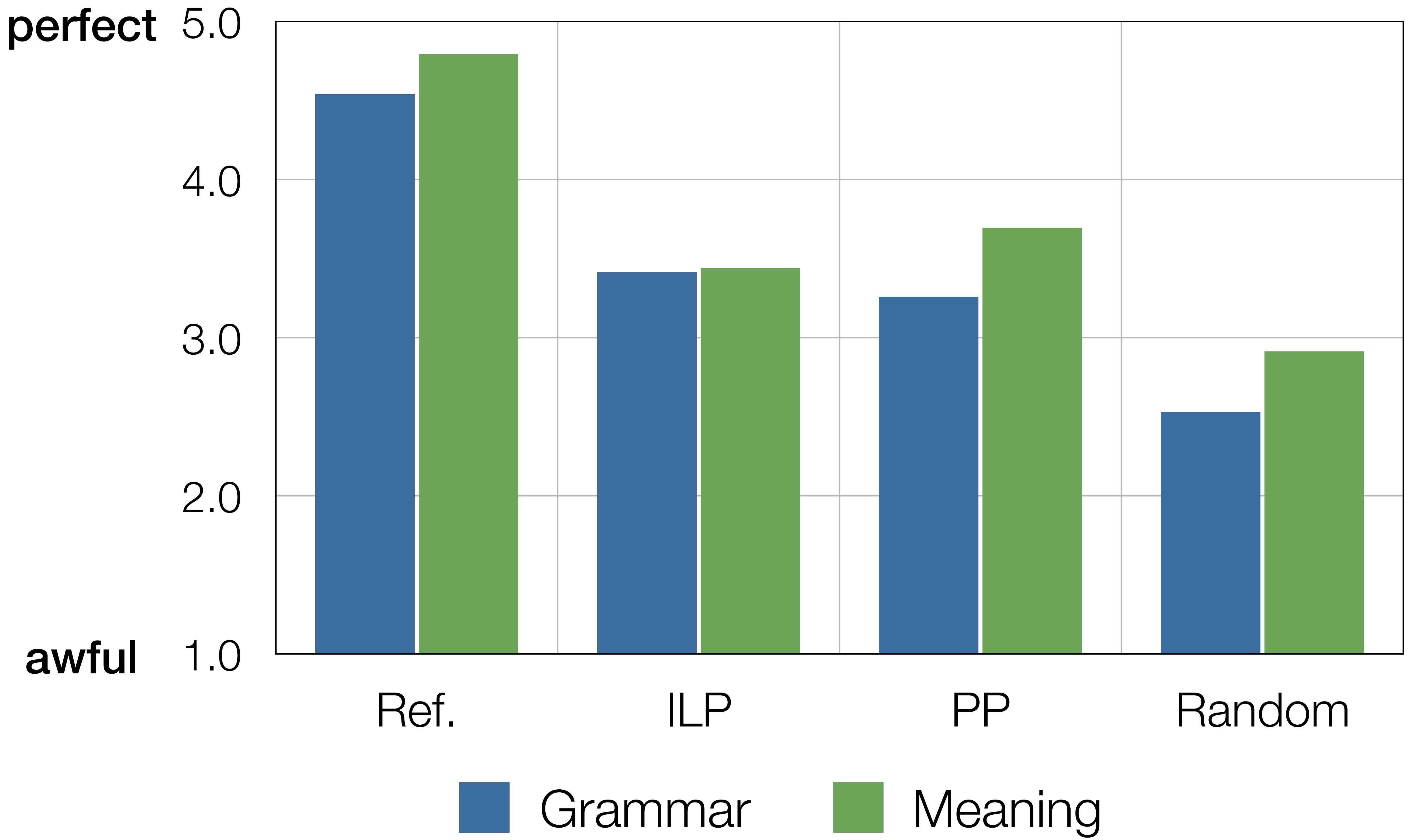
Compression Quality

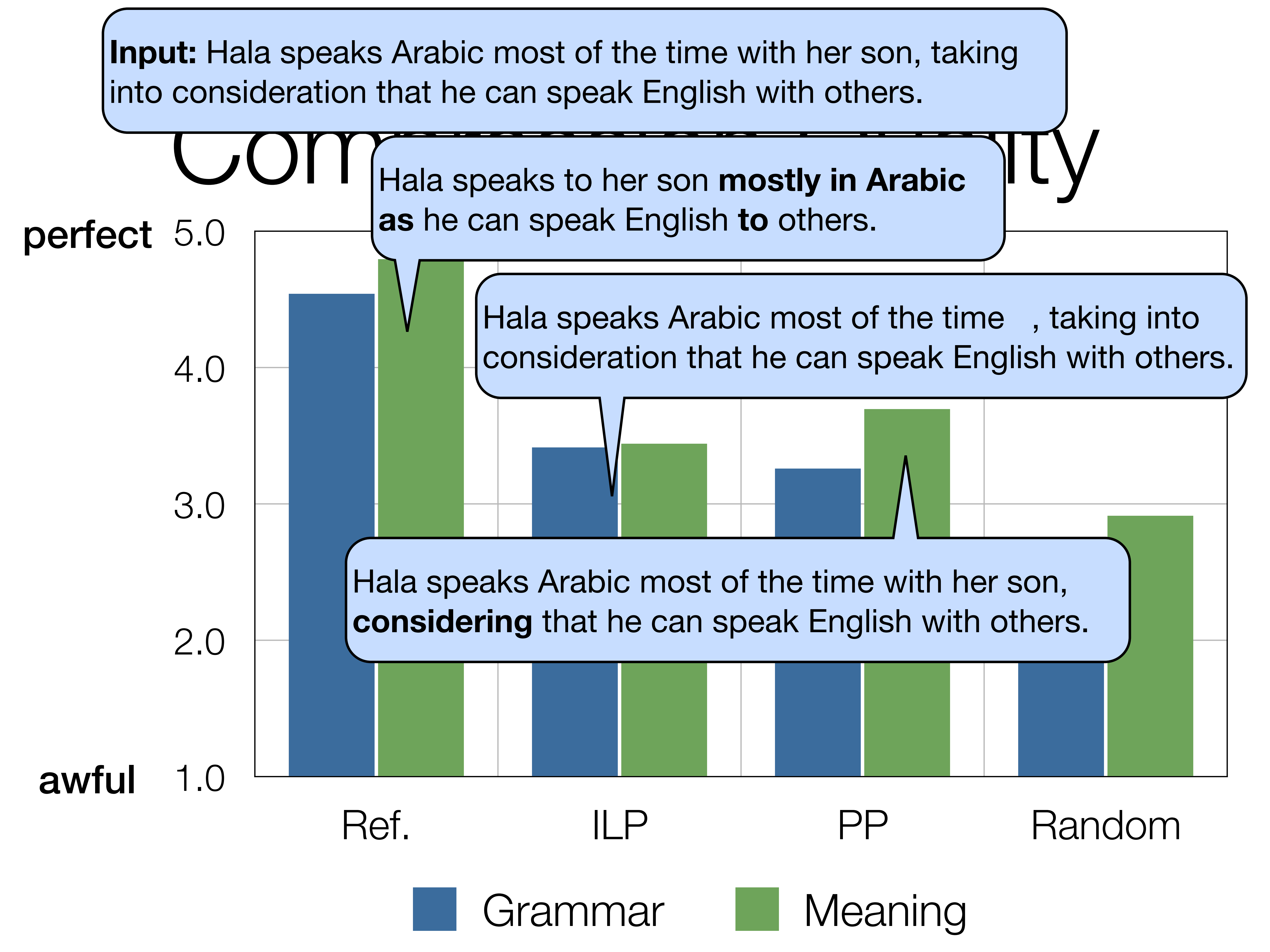


Compression Quality



Compression Quality




Input: Hala speaks Arabic most of the time with her son, taking into consideration that he can speak English with others.
Hala speaks to her son **mostly in Arabic as** he can speak English **to** others.
Hala speaks Arabic most of the time , taking into consideration that he can speak English with others.
Hala speaks Arabic most of the time with her son, **considering** that he can speak English with others.
perfect
awful
5.0
4.0
3.0
2.0
1.0
Ref.
ILP
PP
Random
Grammar
Meaning

Adaptation in 5 easy steps

Step	SMT to T2T Adaptation
1	**Dev data:** Collect a set of sentence pairs that reflects the task that you are trying to model
2	**Objective function:** Create a new objective function that indicates how well the system output the constraints of your task
3	**Task-specific features:** Add new features to the model that will allow it to score its own output for the task
4	**Augment the grammar:** Use your domain knowledge to add any rules that would not normally be contained in a paraphrase grammar.
5	**Other features:** Take advantage of the English to English to add other features that model grammaticality more generally.

Resources

Joshua Decoder

- An open source decoder that synchronous context free grammars to translate
- Implements all algorithms needed for translating with SCFGs
- Now available: "Language Packs" for T2T generation

