Large-Scale Paraphrase Extraction & Application

Juri Ganitkevitch

Jonathan Weese, Courtney Napoles, Benjamin Van Durme,
Matt Post, and Chris Callison-Burch

#tbt

Paraphrase Extraction

Text-to-Text Applications

Scaling to PPDB

Contributions

Paraphrases

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

cup	↔	mug
the king's speech	↔	His Majesty's address
X_1 talks to X_2	↔	X_1 converses with X_2
one JJ instance of NP	↔	a JJ case of NP

..in NLP

Recognition or generation of **meaning**-constrained relationships in **text**, i.e. automatic...

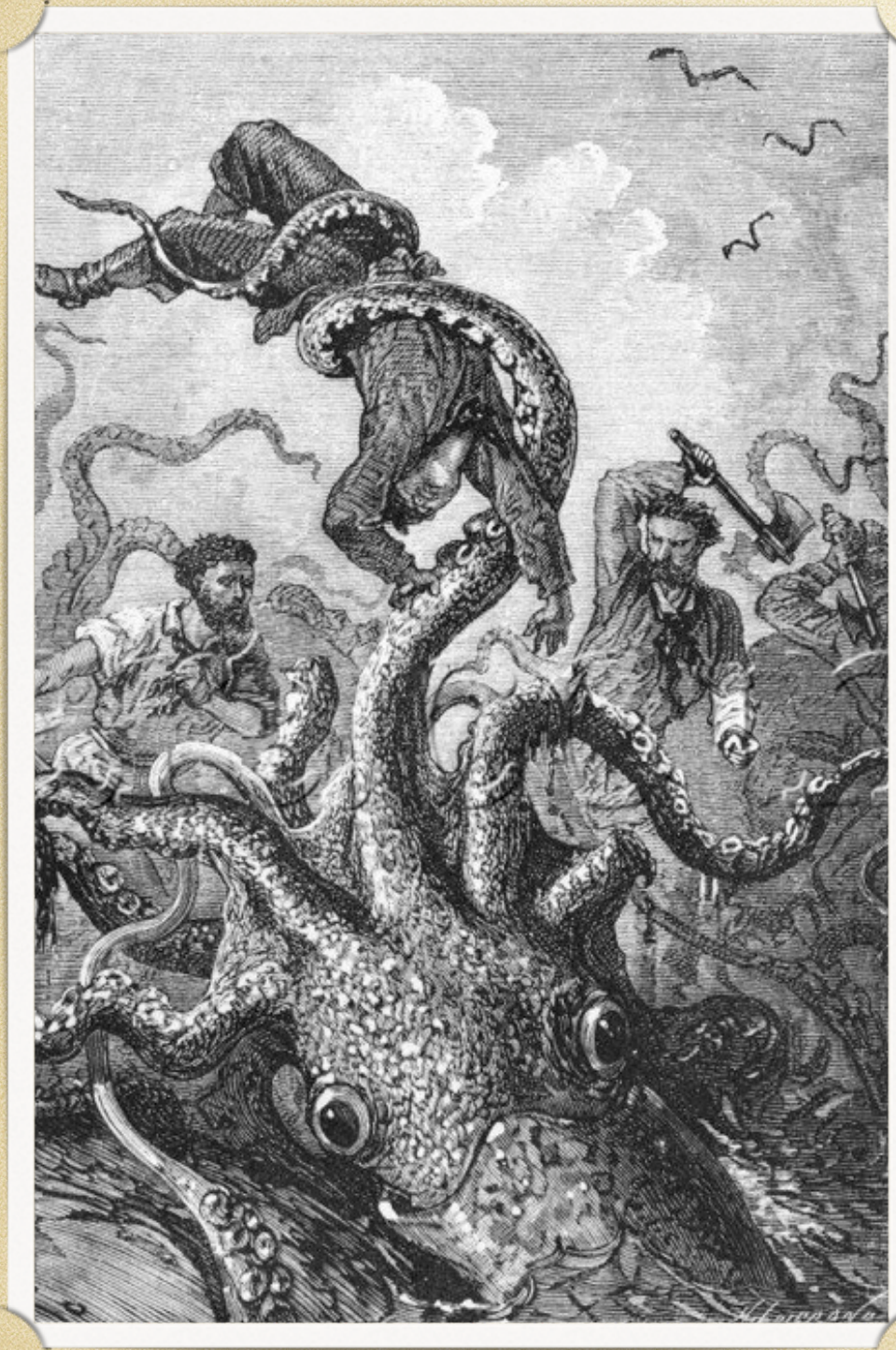
...information extraction, question answering, entailment recognition, summarization, translation, compression, simplification, natural language generation, etc.

Where do paraphrases come from?

Data.

Types of Corpora

Parallel:	English – English
Comparable:	English ~ English
Plain:	English
Bilingual parallel:	English – French



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**.

Types of Corpora

Parallel:	English – English
Comparable:	English ~ English
Plain:	English
Bilingual parallel:	English – French

A staggering 5 million Americans **have been victims of identity theft** in the last five years, according to a federal trade commission survey out this week.

In the last year alone, 1 million people **have had their identity purloined**.

Types of Corpora

Parallel:	English – English
Comparable:	English ~ English
Plain:	English
Bilingual parallel:	English – French

duty | responsibility

Modified by:

additional, administrative, assigned, assumed, collective, congressional, constitutional

Object of:

accept, articulate, assert, assign, assume, attend to, avoid, become, breach

Types of Corpora

Parallel:	English – English
Comparable:	English ~ English
Plain:	English
Bilingual parallel:	English – French

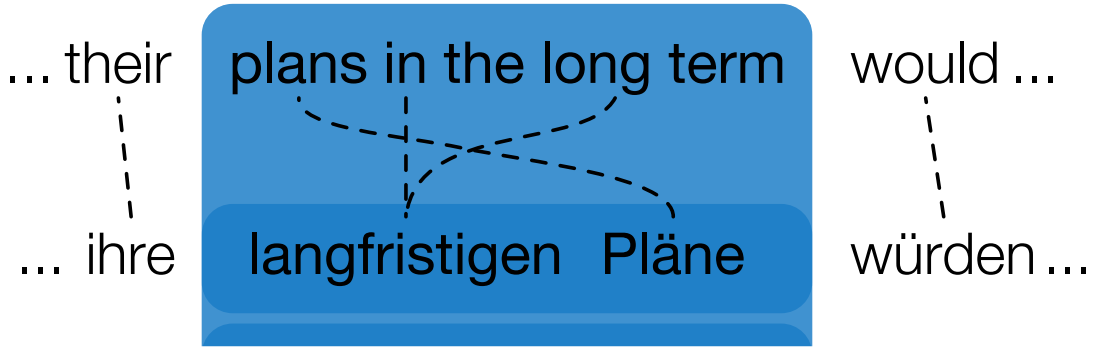
Bilingual Parallel Corpora

Sentence-aligned corpora in English and any foreign language

Strong meaning equivalence signal

Available in large quantities

Pivoting over a Foreign Language

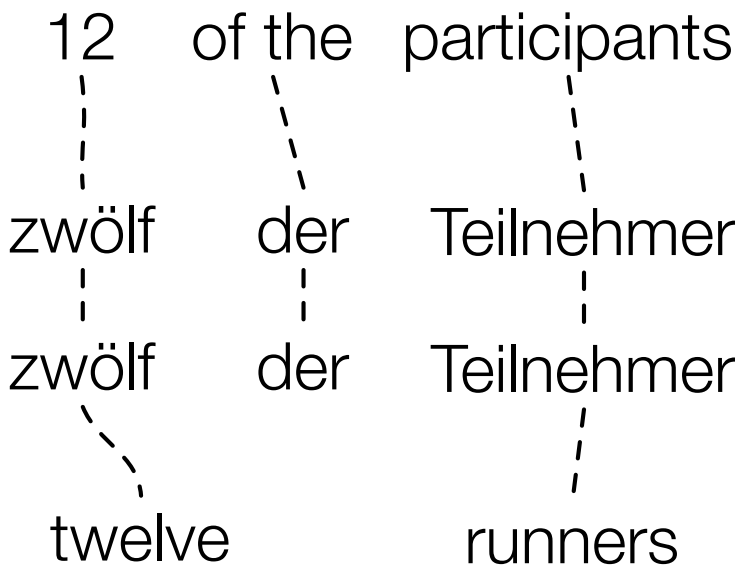


Goal: Expressiveness

Generalization vs. memorization

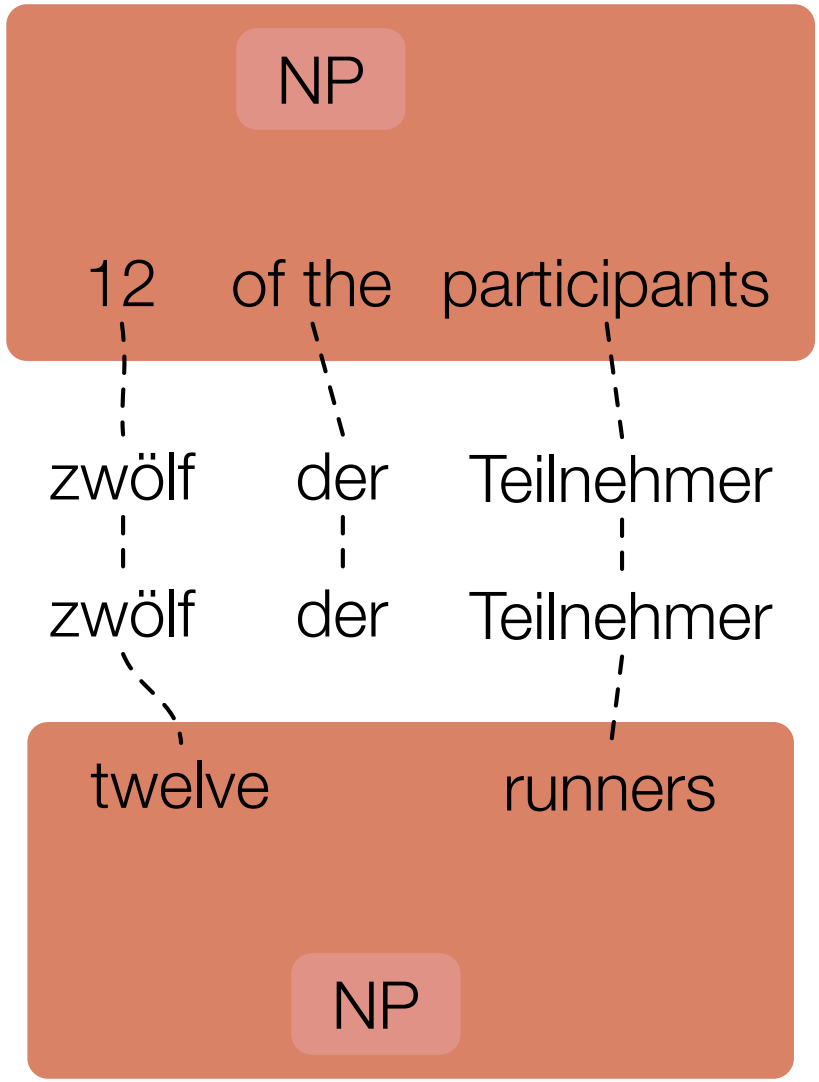
- Prefer general patterns over memorized phrases
- Use linguistic information

Syntactic Pivoting



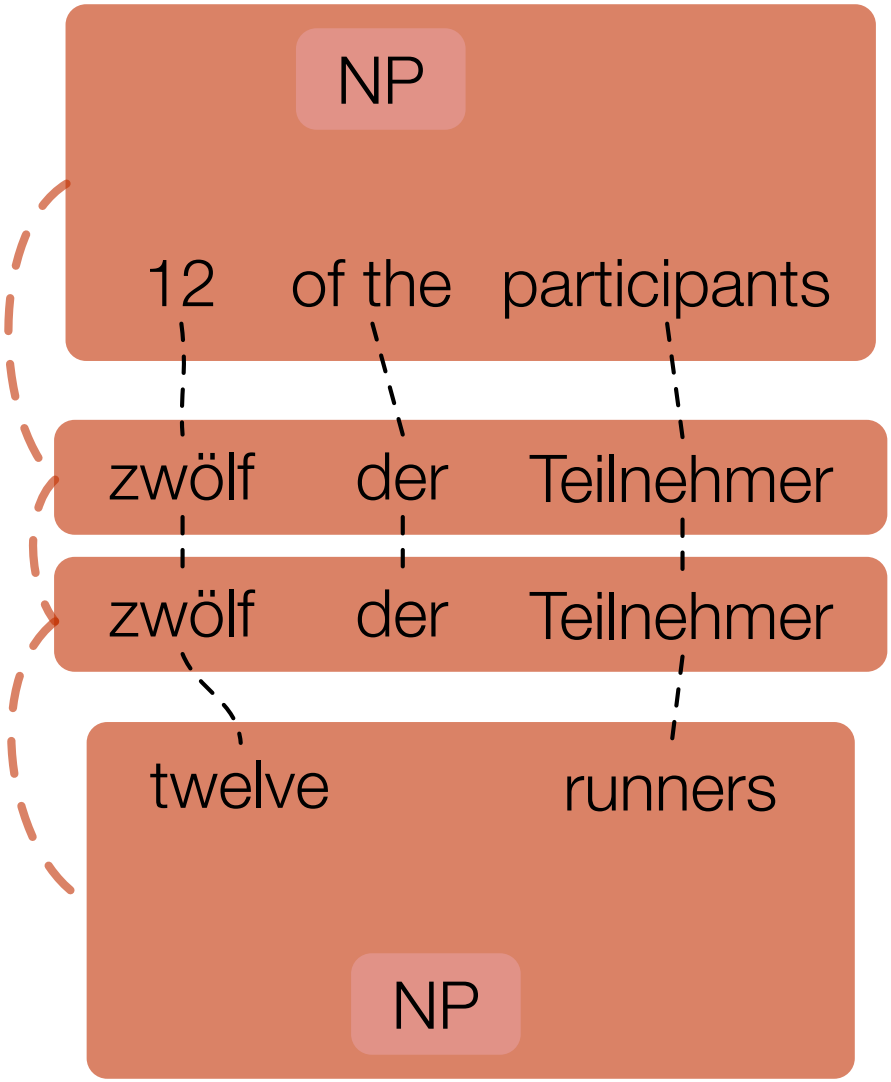
Ganitkevitch et al., 2011

Syntactic Pivoting



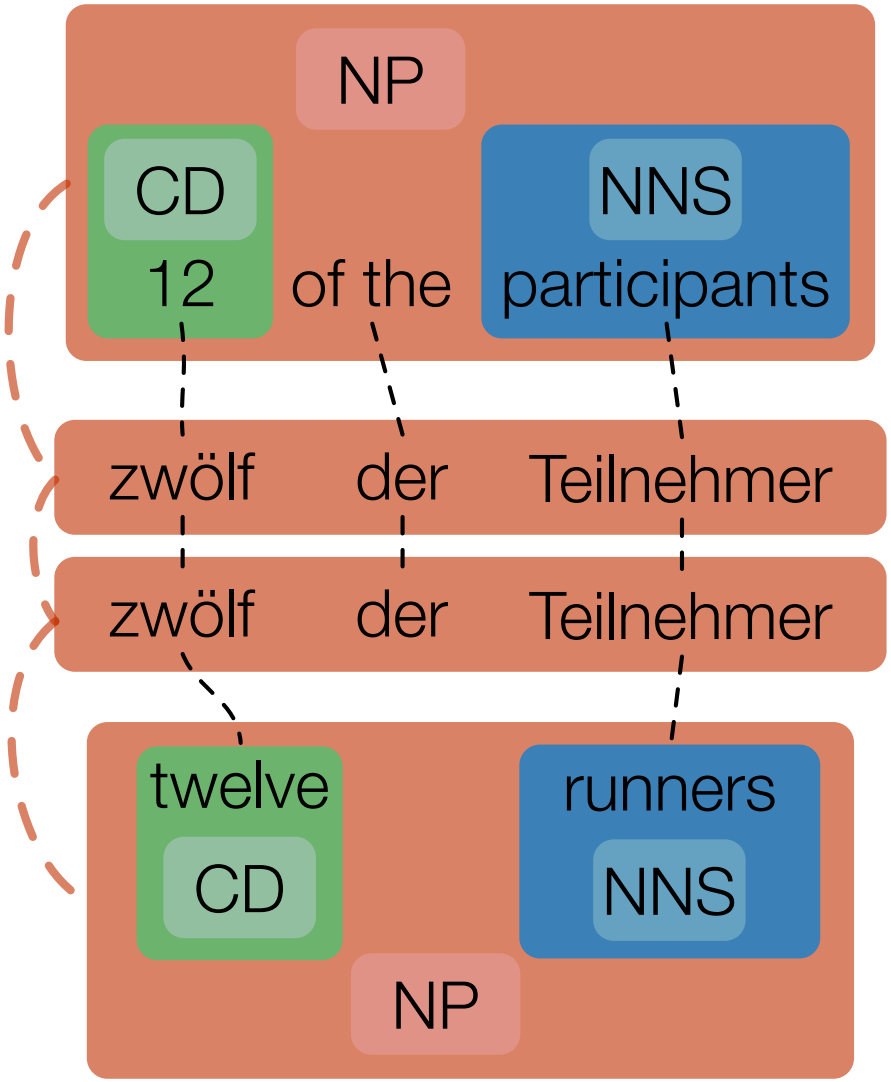
Ganitkevitch et al., 2011

Syntactic Pivoting



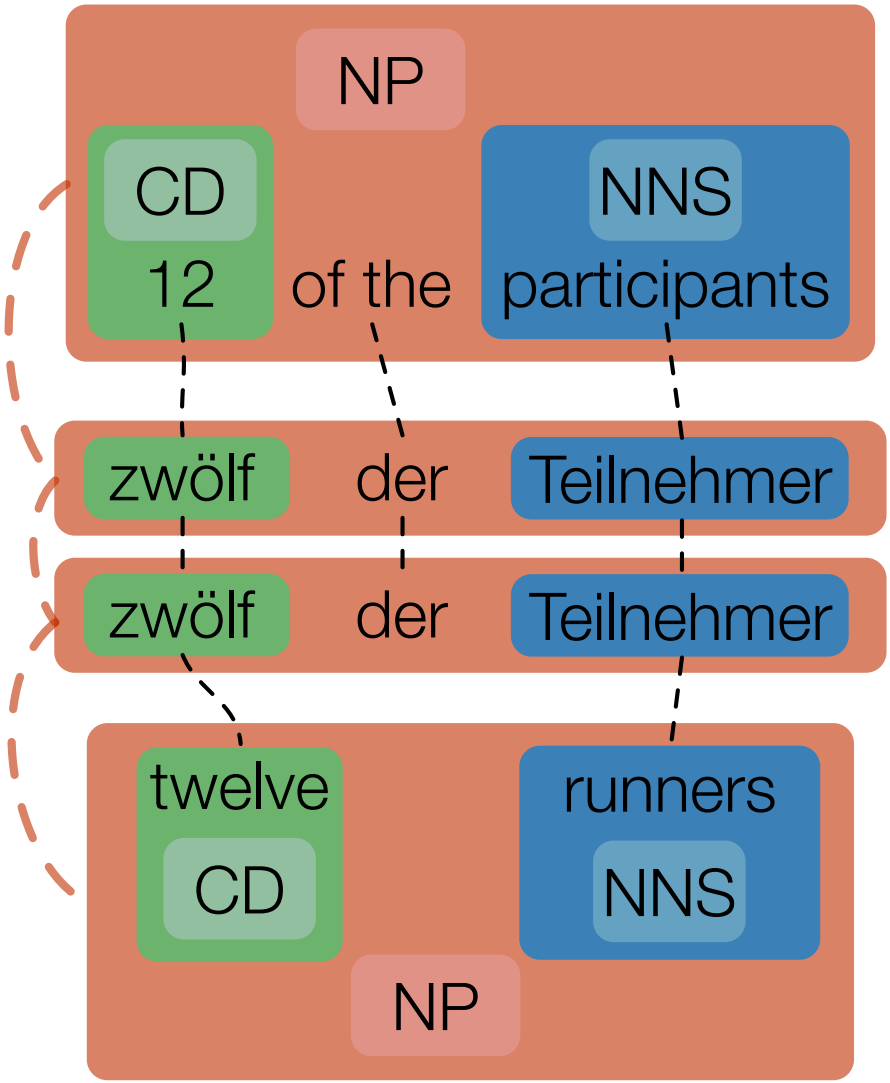
Ganitkevitch et al., 2011

Syntactic Pivoting



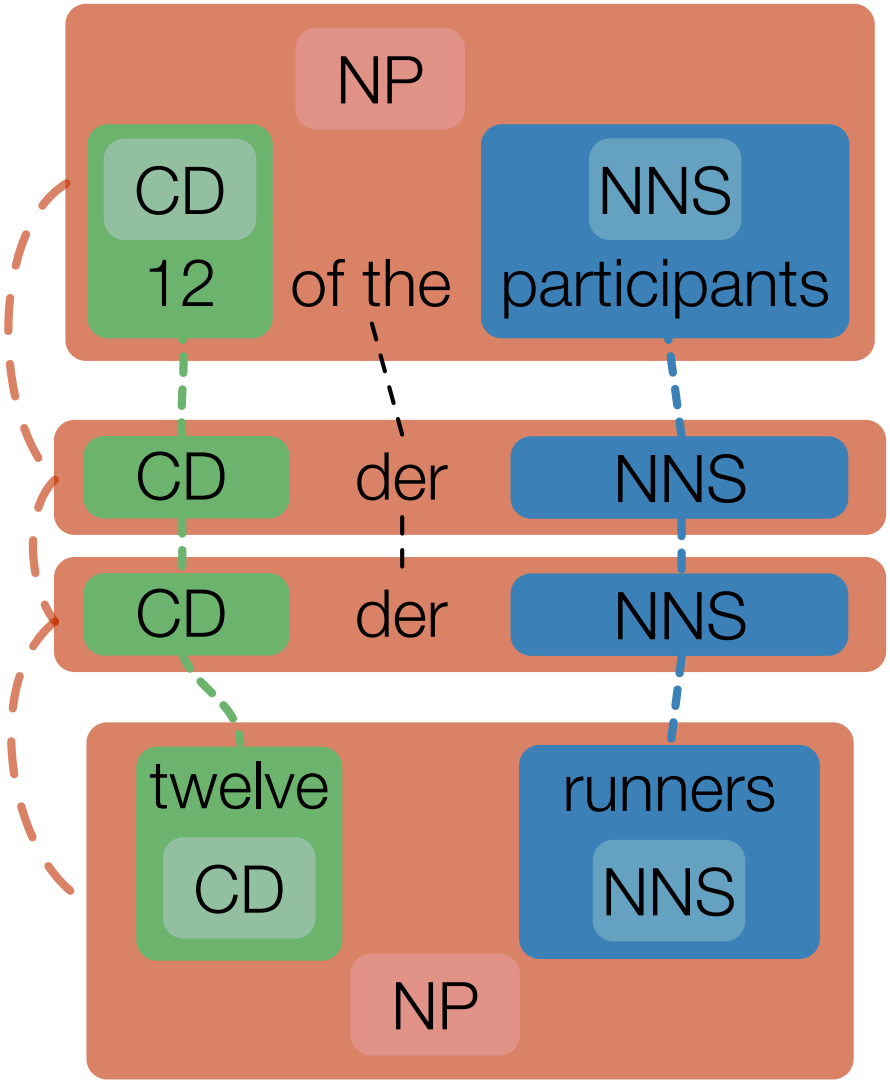
Ganitkevitch et al., 2011

Syntactic Pivoting



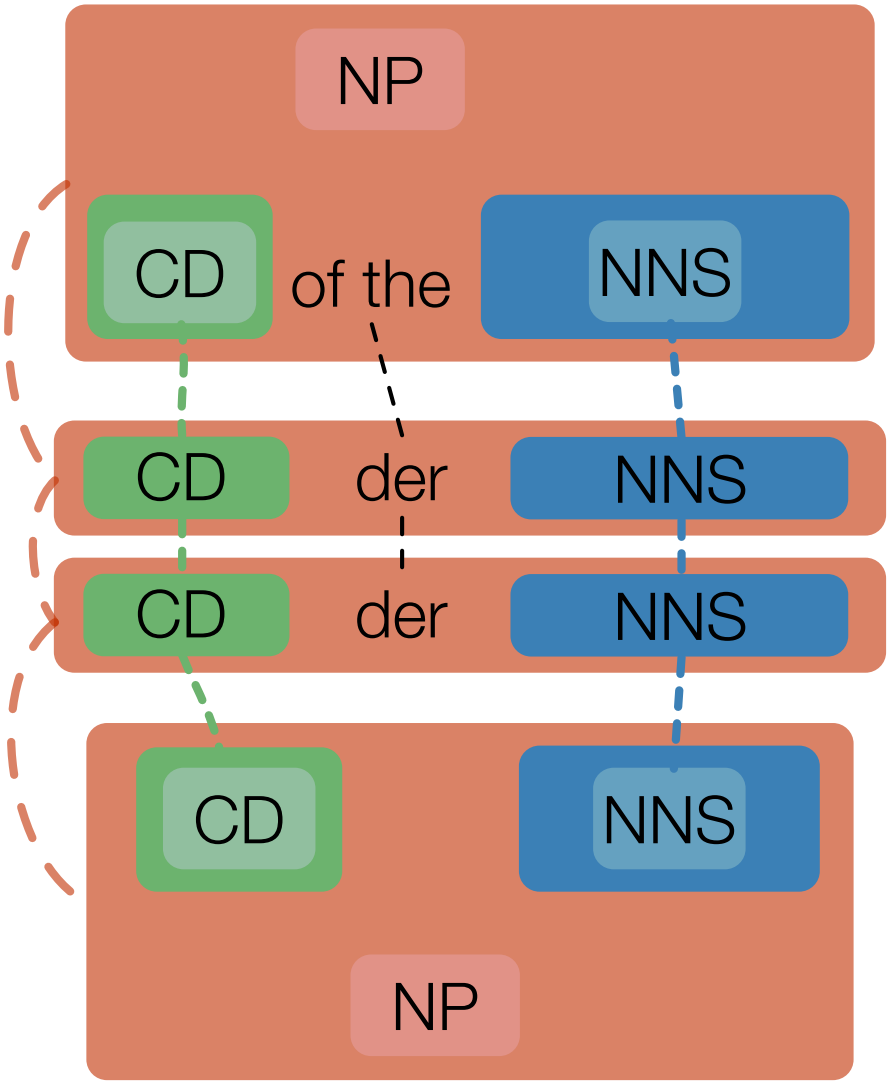
Ganitkevitch et al., 2011

Syntactic Pivoting



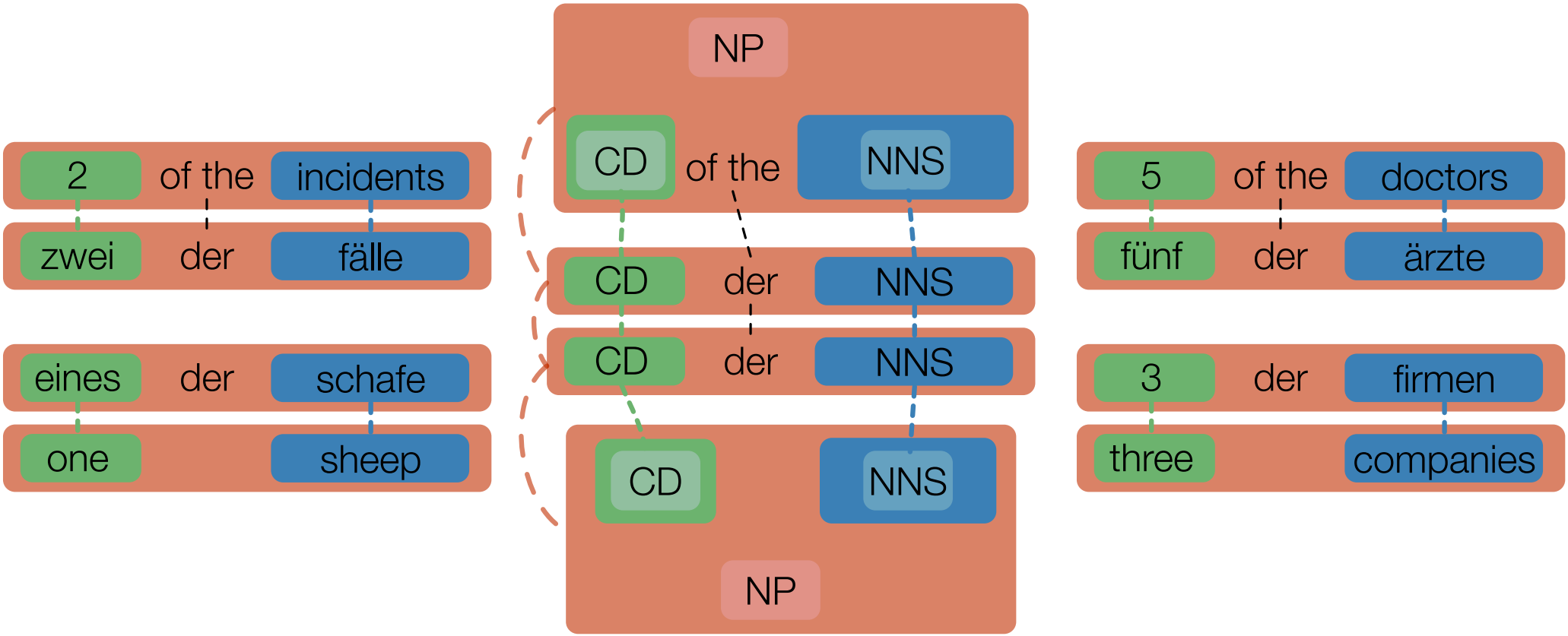
Ganitkevitch et al., 2011

Syntactic Pivoting



Ganitkevitch et al., 2011

Syntactic Pivoting



Ganitkevitch et al., 2011

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 VBP	|	they VBP 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 VBP that	|	although PRP VBP
	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 take a decision PP	|	to decide PP

Ganitkevitch et al., 2011

Paraphrase Quality

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	

Pivoting Probability

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

Paraphrase Extraction

Text-to-Text Applications

Scaling to PPDB

Contributions

Text-to-Text

Paraphrases form an SCFG

“Monolingual translation” via Joshua

Sentence compression, text simplification, poetry generation, lawyer-to-English, etc.