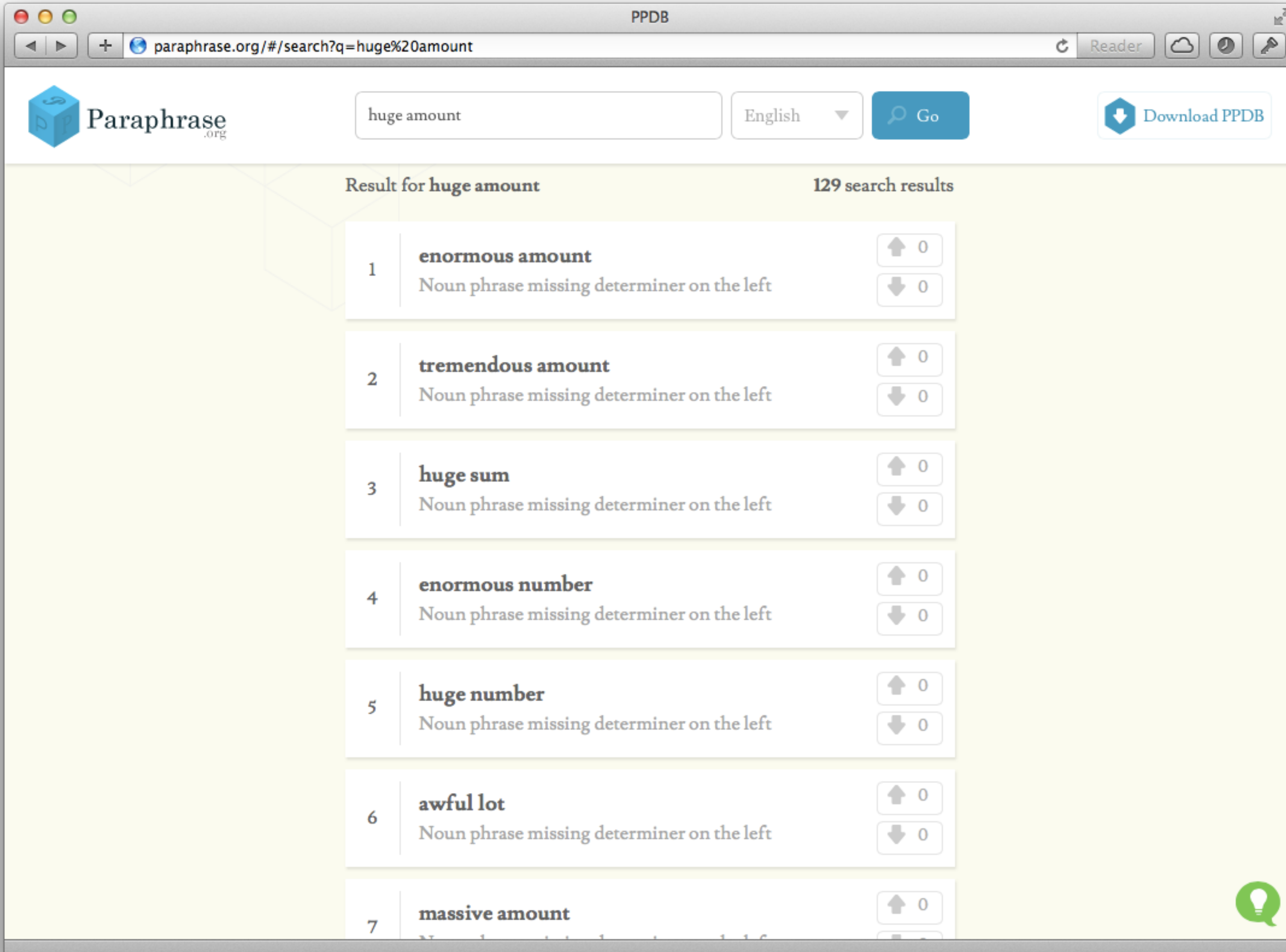
http://joshua-decoder.org

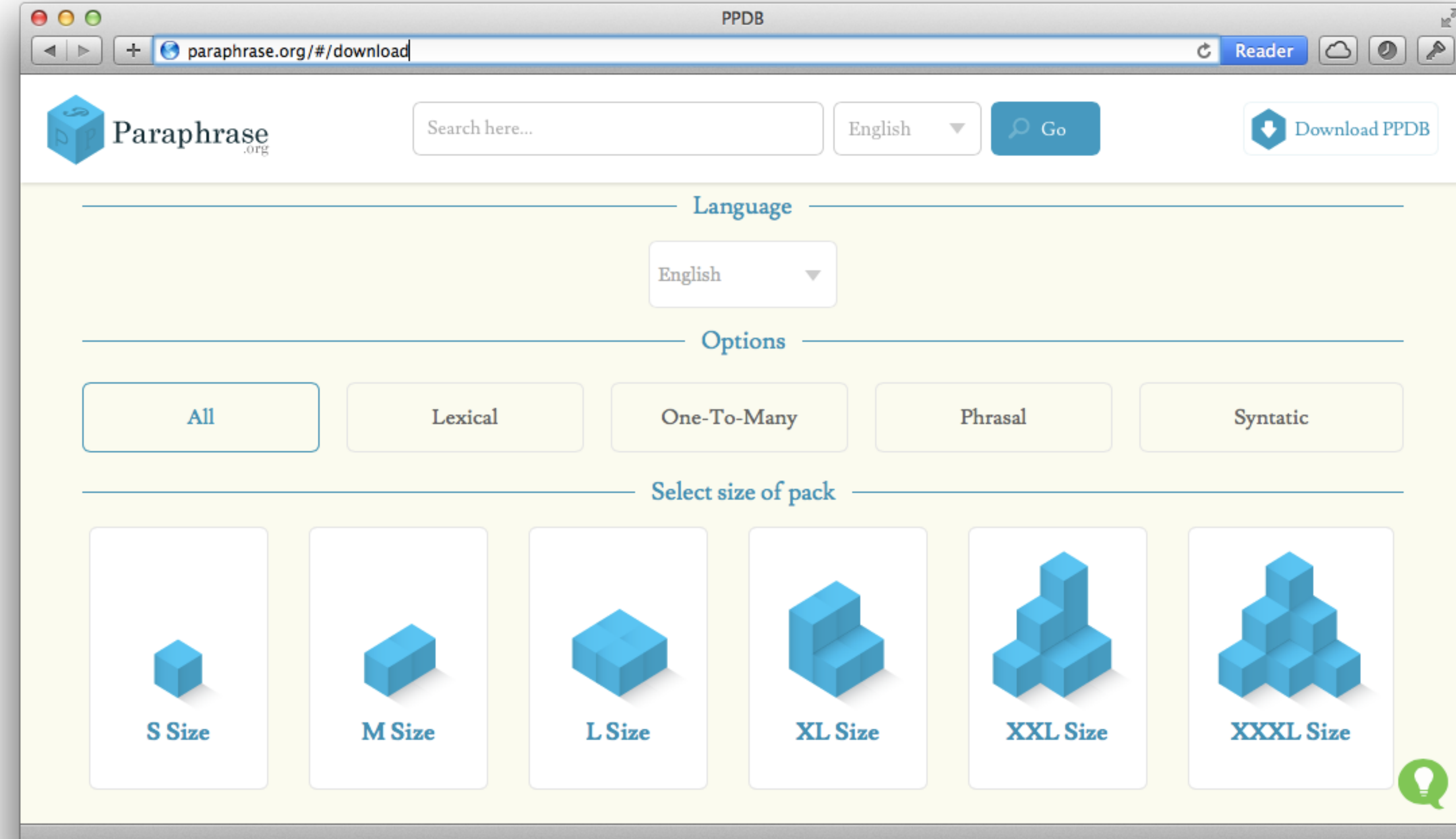
PPDB: The Paraphrase Database

- 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

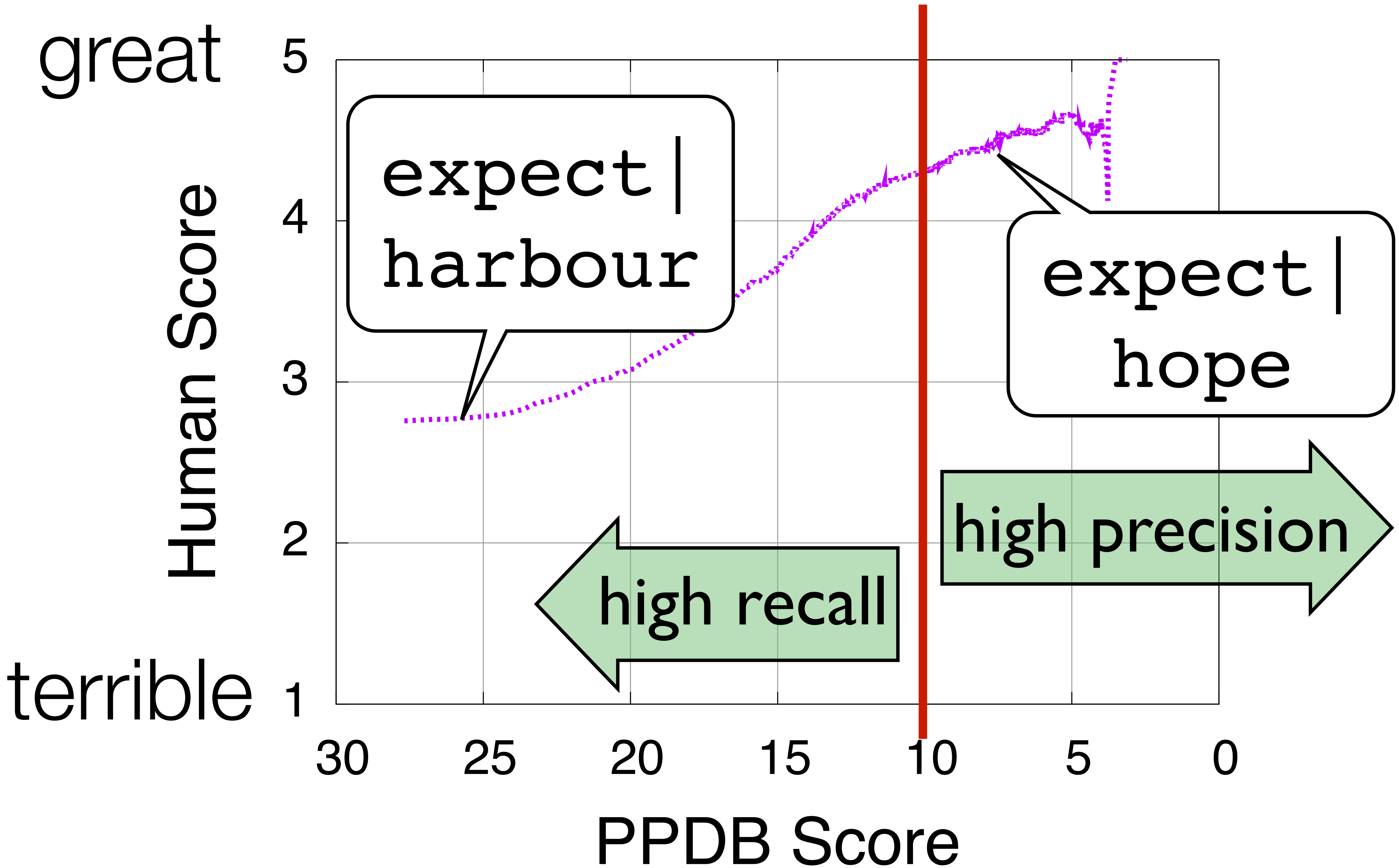
http://paraphrase.org


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



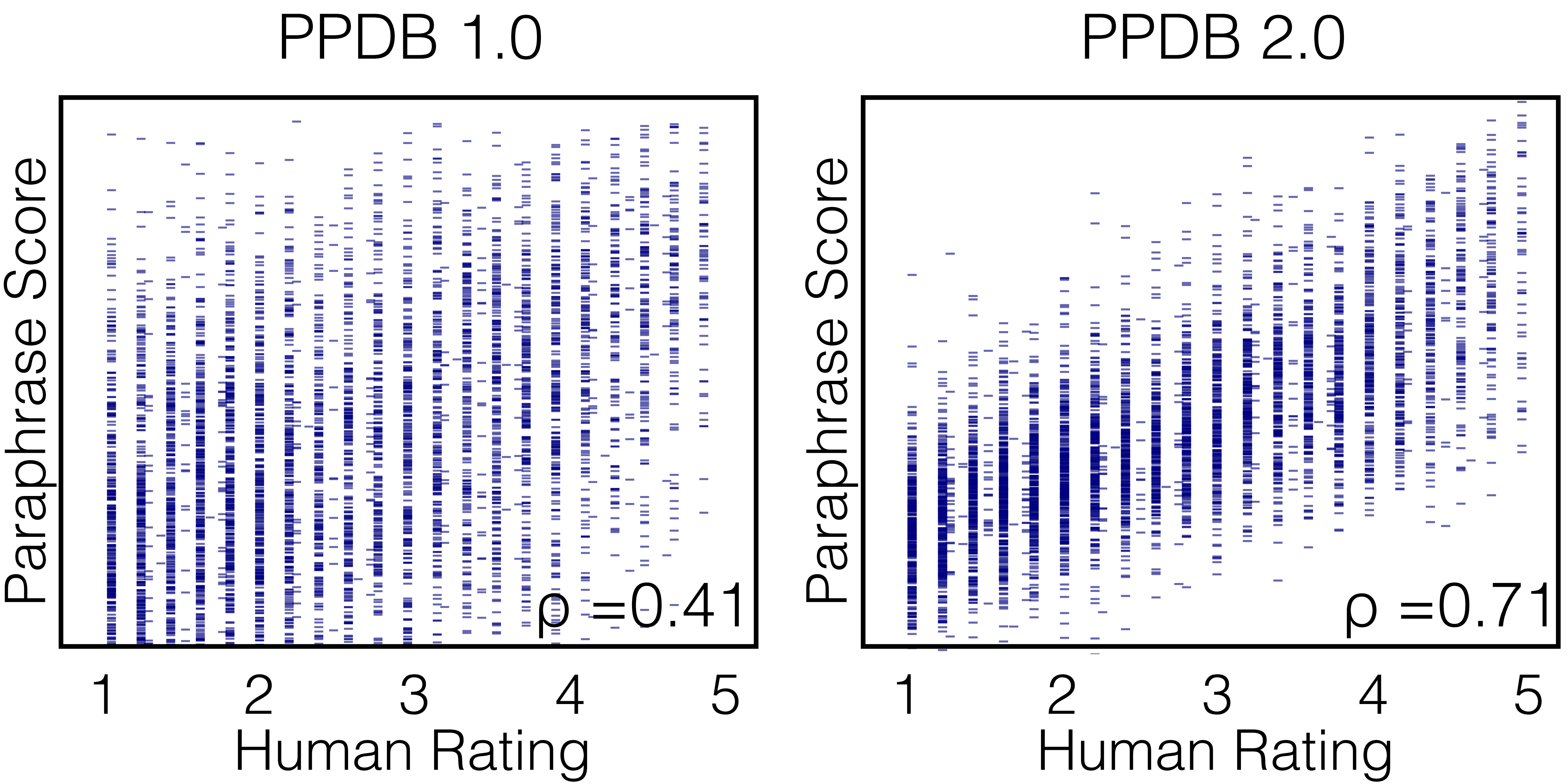
PPDB
paraphrase.org/#/download
Reader
Paraphrase
.org
Search here...
English
Go
Download PPDB
Language
English
Options
All
Lexical
One-To-Many
Phrasal
Syntatic
Select size of pack
S Size
M Size
L Size
XL Size
XXL Size
XXXL Size

Do the Scores Work?



PPDB 2.0

Re-ranked paraphrases better correlate with human judgments



PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, Ben Van Durme, Chris Callison-Burch. ACL-2015.

Fun PPDB Examples

munchies ||| hungry

[redacted]

sex [redacted]

PARENTAL

ADVISORY

EXPLICIT CONTENT

[redacted]

sheeit [redacted]

abso-fucking-lutely ||| indeed

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

Summary

Extraction & Representation

Extended large-scale paraphrase acquisition from bitexts to syntactic paraphrases