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

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

Adaptation Scheme

Tuning Data

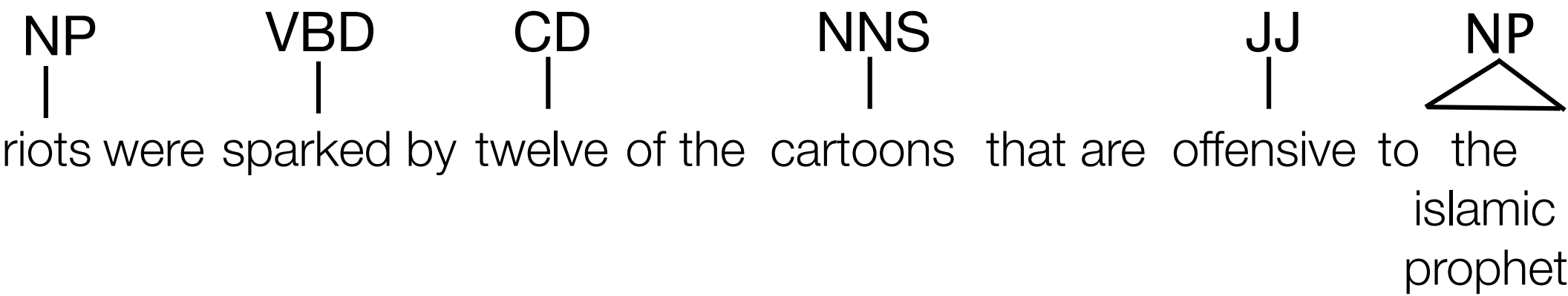
~2k long/short sentence pairs

Objective Function

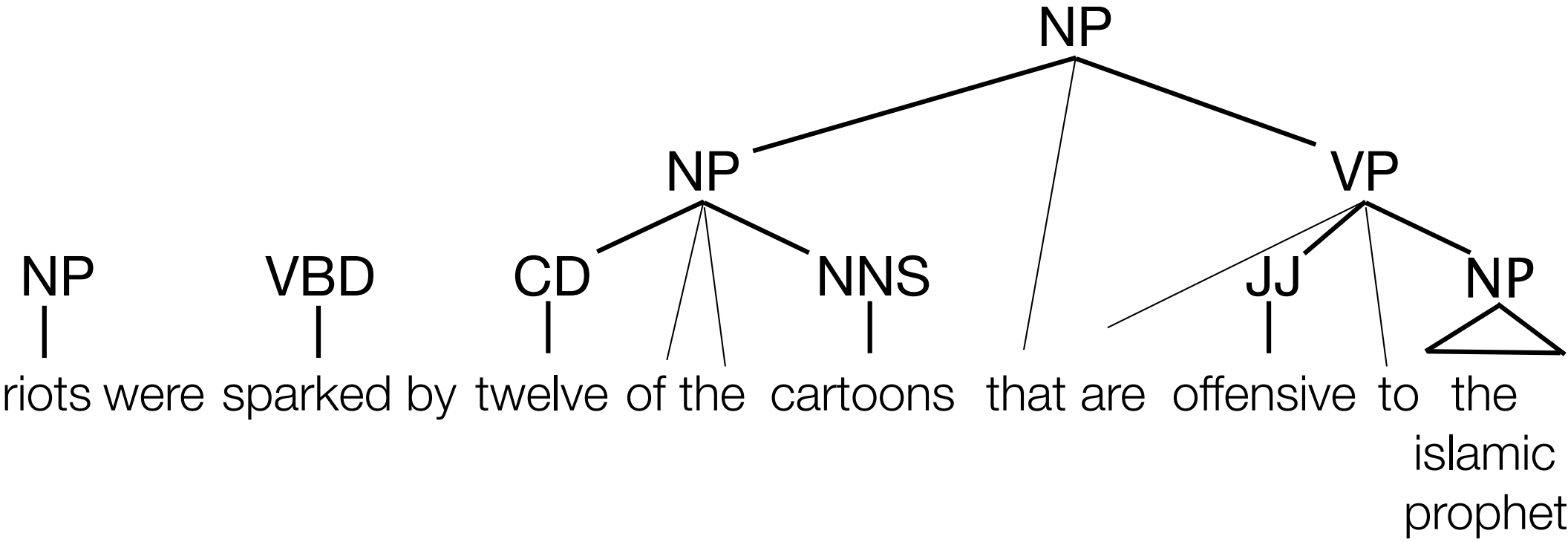
Compression-ratio targeted BLEU

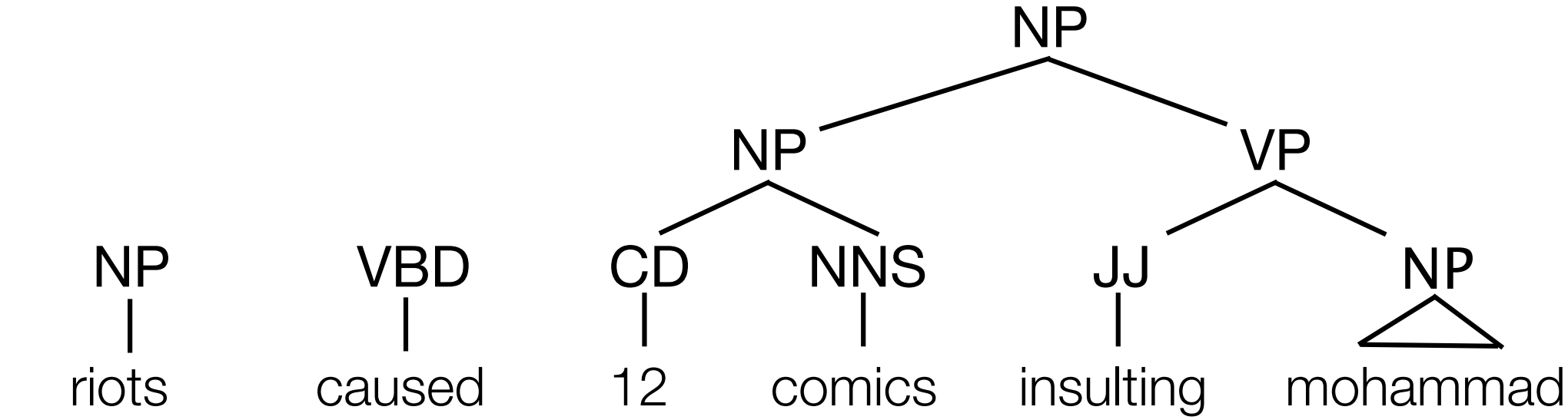
Paraphrase Features

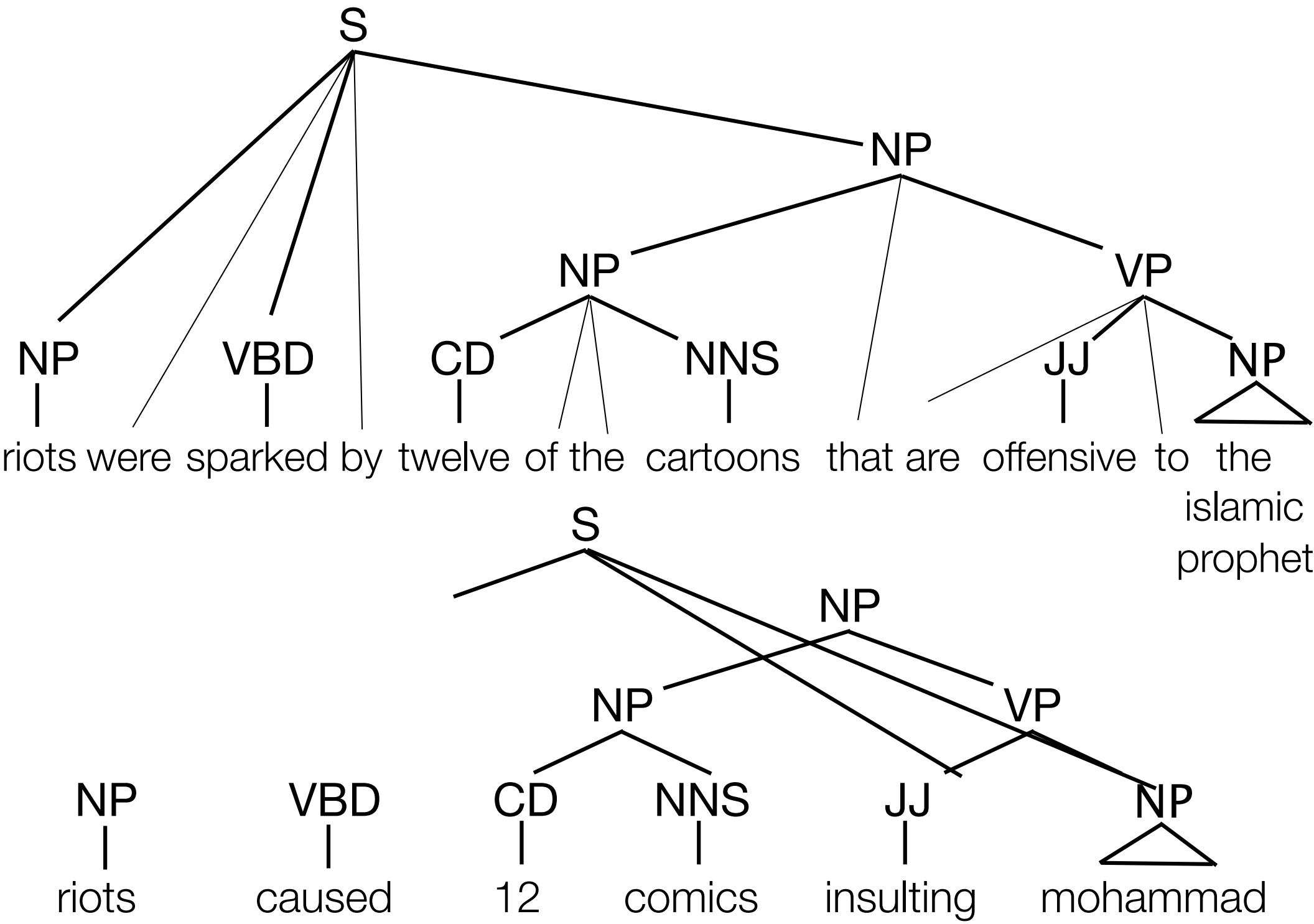
Compression-ratio of paraphrase etc.


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

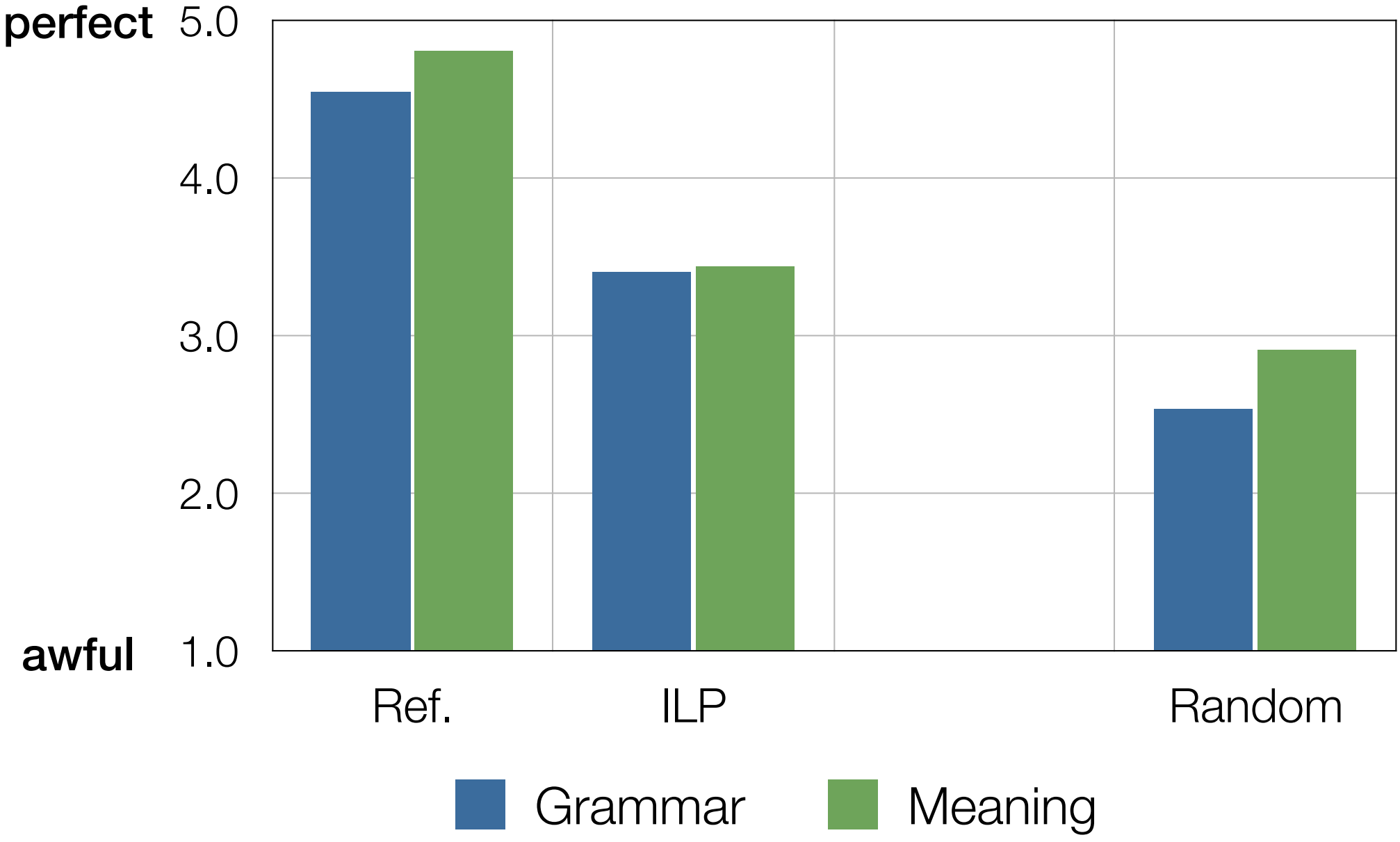

NP
VBD
CD
NNS
JJ
NP
riots
caused
12
comics
insulting
mohammad


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

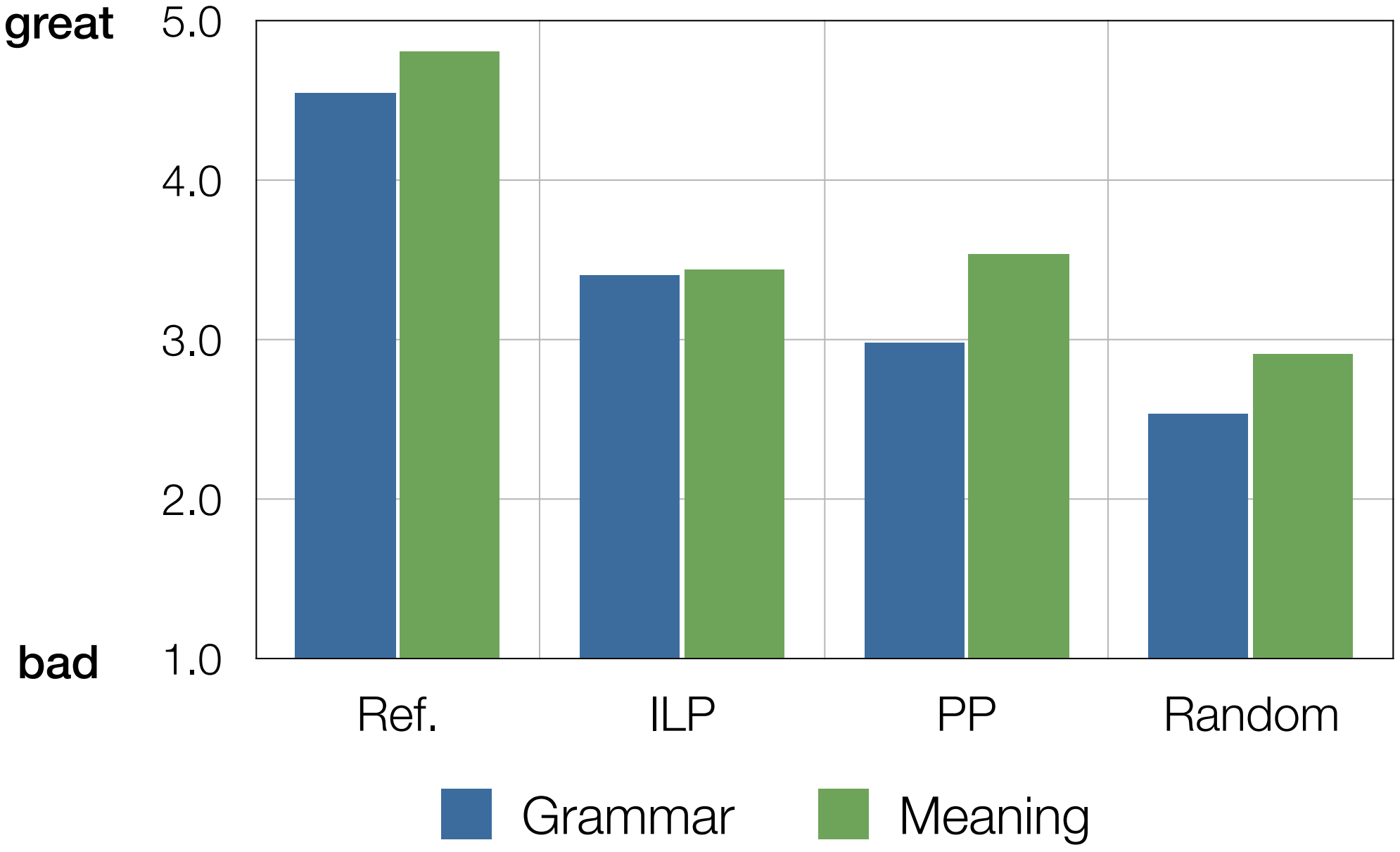

NP
NP
VP
NP
VBD
CD
NNS
JJ
NP
riots
caused
12
comics
insulting
mohammad


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

Compression Results



Compression Results



Paraphrase Extraction

Text-to-Text Applications

Scaling to PPDB

Contributions

Going Beyond Translation

Monolingual data & distributional similarity:

- Orthogonal signal
- Rich contextual features possible
- Vast amounts of data available

Ganitkevitch et al., 2012

duty | responsibility

Modified by:

additional, administrative, assigned, assumed, collective, congressional, constitutional

Object of:

accept, articulate, assert, assign, assume, attend to, avoid, become, breach

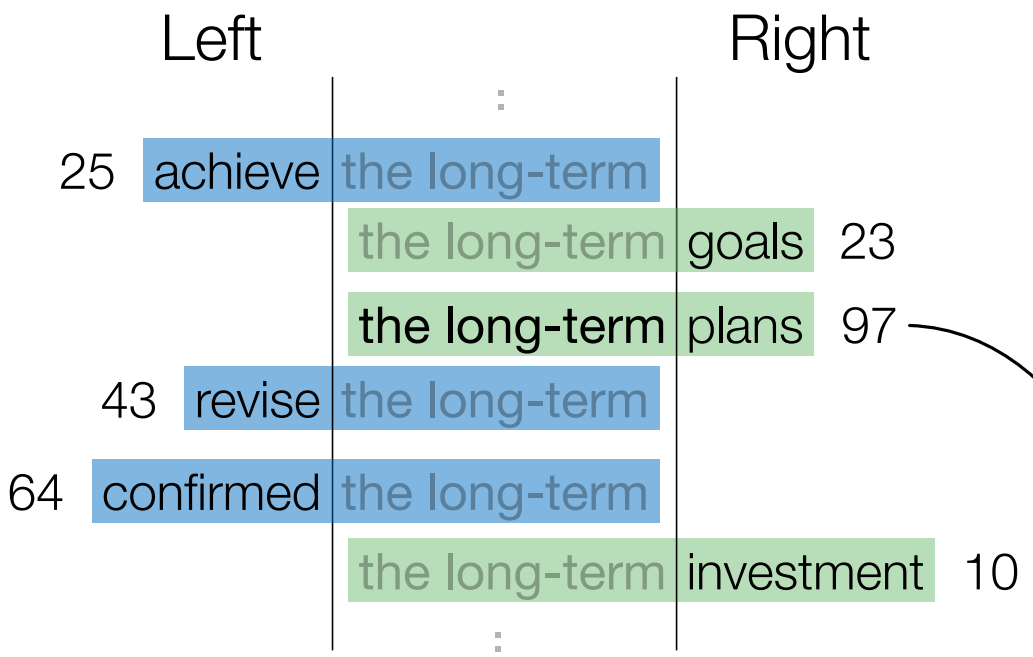
Monolingual Data

Google *n*-grams

Annotated Gigaword

Ganitkevitch et al., 2012

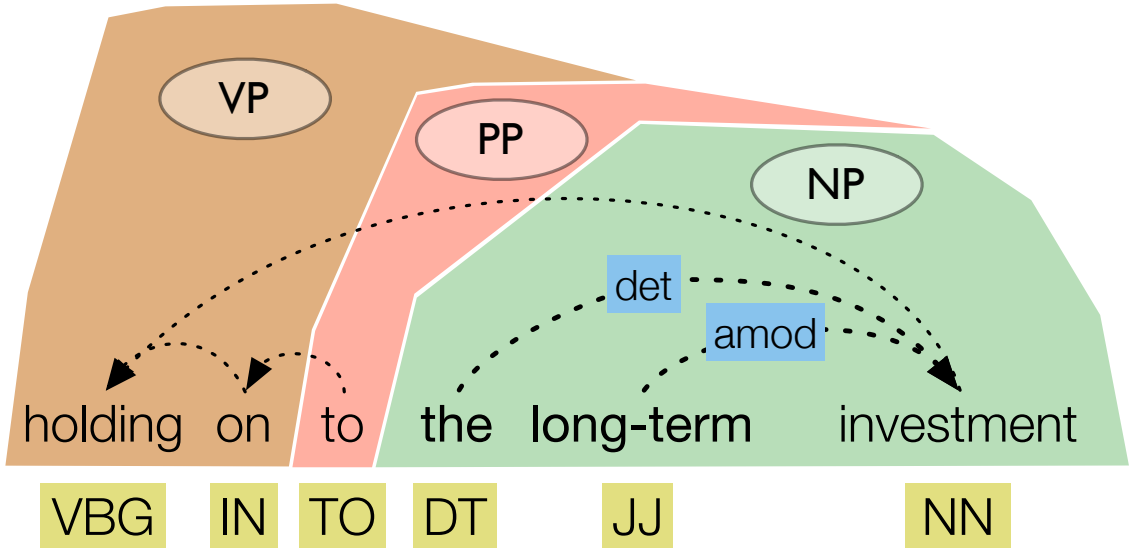
n-gram Context



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

Ganitkevitch et al., 2012

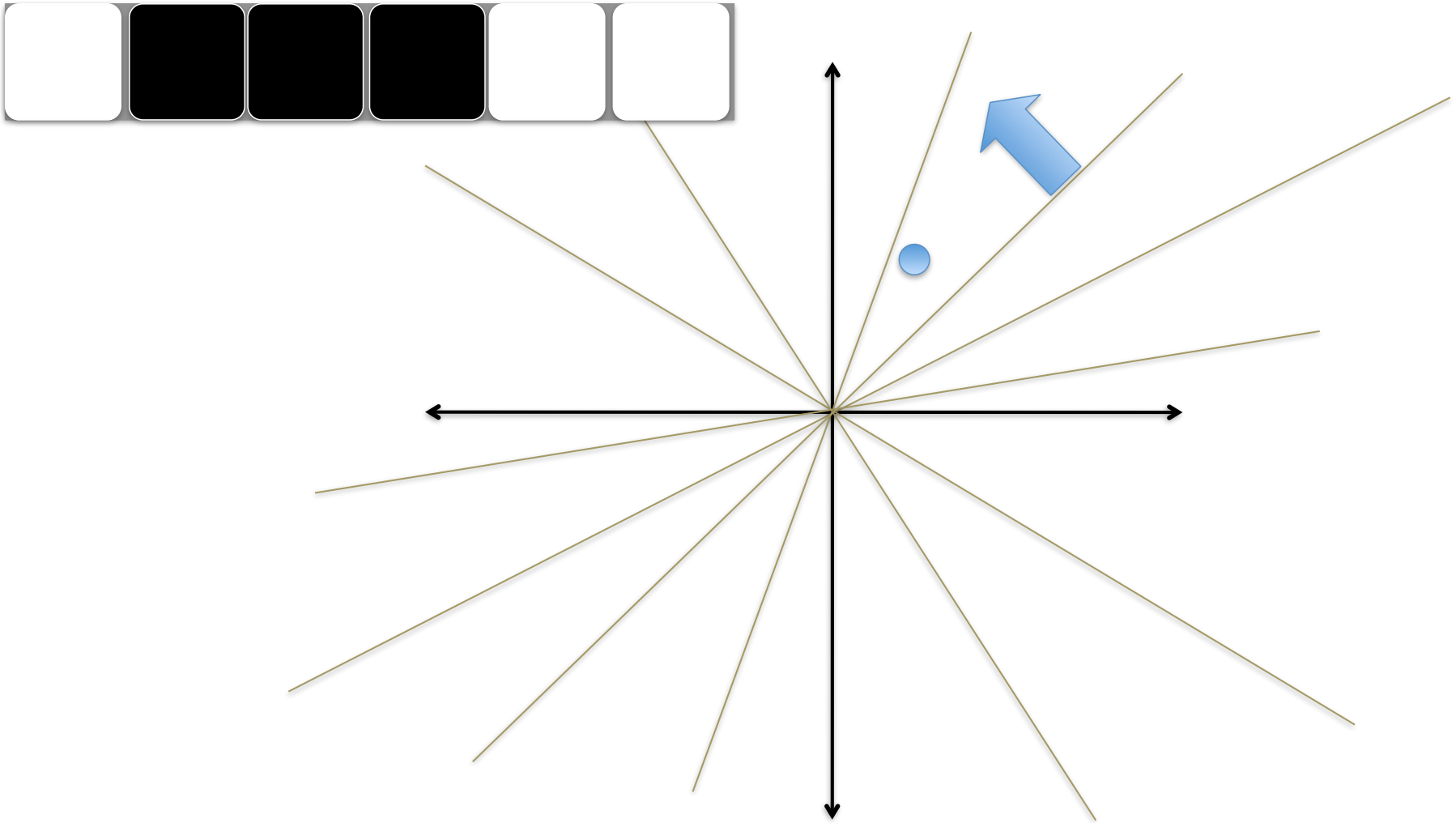
Annotated Context



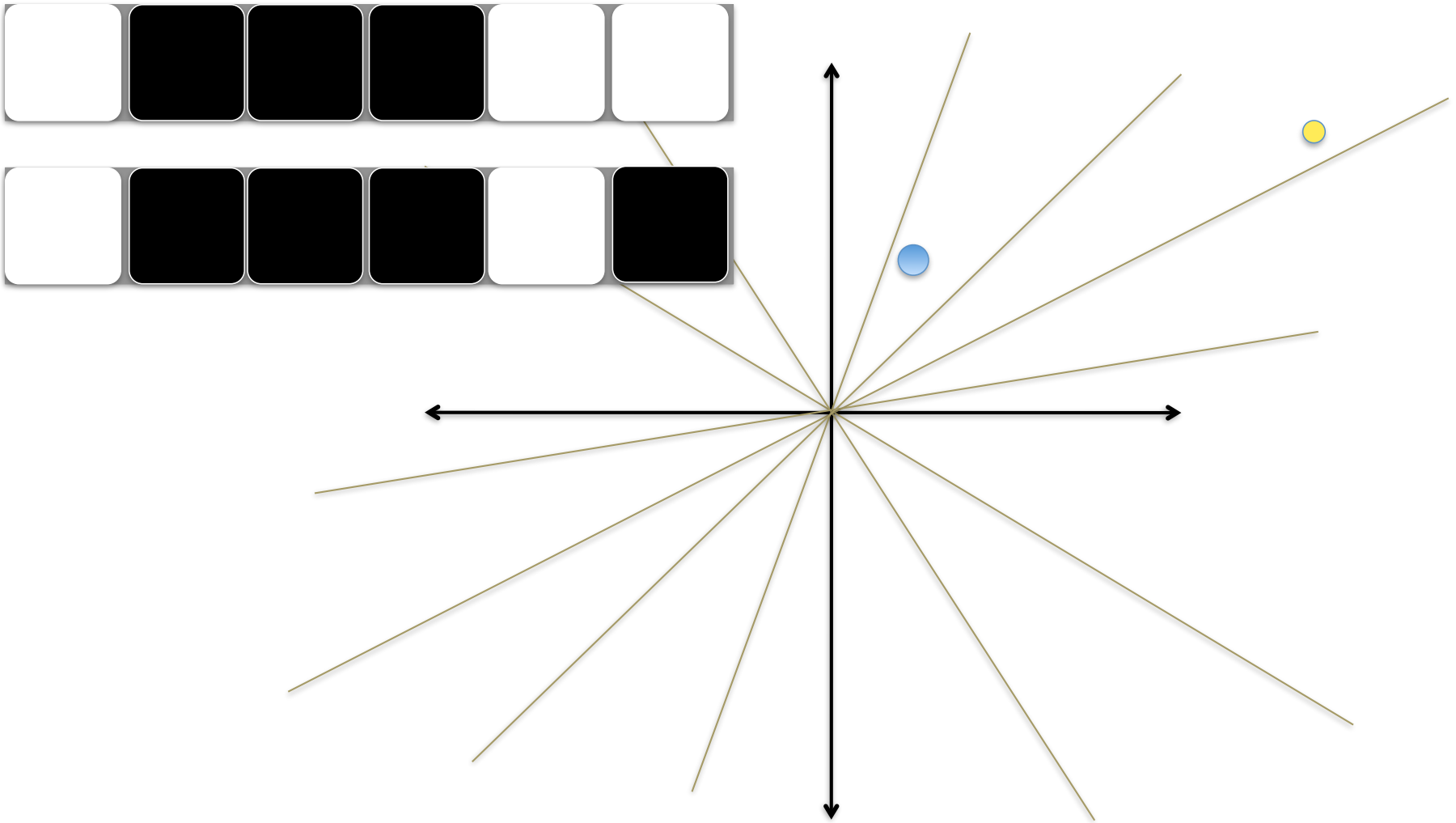
$$\vec{sig}_{syntax}\left(\text{the long-term}\right) = \left(\begin{array}{l} \text{lex-R-investment} \quad \text{lex-L-on-to} \\ \text{pos-L-IN-TO} \quad \text{pos-L-TO} \quad \text{lex-L-to} \\ \text{dep-det-R-investment} \quad \text{pos-R-NN} \\ \text{dep-amod-R-investment} \\ \text{dep-det-R-NN}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