Generation

Introduced a straightforward and effective adaptation framework

Extensions beyond SMT

Improved performance by using monolingual information

Extensions

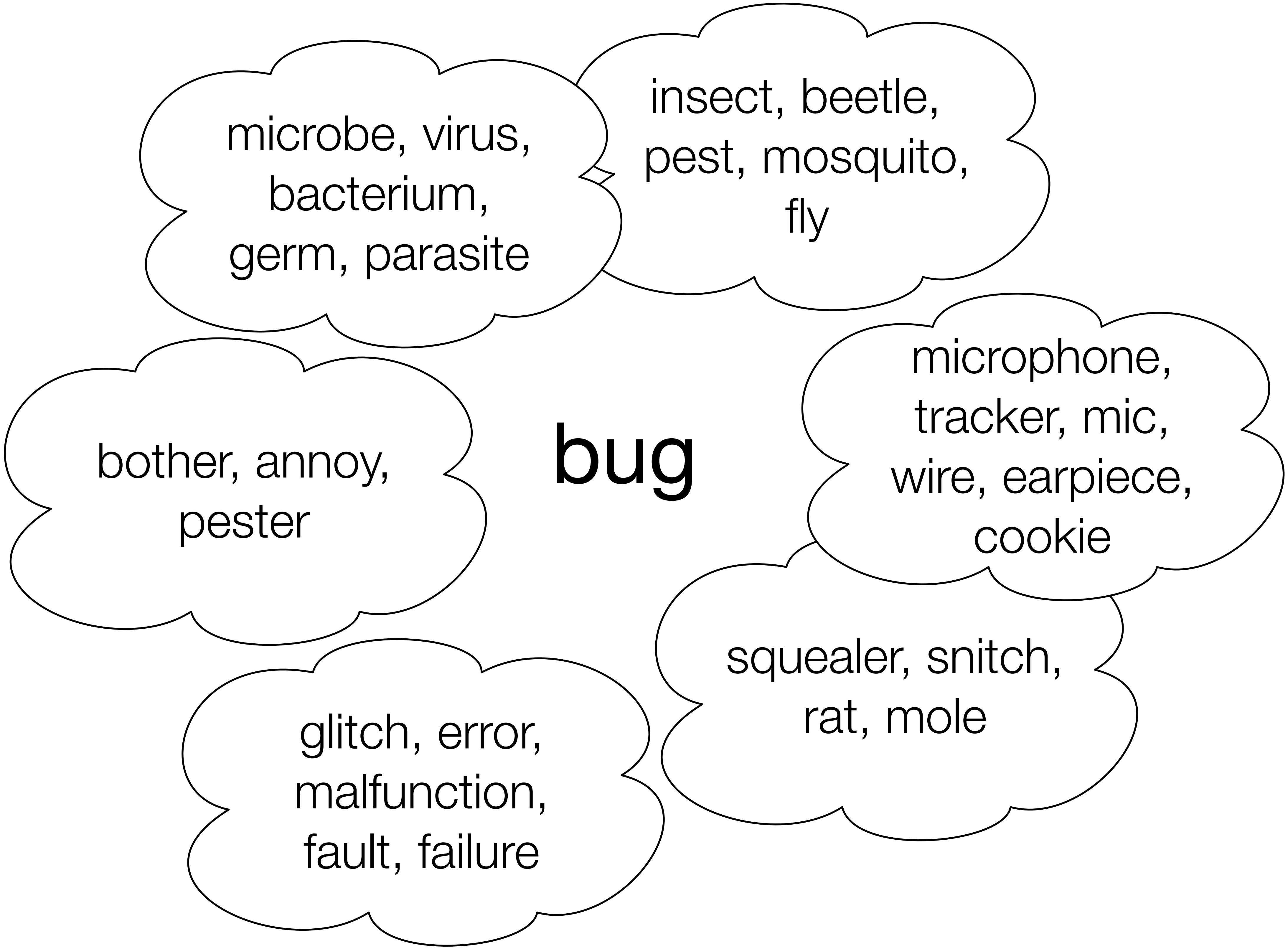
Polysemy of paraphrases

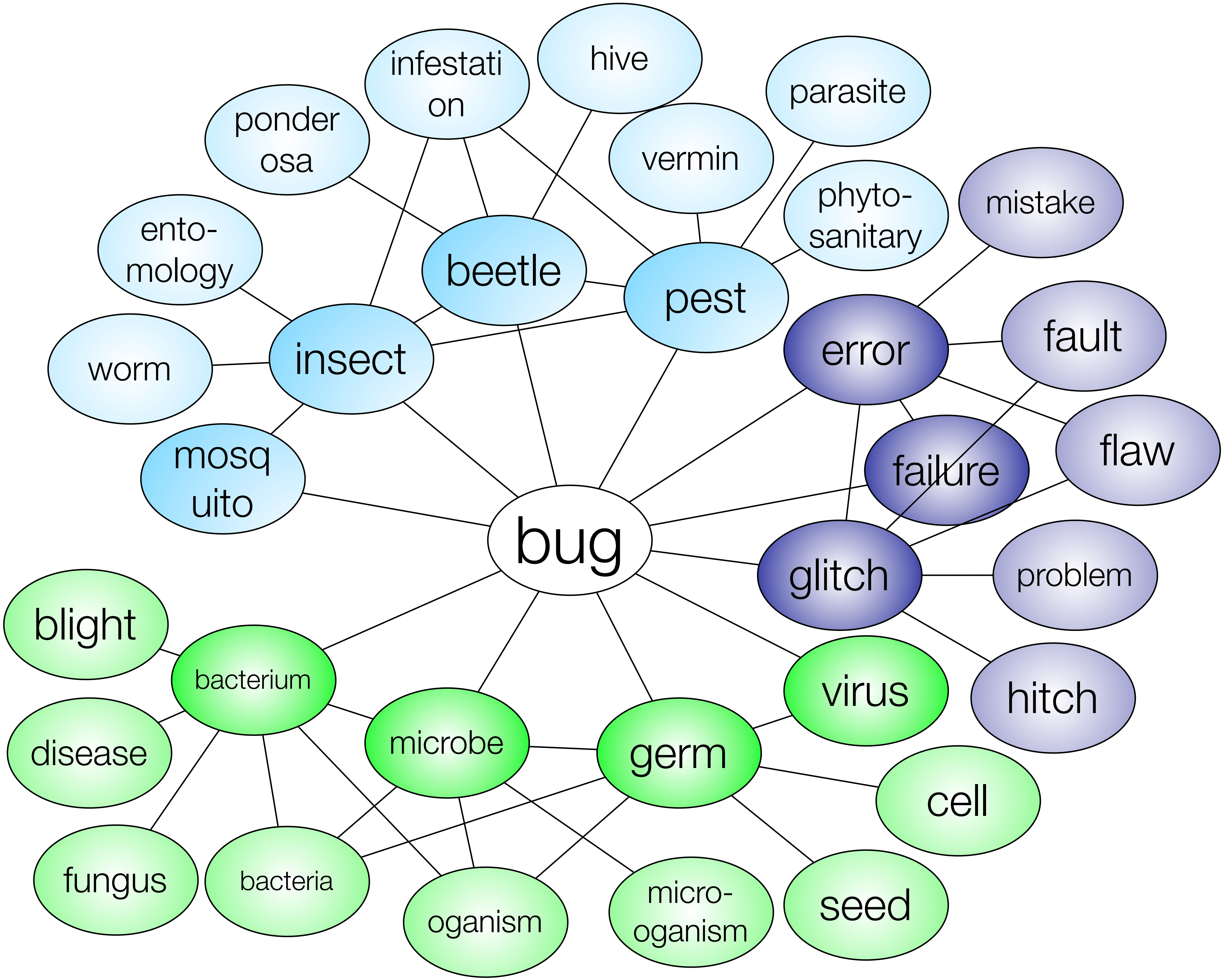
Our method sometimes groups paraphrases that correspond to different senses of the input phrase. How can we partition them into sets?

Paraphrase recognition and entailment