Ganitkevitch et al., 2012

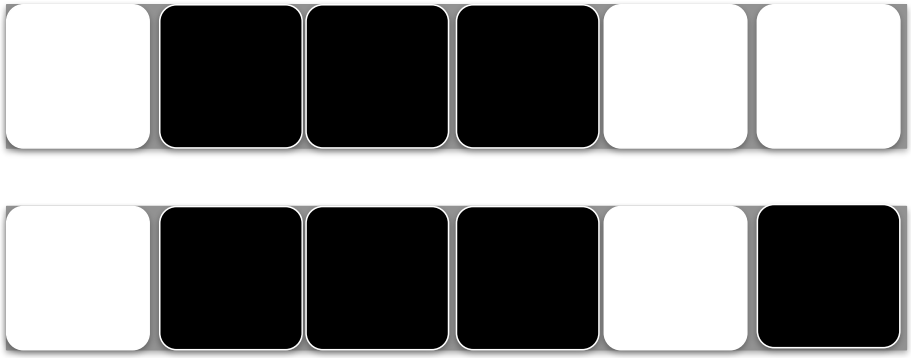
Locality Sensitive Hashing

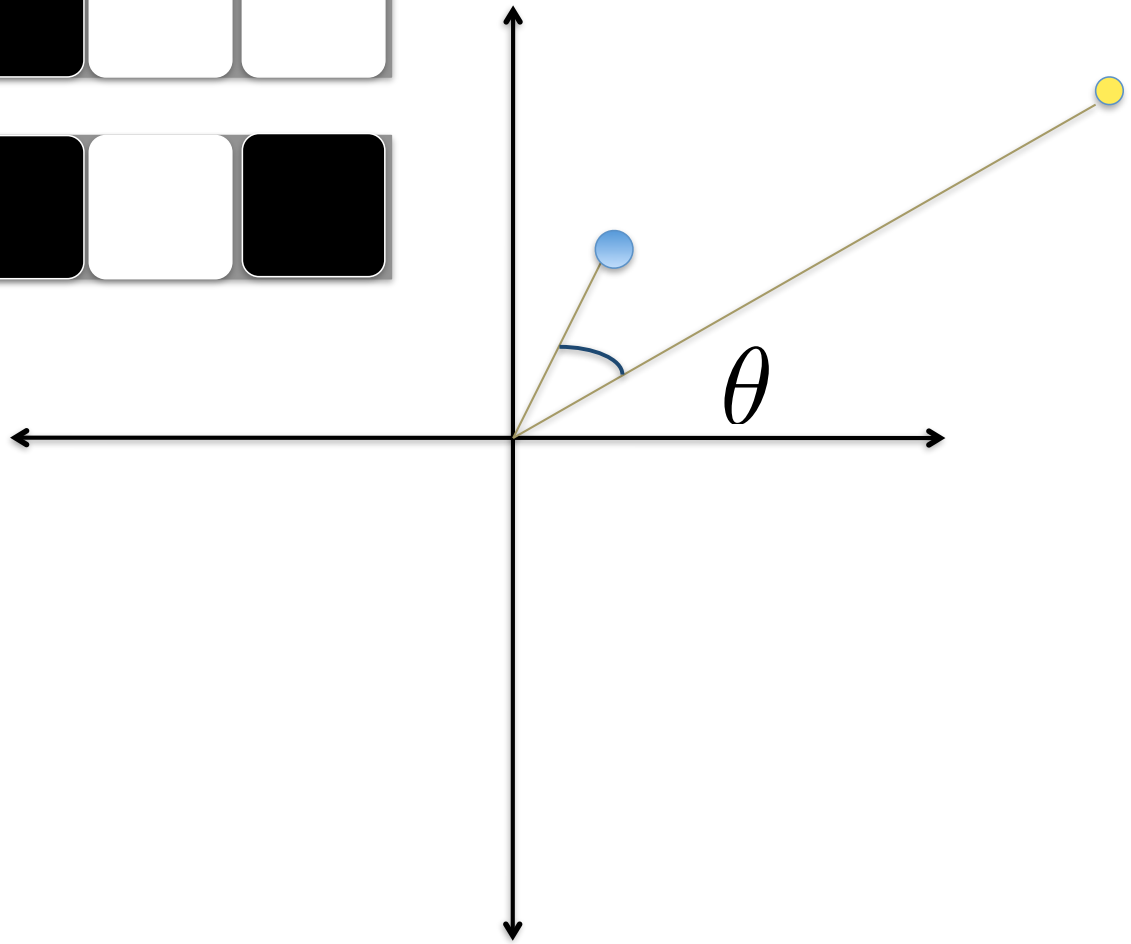


Locality Sensitive Hashing

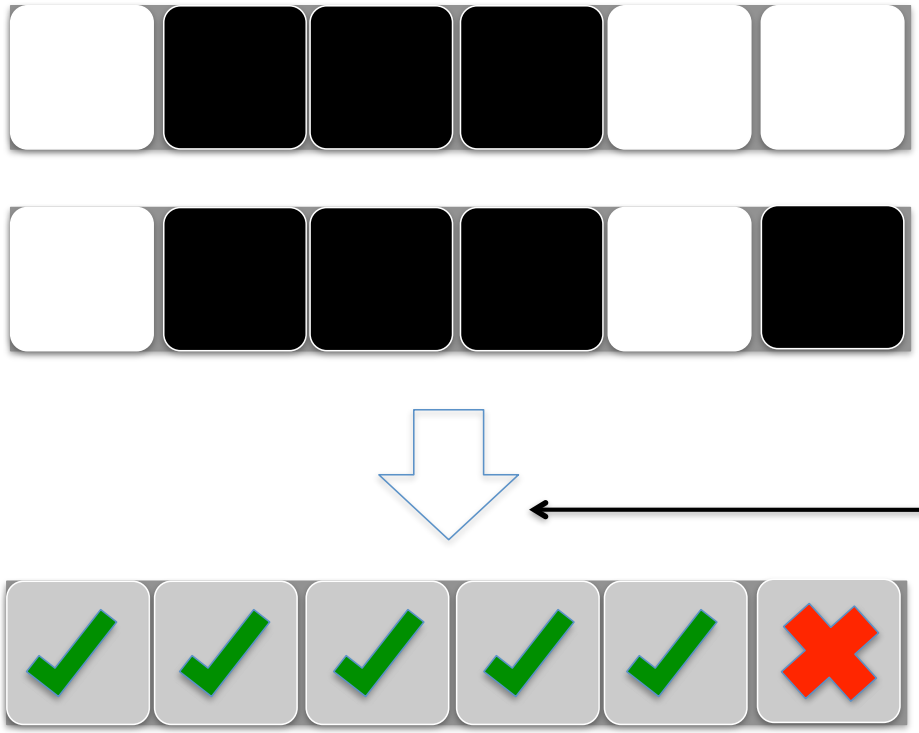


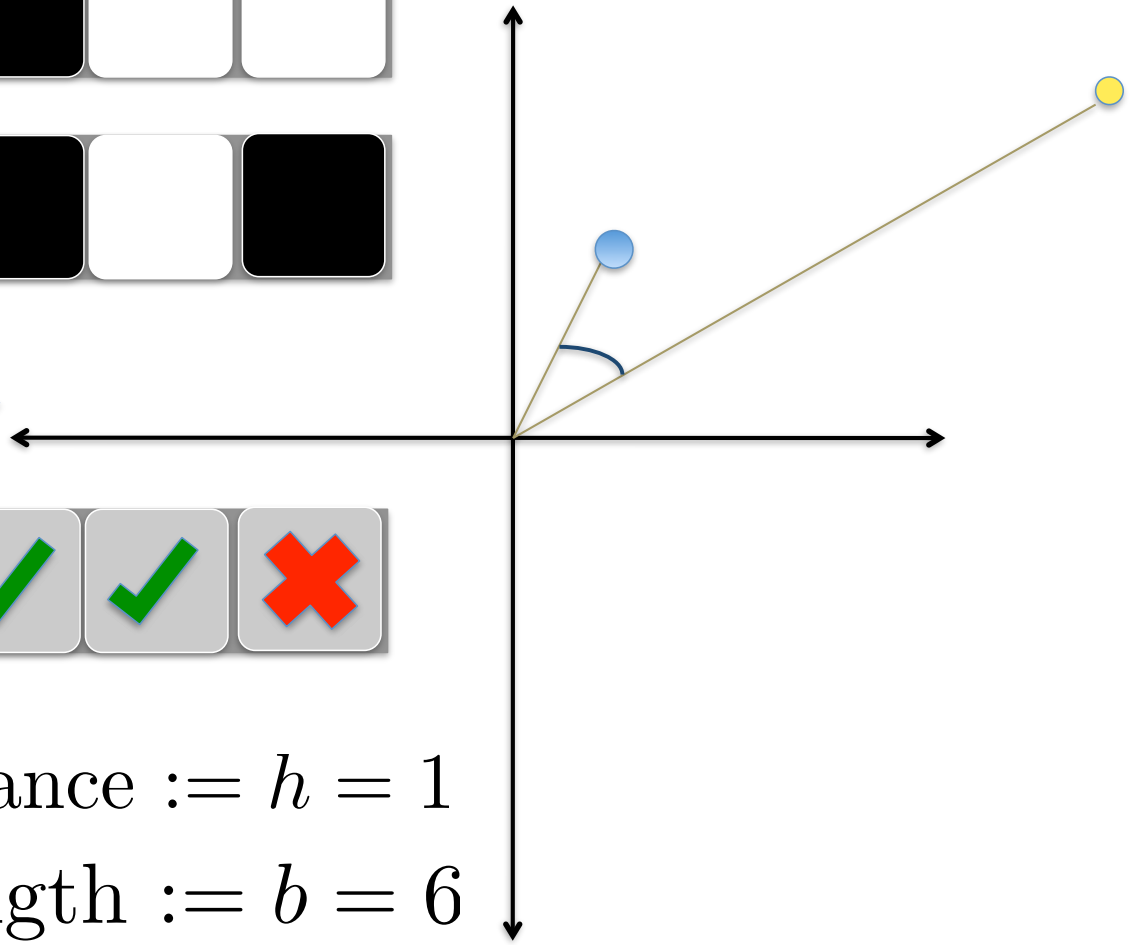
Locality Sensitive Hashing





Locality Sensitive Hashing



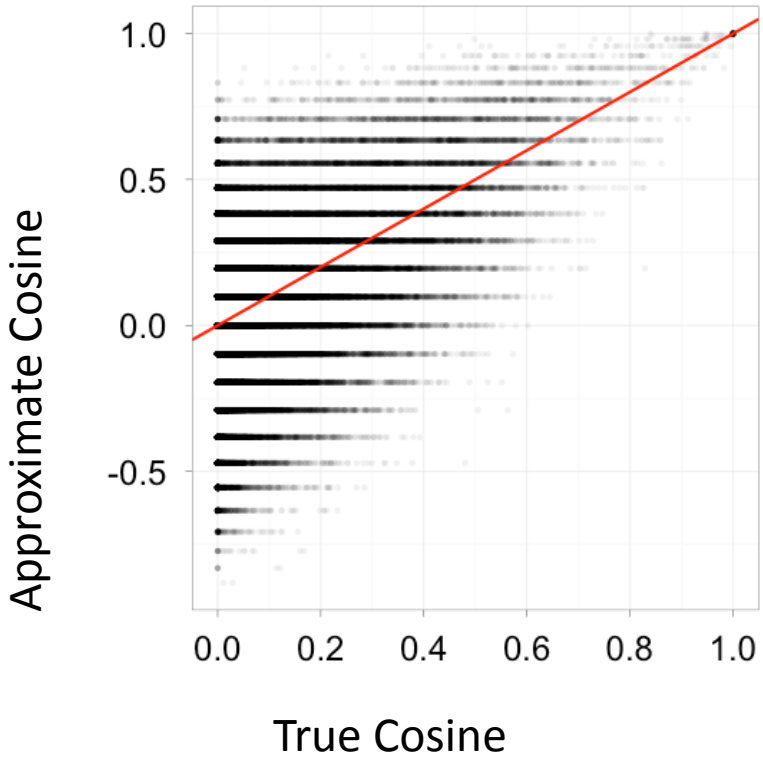


$$\text{Hamming Distance} := h = 1$$

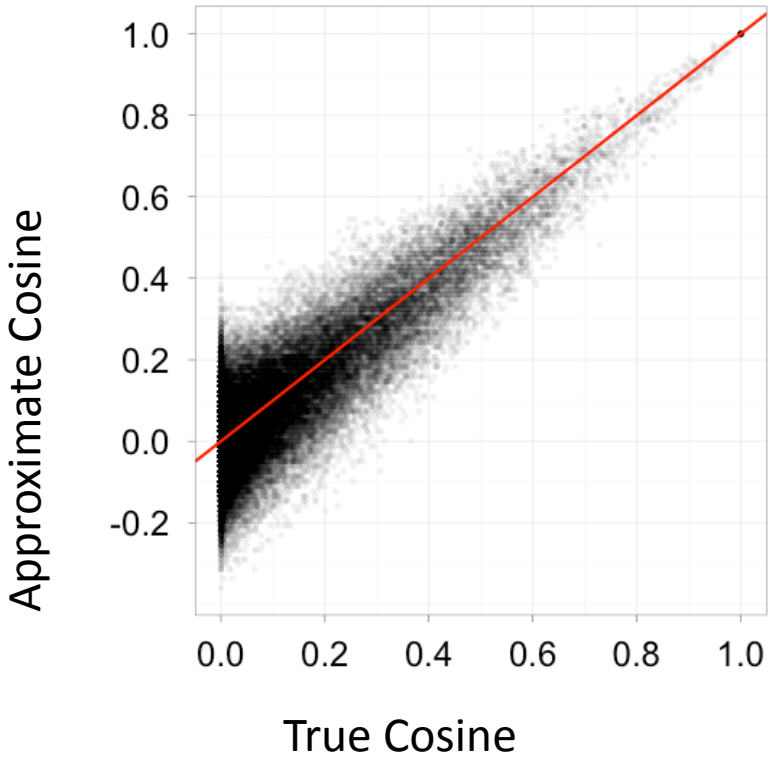
$$\text{Signature Length} := b = 6$$

Accuracy as function of bit length

32 bit signatures



256 bit signatures

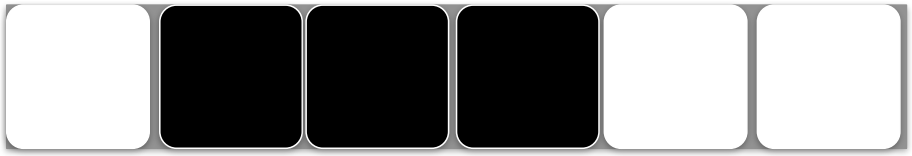


Cheap

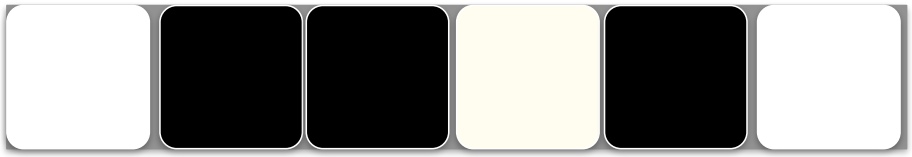
Accurate

Signature Collection

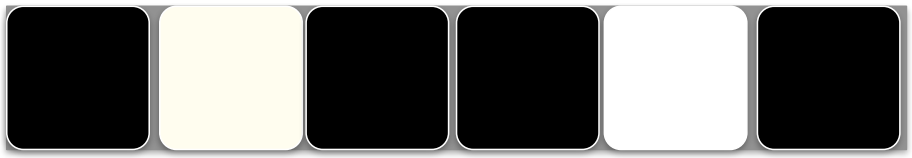
of



‘s



the long term



in the long term

Syntactic Paraphrase Similarity

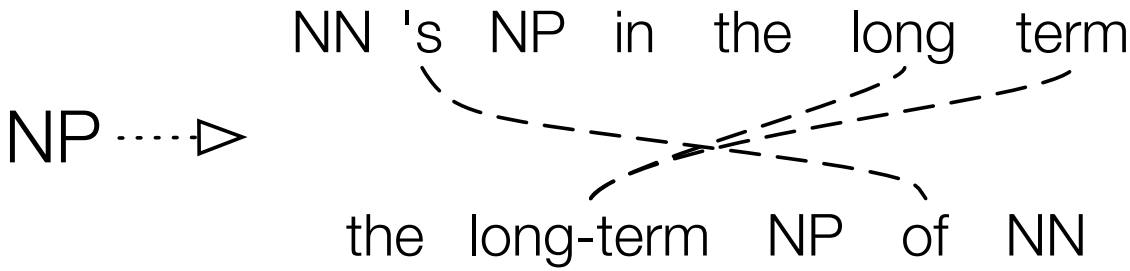
NN 's NP in the long term

NP ⇢

the long-term NP of NN

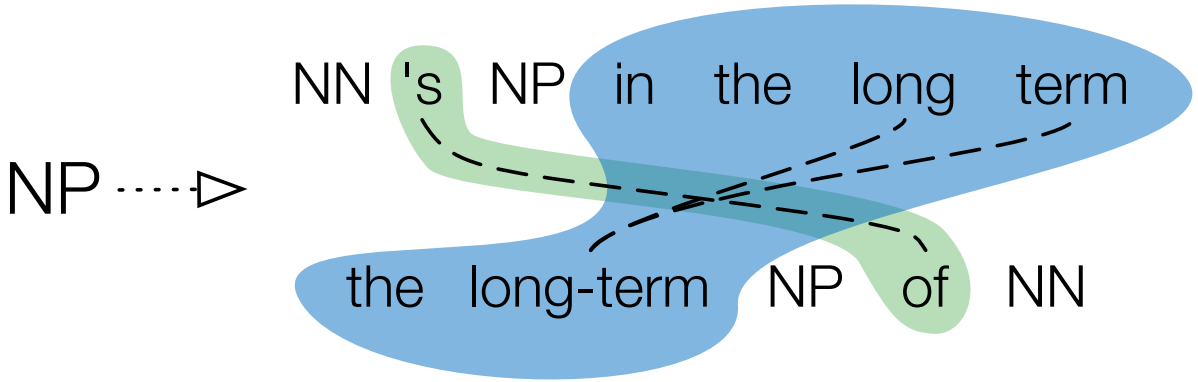
Ganitkevitch et al., 2012

Syntactic Paraphrase Similarity



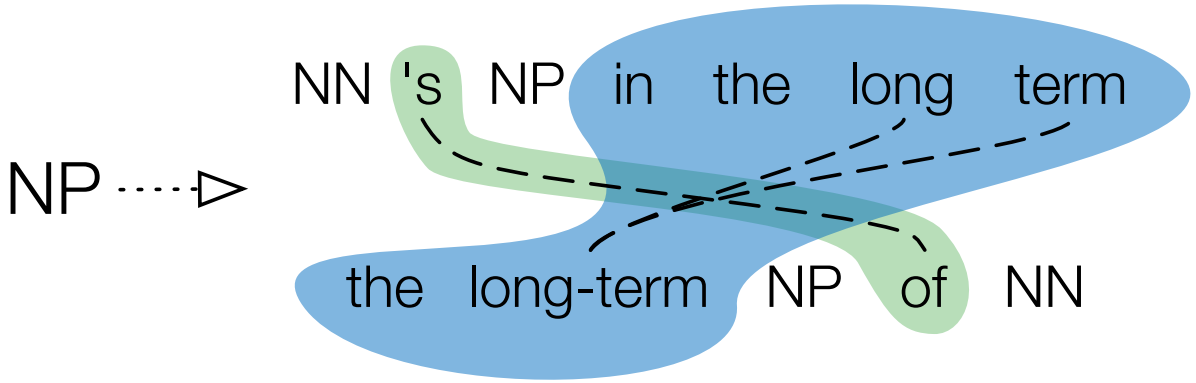
Ganitkevitch et al., 2012

Syntactic Paraphrase Similarity



Ganitkevitch et al., 2012

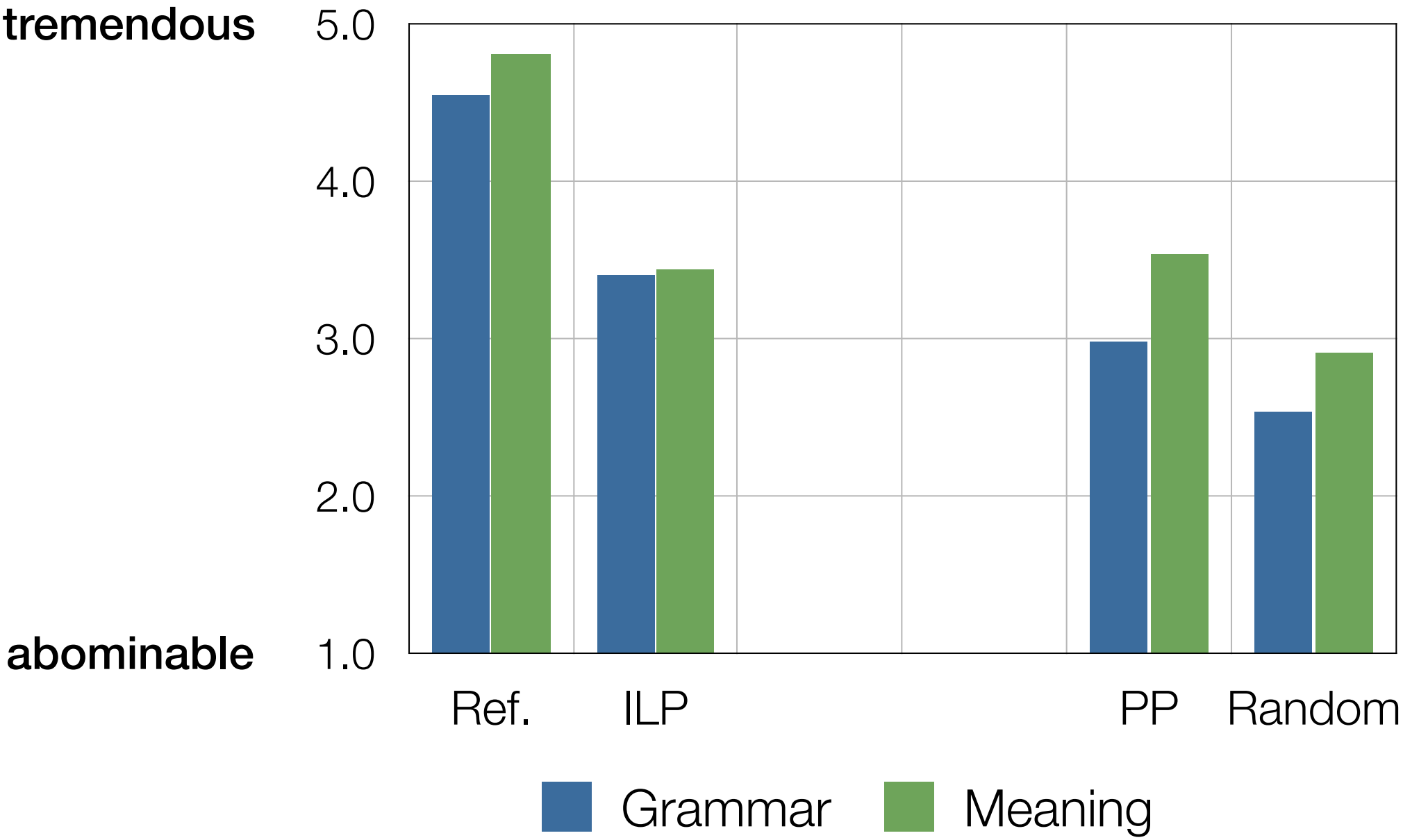
Syntactic Paraphrase Similarity



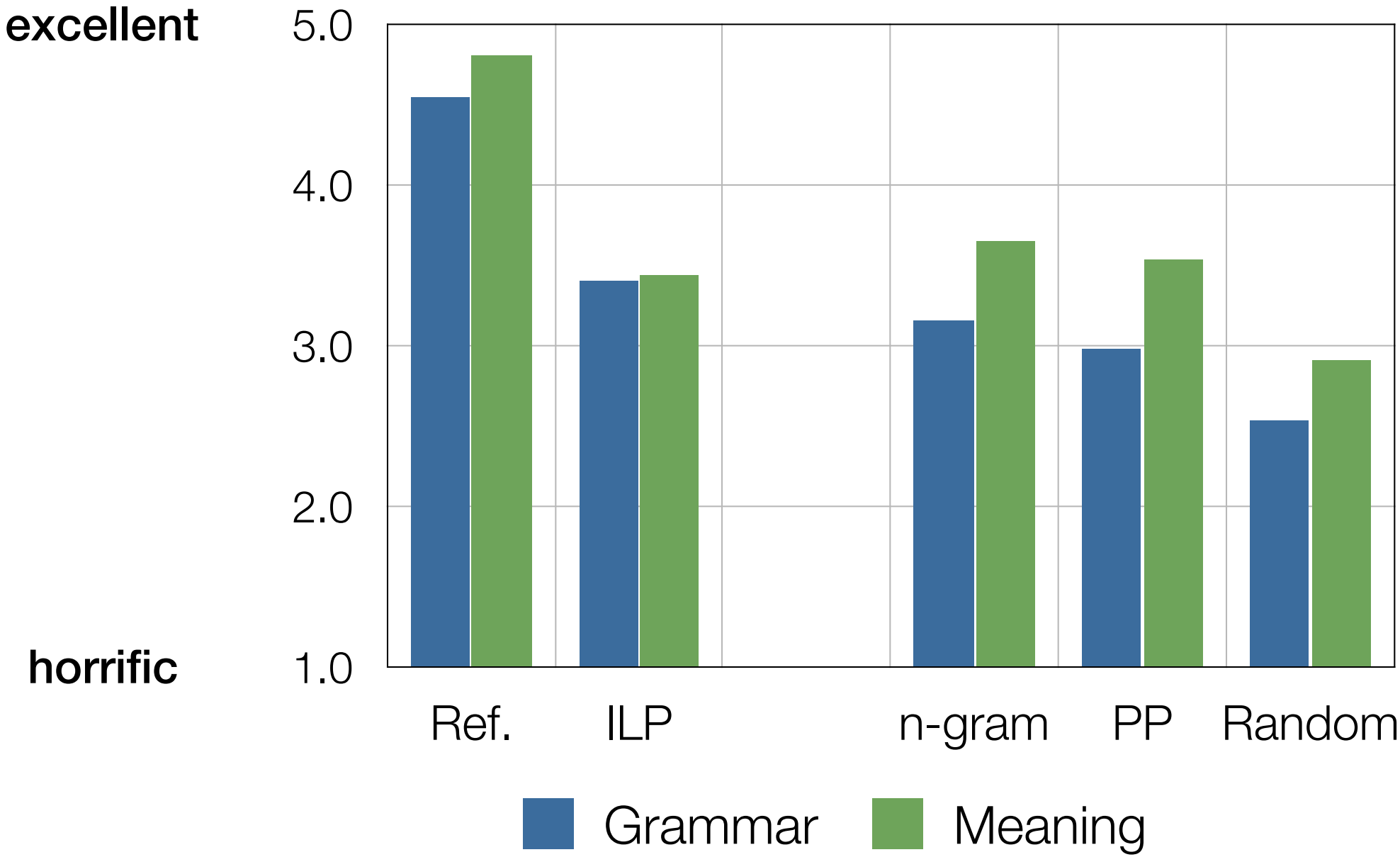
$$sim(\mathbf{r}) = \frac{1}{2}\left(sim\begin{pmatrix}\text{the long-term}\\ \text{in the long term}\end{pmatrix} + sim\begin{pmatrix}\text{'s}\\ \text{of}\end{pmatrix} \right)$$

Ganitkevitch et al., 2012

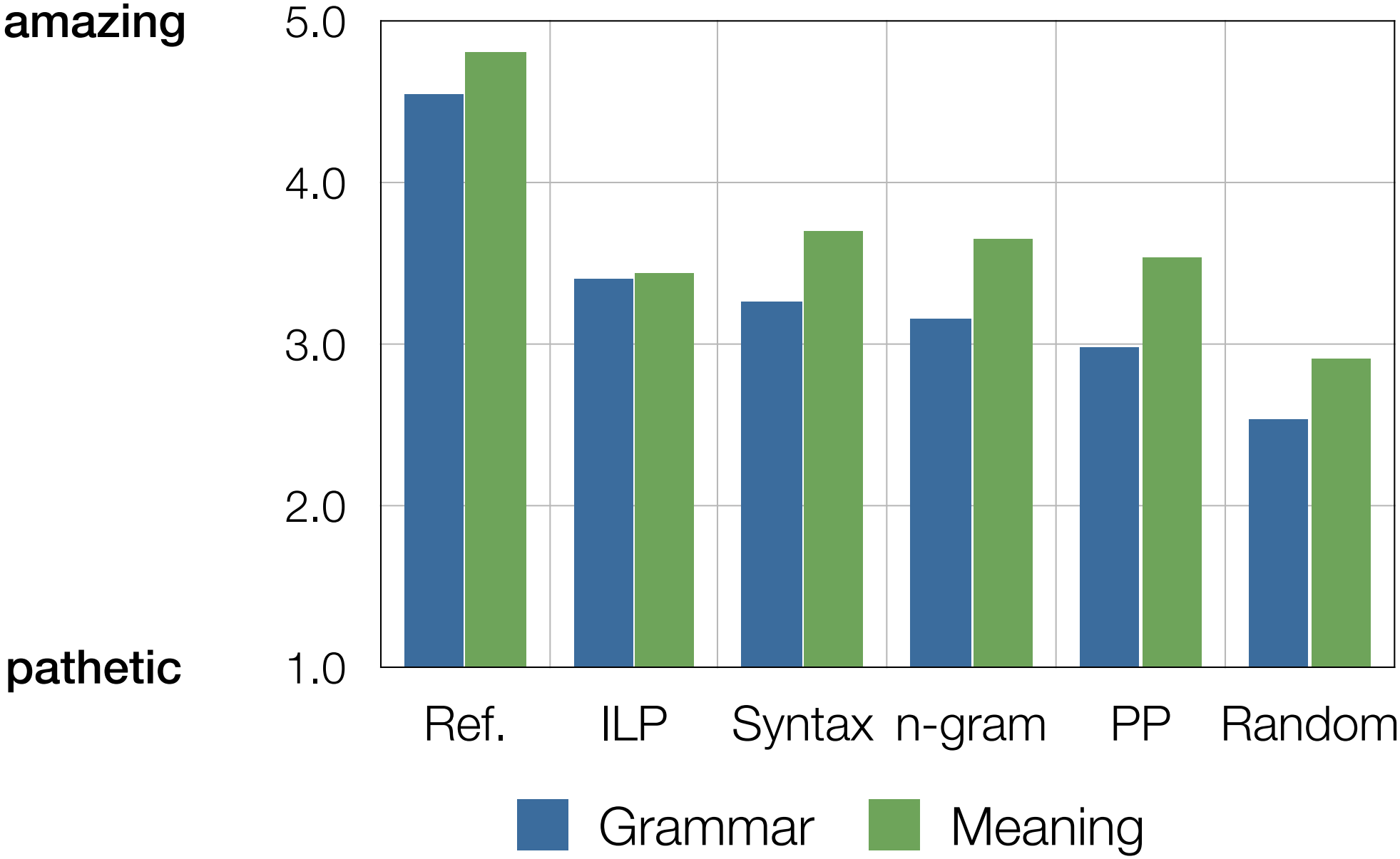
Compression Results



Compression Results



Compression Results



Paraphrase Extraction

Text-to-Text Applications

Scaling to PPDB

Contributions

Bilingual Data

Europarl, Fr-En Gigaword, News Commentary, UN, JRC Acquis, GALE, NIST Urdu, OpenST, TAUS

106 million sentence pairs

2 billion English words

22 pivot languages

Ganitkevitch et al., 2013

Monolingual Data

Google *n*-grams

- Collection of 1 trillion tokens with counts
- Based on vast amounts of text

Annotated Gigaword (Napoles et al., 2012)

- Collection of 4 billion words, parsed and tagged

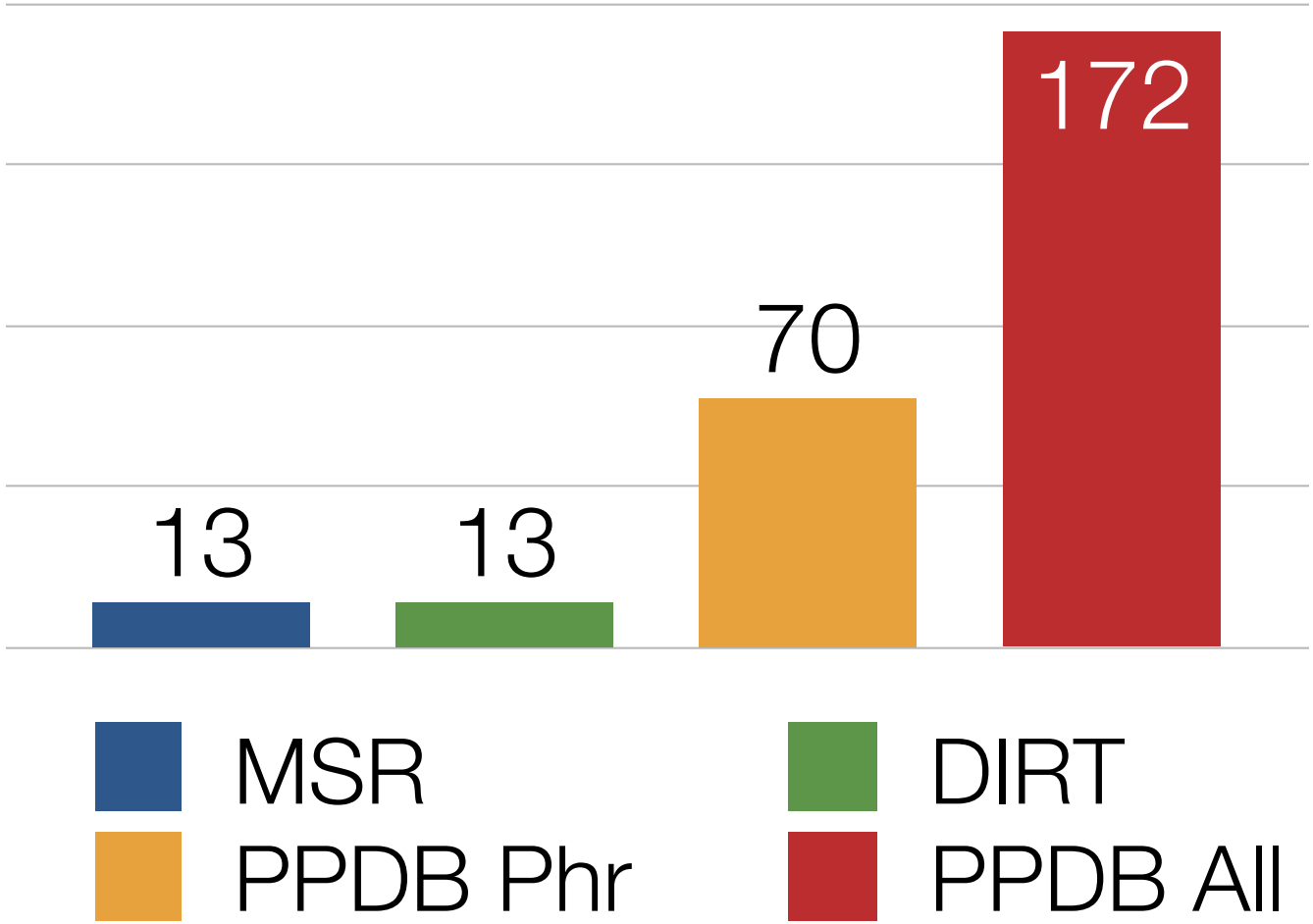
Ganitkevitch et al., 2013

PPDB:Eng

	Paraphrases
Lexical	7.6 M
One-Many	2.8 M
Phrasal	68.4 M
Syntactic	93.6 M
Total	172.4 M

Ganitkevitch et al., 2013

Other Resources



Ganitkevitch et al., 2013

Machinery

Joshua & Thrax

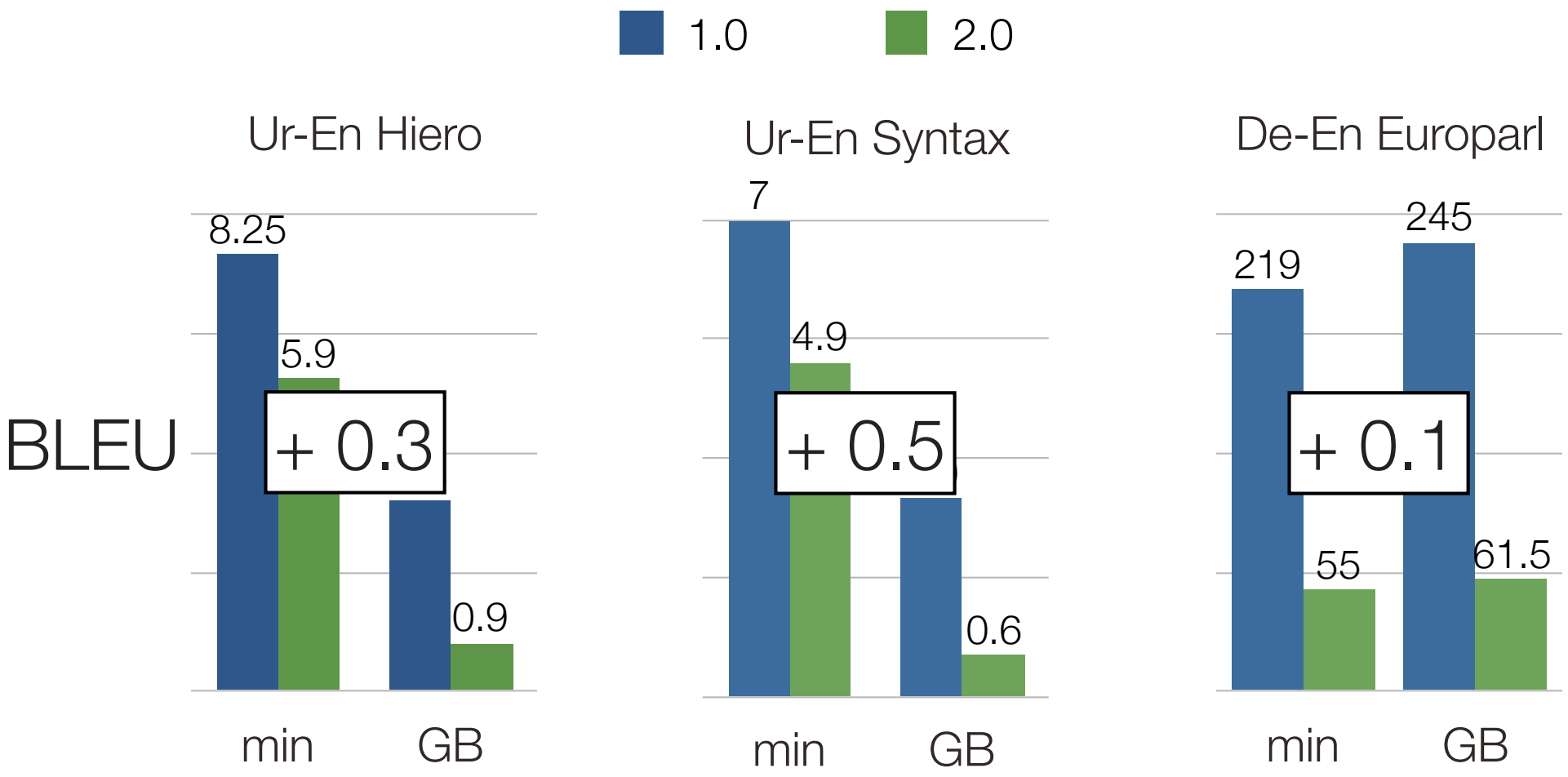
SCFG-based decoder for translation and paraphrasing

Includes: parallelized extraction of paraphrase grammars & distributional signatures

Jerboa

Toolkit for randomized and streaming algorithms (LSH, Bloom filters, etc.)

Thrax 2.0



75% slimmer, up to 300% faster

Ganitkevitch et al., 2012

Extraction on Amazon

PPDB:Spa

100 machines, 4 cores, 15GB RAM

~ 8 hours on spot instances

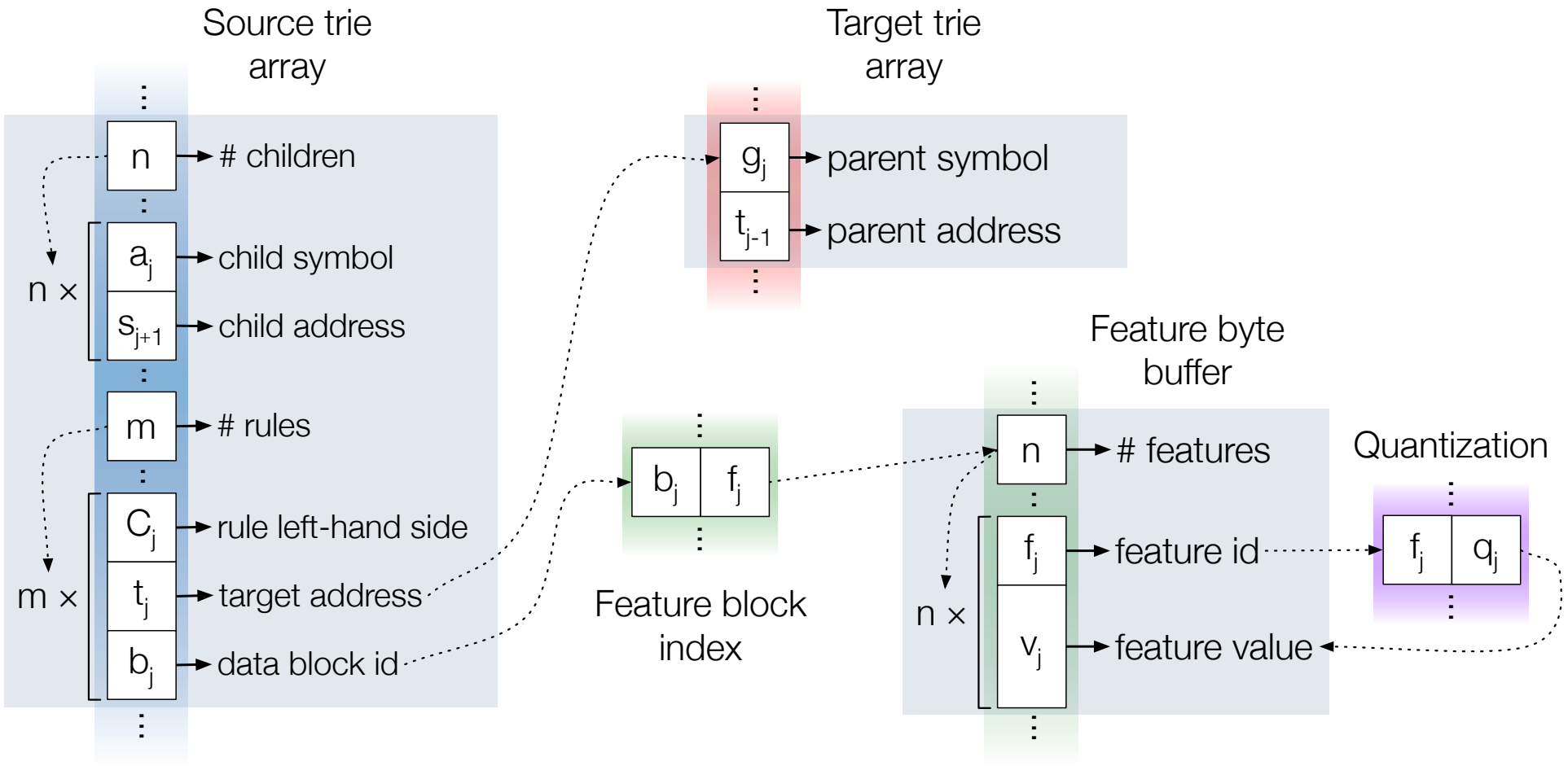
~ $100

Text-to-Text with PPDB?