The RTE problem diverges in interesting ways from paraphrasing. We are combining natural language inference and data-driven paraphrasing.

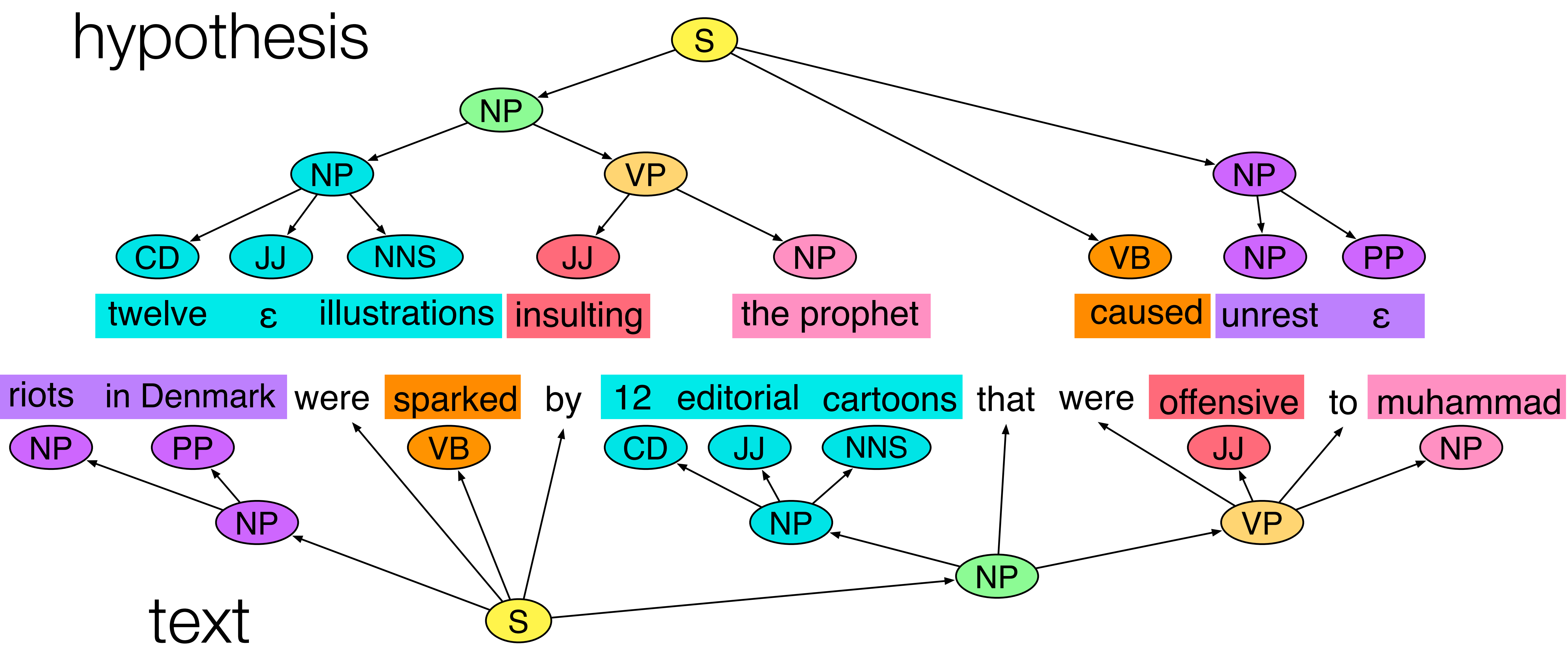
Word Sense


microbe, virus, bacterium, germ, parasite
insect, beetle, pest, mosquito, fly
bother, annoy, pester
bug
microphone, tracker, mic, wire, earpiece, cookie
squealer, snitch, rat, mole
glitch, error, malfunction, fault, failure