Filtering PPDB to a dev & test set:

- 118 million paraphrases
- Needs ~192 GB of memory
- Takes hours to load

Packed Grammars



Ganitkevitch et al., 2012

Smaller & Faster

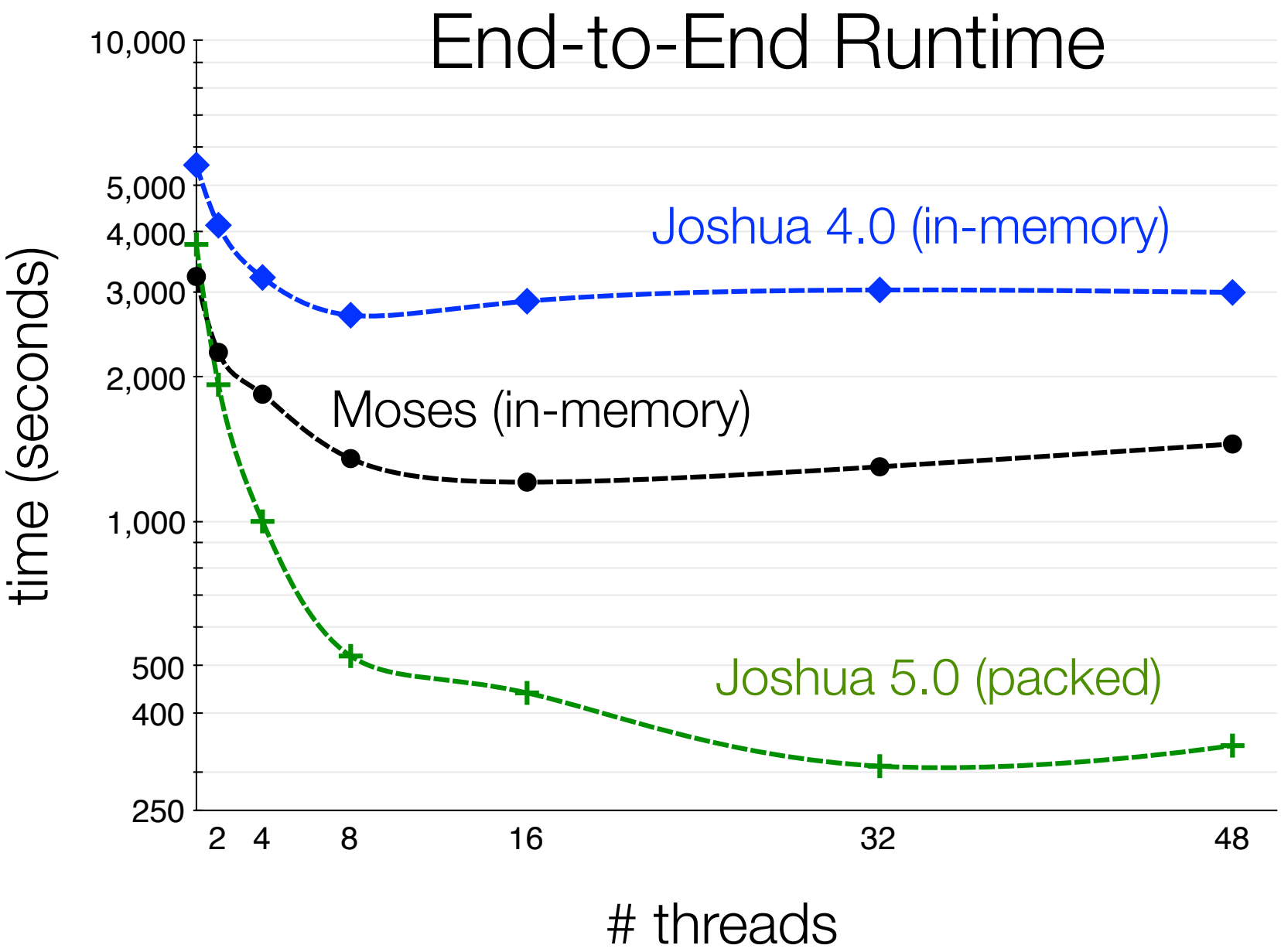
Takes up ~6.5 GB of space (~3% of unpacked)

Little over 2 minutes to load

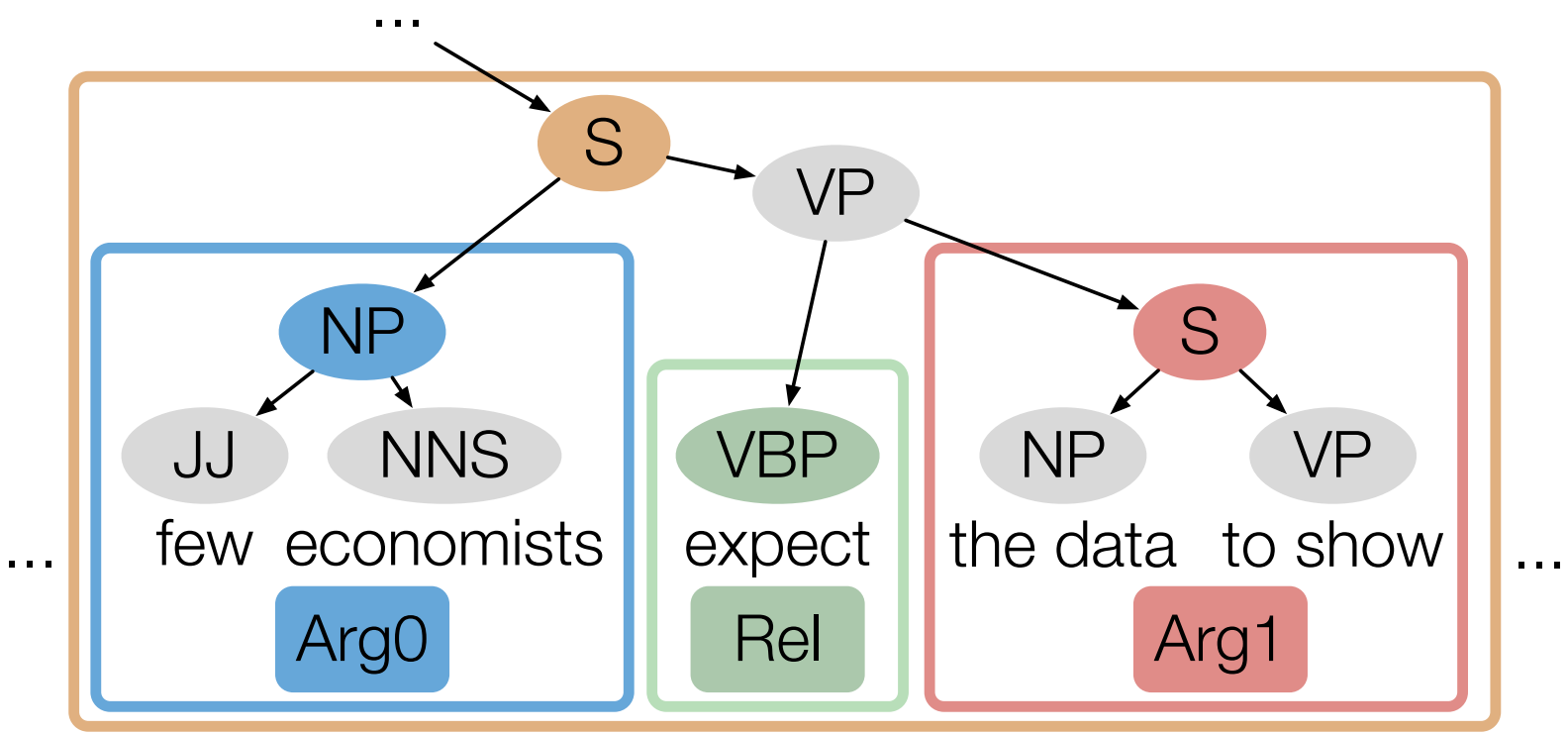
Slightly slower decoding

Ganitkevitch et al., 2012

Joshua


End-to-End Runtime
10,000
5,000
4,000
3,000
2,000
1,000
500
400
250
time (seconds)
Joshua 4.0 (in-memory)
Moses (in-memory)
Joshua 5.0 (packed)
2
4
8
16
32
48
threads

Expanding Propbank



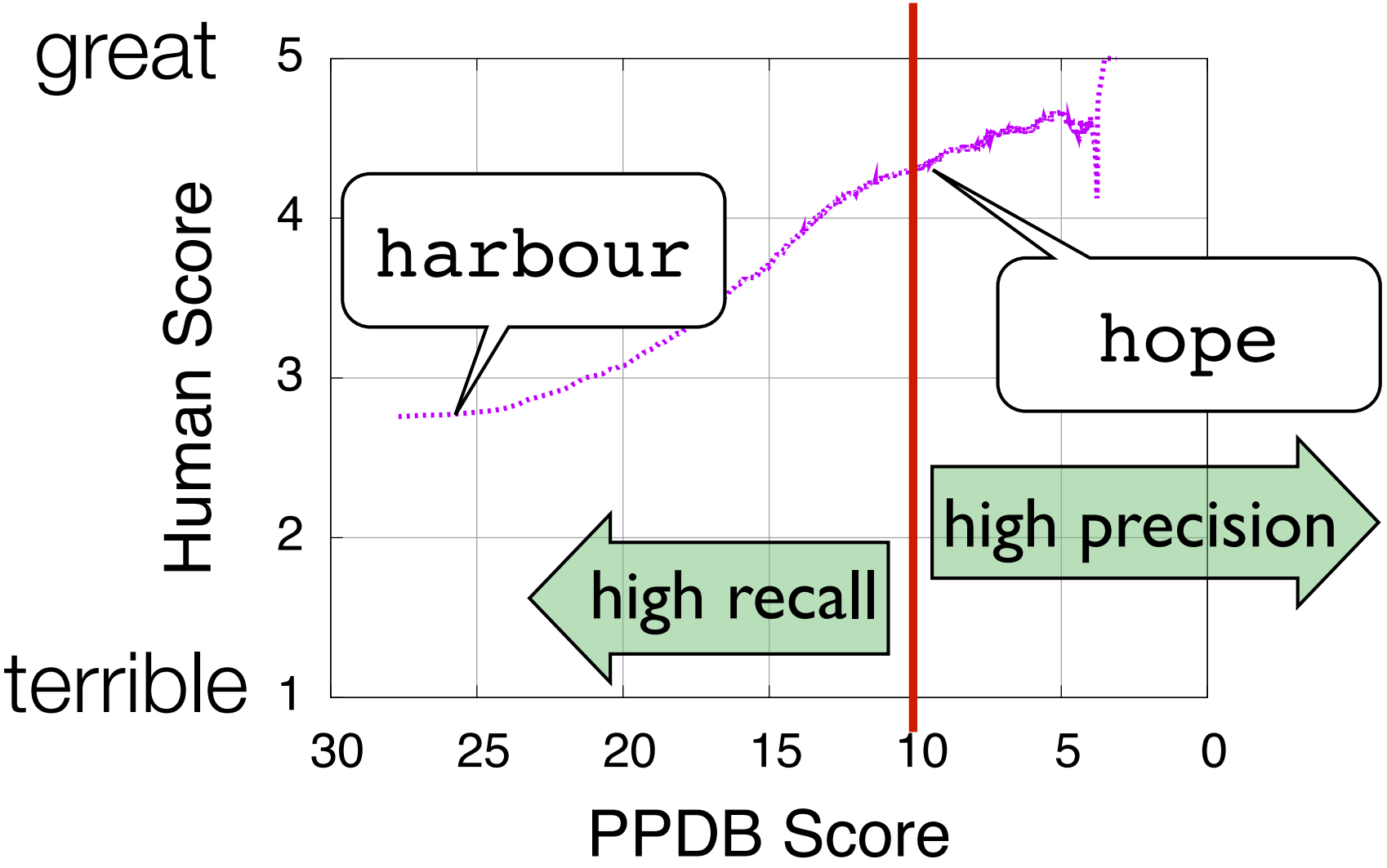
Ganitkevitch et al., 2013

expect

```
7.40909     [VBP]     await
7.48333     [VBP]     wait
8.55643     [VBP]     hope
9.21413     [VBP]     anticipate
...
26.7452     [VBP]     grant
26.8336     [VBP]     let
26.8438     [VBP]     save
26.8846     [VBP]     harbour
```

Ganitkevitch et al., 2013

Do the Scores Work?