Clustering Paraphrases by Word Sense. Anne Cocos and Chris Callison-Burch. NAACL 2016.

Textual Inference



Semantic Relationships

twelve	12	equivalence
cartoons	illustrations	forward entailment
ε	in Denmark	reverse entailment
caused	prevented	negation
Europe	the middle East	alternation

Attaching a Semantics

twelve	12	equivalence
cartoons	illustrations	forward
caused	prevented	negation
Europe	the middle East	alternation

Riots in Greece → Civil unrest in Europe

~~**Civil unrest in Europe → Riots in Greece**~~

thank you for your time

many thanks

anyway , thanks

here you go

leave a message

gee , thanks

thanks , man

you look amazing

bless you

diet coke

Thank you!

thank you very much

thank you for your attention

keep the change

uh , thanks

why , thank you

don't thank me

hey , thanks

thank you , frank

Bibliography

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

Improved Statistical Machine Translation Using Paraphrases. Chris Callison-Burch, Philipp Koehn and Miles Osborne, 2006. In Proceedings NAACL 2006.

Paraphrase Substitution for Recognizing Textual Entailment. Wauter Bosma and Chris Callison-Burch. Lecture Notes in Computer Science, 2007.

Paraphrasing and Translation. Chris Callison-Burch, 2007. PhD Thesis, University of Edinburgh.

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

Constructing Corpora for the Development and Evaluation of Paraphrase Systems. Trevor Cohn, Chris Callison-Burch, Mirella Lapata, 2008. Computational Linguistics: Volume 34, Number 4.

ParaMetric: An Automatic Evaluation Metric for Paraphrasing. Chris Callison-Burch, Trevor Cohn and Mirella Lapata. COLING 2008

Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity. Charley Chan, Chris Callison-Burch, and Benjamin Van Durme. GEMS 2011.

Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation. Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. EMNLP 2011.

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

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

The Multilingual Paraphrase Database. Juri Ganitkevitch and Chris Callison-Burch. LREC 2014.

PARADIGM: Paraphrase Diagnostics through Grammar Matching. Jonny Weese, Juri Ganitkevitch, and Chris Callison-Burch. EACL 2014

Bibliography

Adding Semantics to Data-Driven Paraphrasing. Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. ACL-2015.

PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, Ben Van Durme, Chris Callison-Burch. ACL-2015.

Domain-Specific Paraphrase Extraction. Ellie Pavlick, Juri Ganitkevich, Charley Chan, Xuchen Yao, Ben Van Durme, Chris Callison-Burch. ACL-2015.

FrameNet+: Fast Paraphrastic Tripling of FrameNet. Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Drezde, Ben Van Durme. ACL-2015.

SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter. Wei Xu, Chris Callison-Burch, and Bill Dolan. SemEval-2015.

Problems in Current Text Simplification Research: New Data Can Help. Wei Xu, Chris Callison-Burch, and Courtney Napoles. TACL-2015.

Optimizing Statistical Machine Translation for Text Simplification. Wei Xu, Courtney Napoles, Ellie Pavlick, Jim Chen, and Chris Callison-Burch. TACL-2016.

Sentential Paraphrasing as Black-Box Machine Translation. Courtney Napoles, Chris Callison-Burch, and Matt Post. NAACL-2016.

Simple PPDB: A Paraphrase Database for Simplification. Ellie Pavlick and Chris Callison-Burch. ACL-2016.

Clustering Paraphrases by Word Sense. Anne Cocos and Chris Callison-Burch. NAACL-2016.

Entailment relations

Hypernym	Synonym	Antonyms	Alternations	Independent
beetle | insect	icebox | refrigerator	advantage | disadvantage	cheese | butter	advocacy | spokesman
honeybee | bee	impasse | deadlock	competence | incompetence	cliff | cave	aircraft | sky
fees | spending	infirmary | hospital	continuity | discontinuity	clothing | equipment	actor | arena
know-how | knowledge	insurrection | revolt	inflow | outflow	clothing | housing	actor | maker
pond | lake	jewel | gem	insanity | sanity	coating | asphalt	actor | movie
fertilizer | manure	john | lavatory	legitimacy | illegitimacy	columnist | newspaperman	actor | singer
actor | entertainer	kale | cabbage	niece | nephew	commentator | reporter	actor | spokesman
actor | performer	labyrinth | maze	descendants | ancestors	competence | productivity	advantage | equipment
acquisition | buying	laundry | washing	husbands | wives	compliance | enforcement	ambassador | delegation