Ganitkevitch et al., 2013

Paraphrase Extraction

Text-to-Text Applications

Scaling to PPDB

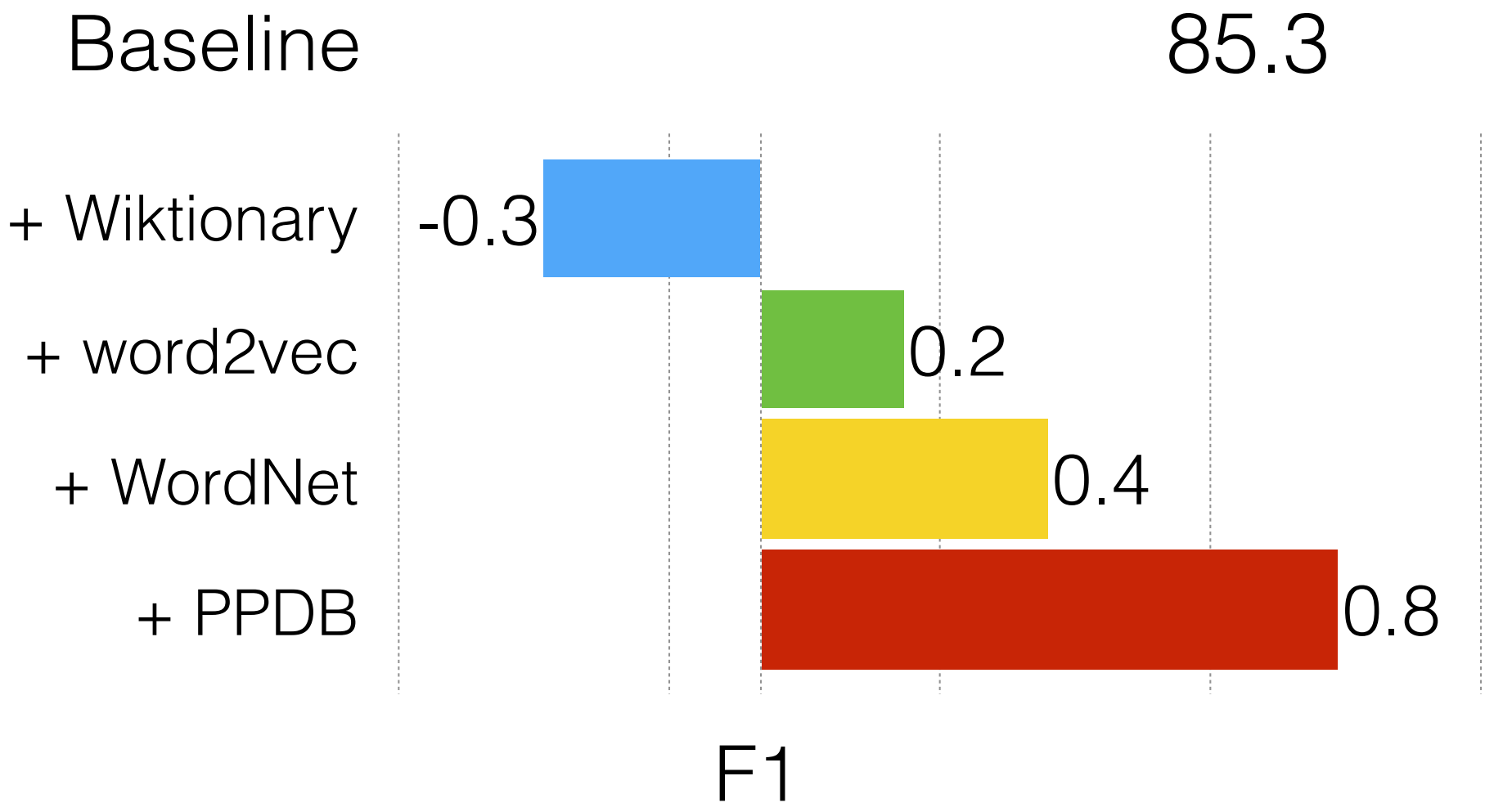
Contributions

Monolingual Alignment

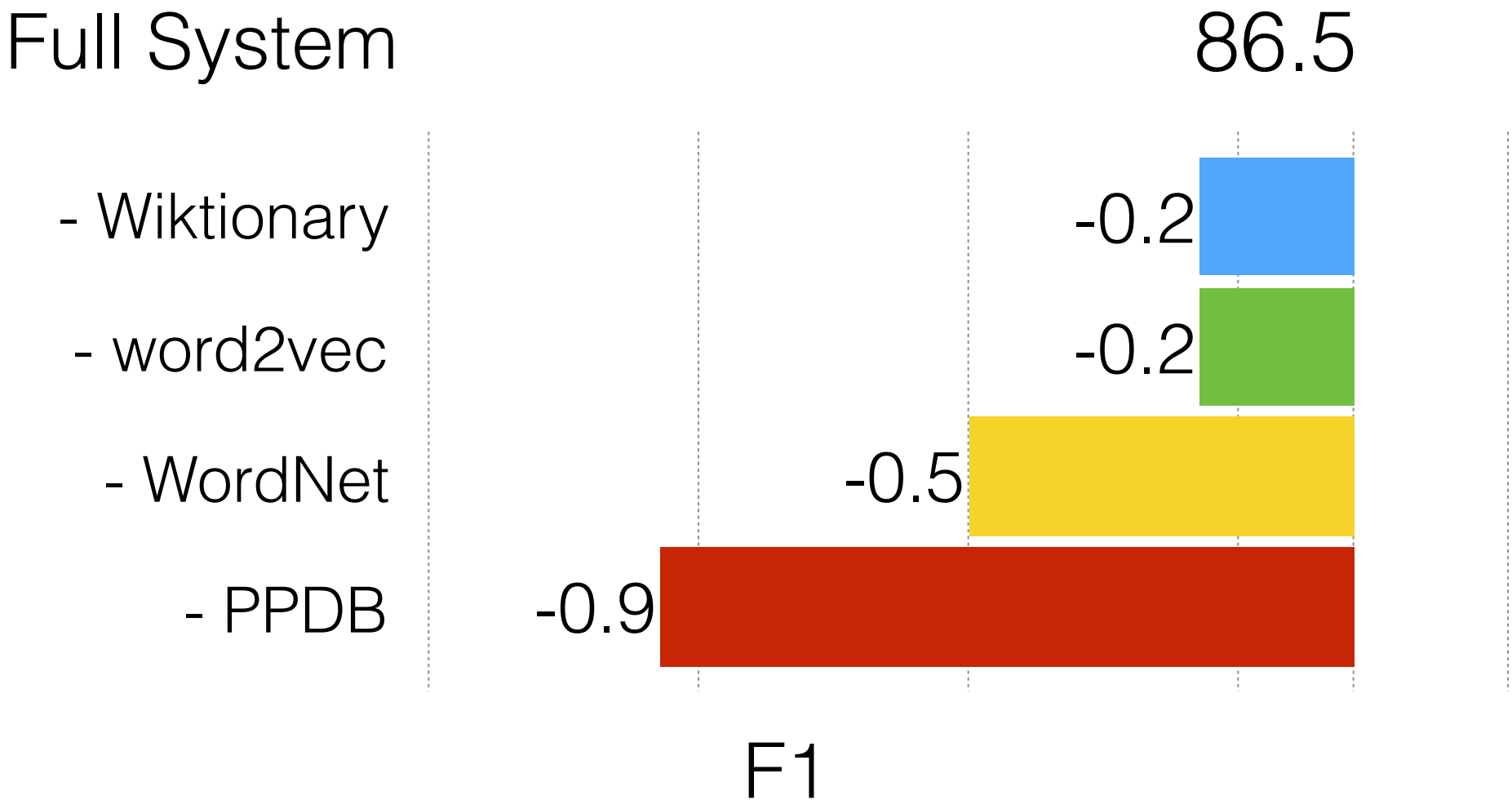
Jacana Aligner (Yao et al., ACL 2013)

- Wiktionary (90k)
- Snow's 400k-WordNet
- word2vec (trained on enwik9)
- PPDB (270k)

Monolingual Alignment



Monolingual Alignment



Retrofitting Embeddings

Semantic Relatedness (Faruqui et al., ’14)

- WordNet
- FrameNet
- PPDB (XL)

Other Languages

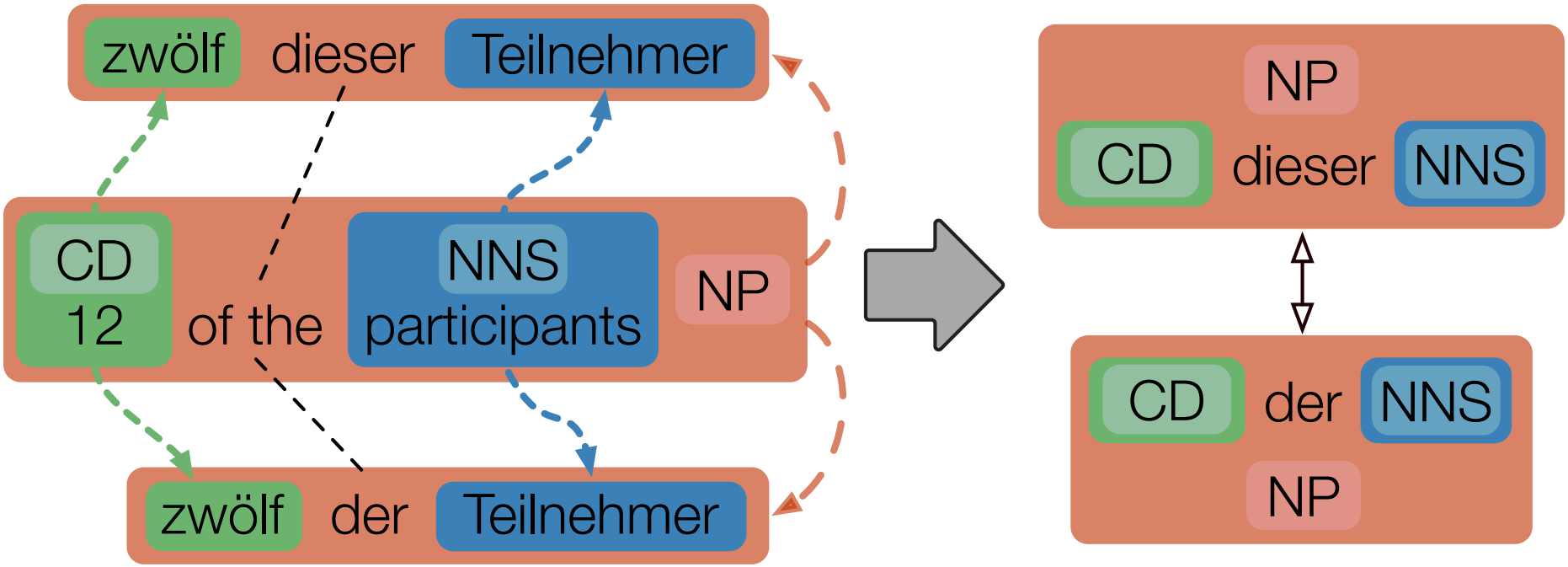
PPDB bilingual data:

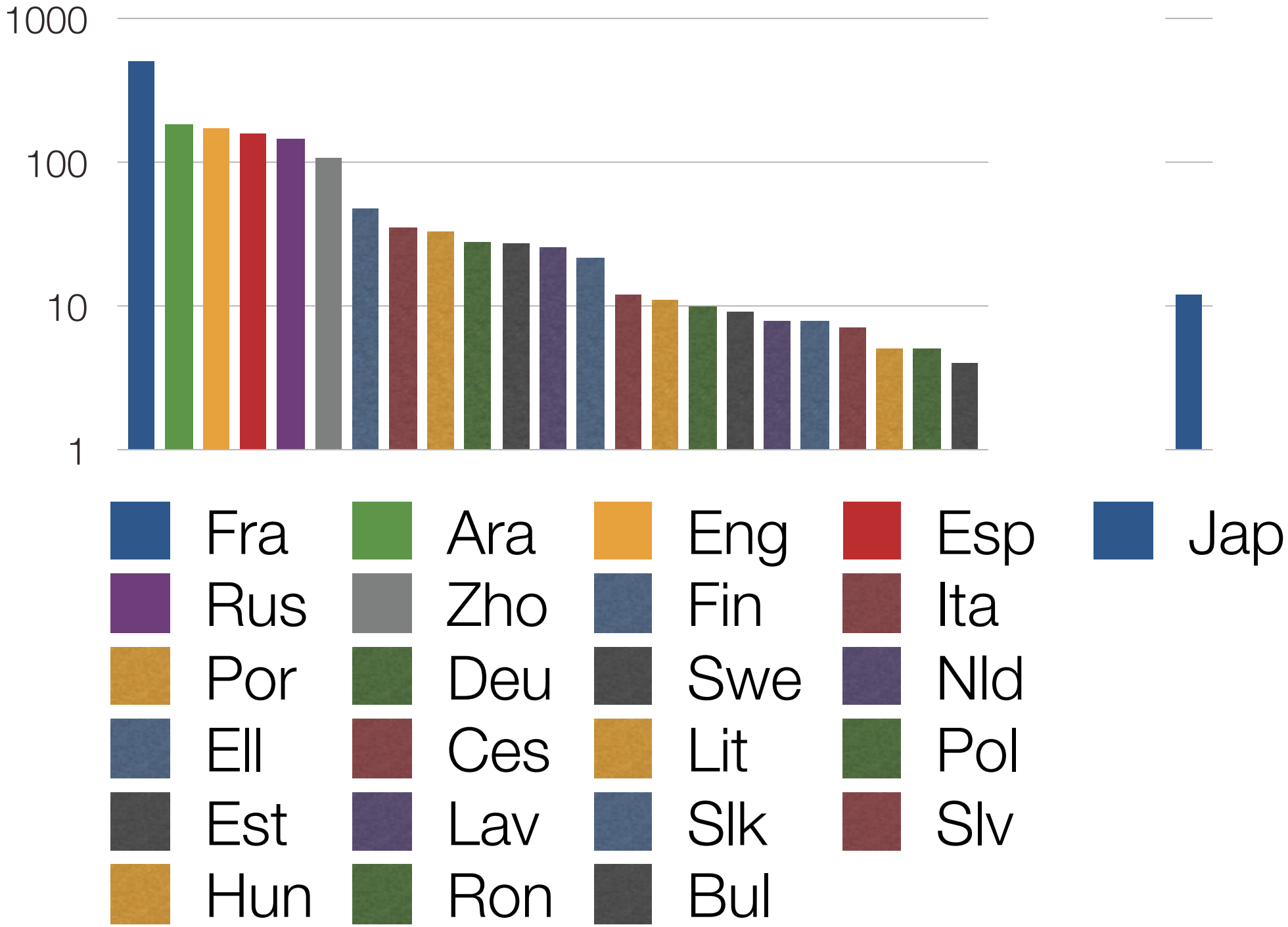
106 million sentence pairs

2 billion English words

22 pivot languages

Projected Labels




1000
100
10
1
Fra
Ara
Eng
Esp
Jap
Rus
Zho
Fin
Ita
Por
Deu
Swe
Nld
Ell
Ces
Lit
Pol
Est
Lav
Slk
Slv
Hun
Ron
Bul

Coverage Expansion

Plug-in paraphrasing for NLP:

Query expansion, feature mapping, input normalization,...


AT&T
2:47 AM
“Siri, tell me some astounding facts about potatoes”


AT&T
2:47 AM
“Siri, tell me some astounding facts about potatoes”

astounding

astonishing amazing staggering surprising stunning startling mind-boggling incredible striking breathtaking impressive unbelievable uncanny perplexing narcotic stunning shocking appalling remarkable astonishing extraordinary upside-down dreadful puzzling awesome miraculous fantastic meteoric horrific hideous dazzling alarming surprised horrendous dramatic gruesome overwhelming frightening phenomenal terrific tremendous spectacular dire strange narcotics reversal

astounding

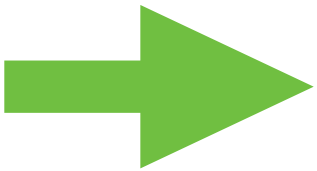
awful unexpected wonderful atrocious precipitous non-recurring paradoxical odd whopping growth intimidating terrible unusual horrible awe-inspiring great embarrassing sensational runaway gorgeous wacky intriguing unreal horrifying exciting splendid special unthinkable surprising daunting egregious heinous marvellous shocking inconceivable disconcerting explosive notable brilliant noticeable incredible rapid fabulous excellent superb bizarre outrageous impossible

astounding

confusing troubling fascinating unlikely noteworthy sheer massive unprecedented non-routine exceptional unbelievable memorable amazing extra-ordinary curious formidable groovy growing peculiar mighty insane distinguished spectacular crazy weird wondrous impressed compelling cool sterling stellar extra suspicious commendable freaky sizeable admirable immense singular unspeakable unmatched magnificent conspicuous unconventional scary outstanding jaw-dropping wretched

Canonicalization

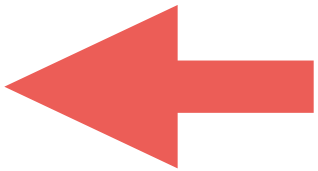
astonishing
amazing
staggering
surprising
stunning
startling
mind-boggling
incredible
striking
breathtaking
impressive
unbelievable
uncanny
perplexing
narcotic
stunning
shocking
appalling
remarkable
astonishing
extraordinary
upside-down
dreadful
puzzling
awesome
miraculous
fantastic
meteoric
horrific
hideous
dazzling
alarming
surprised
horrendous
dramatic
gruesome
overwhelming
frightening
phenomenal



astounding

Expansion

astonishing
amazing
staggering
surprising
stunning
startling
mind-boggling
incredible
striking
breathtaking
impressive
unbelievable
uncanny
perplexing
narcotic
stunning
shocking
appalling
remarkable
astonishing
extraordinary
upside-down
dreadful
puzzling
awesome
miraculous
fantastic
meteoric
horrific
hideous
dazzling
alarming
surprised
horrendous
dramatic
gruesome
overwhelming
frightening
phenomenal



astounding

Authoring

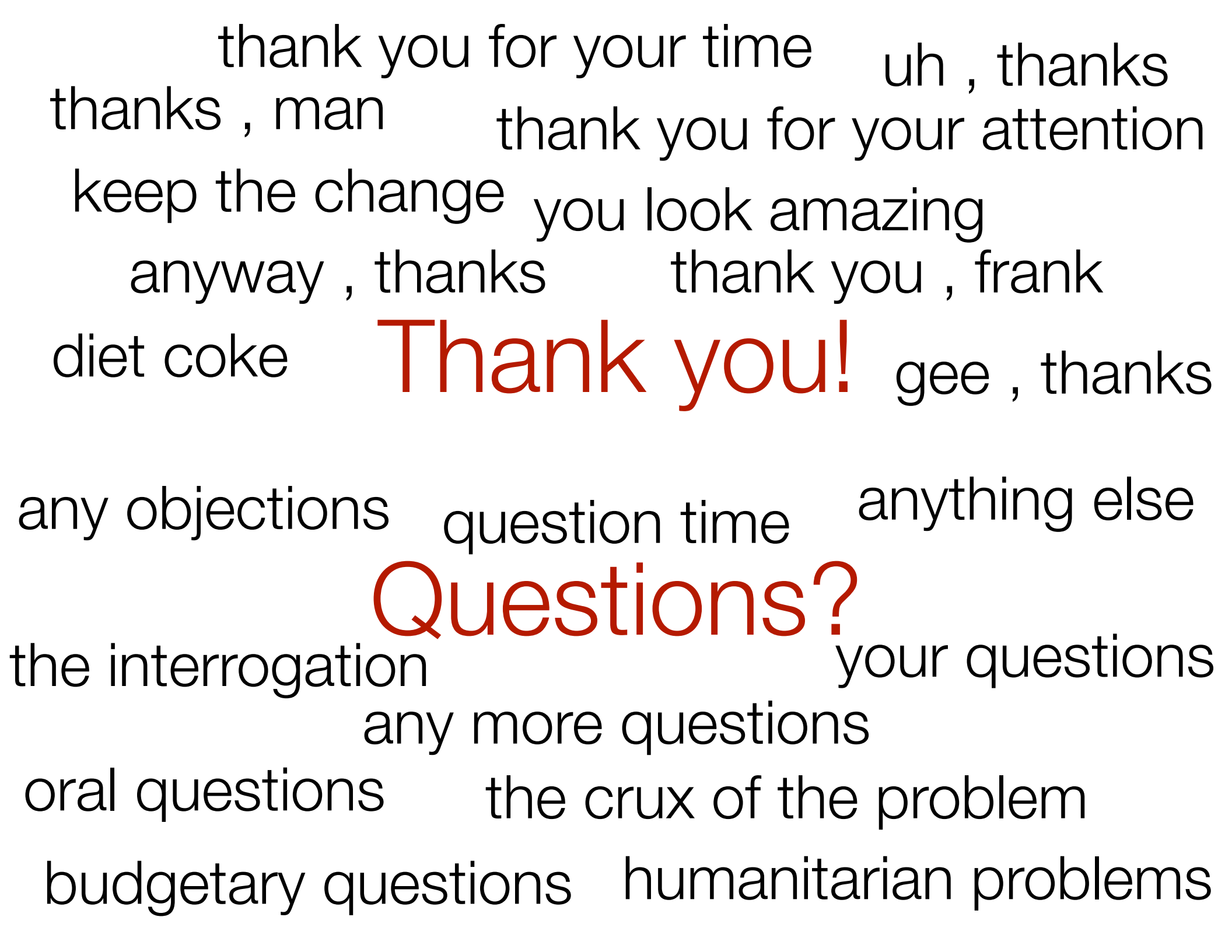
Facebook, 2014

Facebook, 2014

The Real World™

To Summarize

- Scaled rich paraphrase extraction to large bitext corpora
- Integrated distributional signal with bitext-derived paraphrases
- Open source infrastructure to enable text-to-text with PPDB
- Created a resource for multilingual coverage expansion in NLP


thank you for your time
uh , thanks
thanks , man
thank you for your attention
keep the change
you look amazing
anyway , thanks
thank you , frank
diet coke
Thank you!
gee , thanks
any objections
question time
anything else
Questions?
the interrogation
your questions
any more questions
oral questions
the crux of the problem
budgetary questions
humanitarian problems