Syntactic Annotations for the Google Books Ngram Corpus

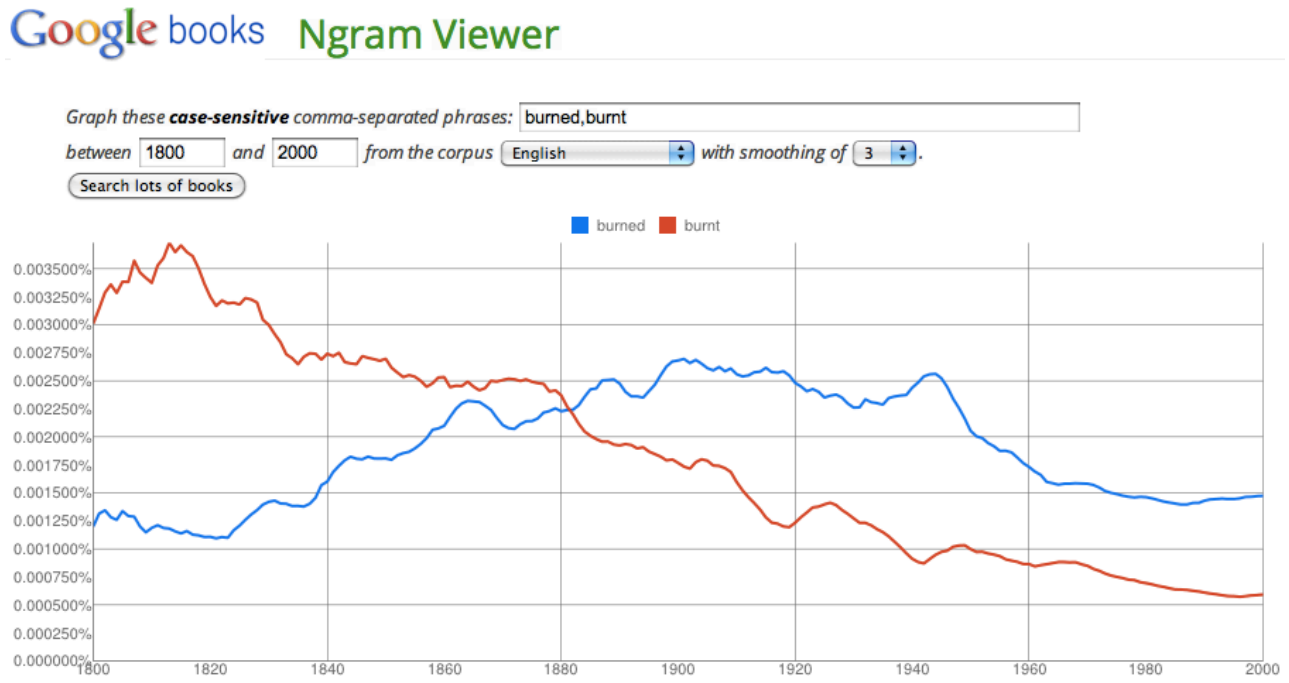
Research at Google

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden,
Jon Orwant, Will Brockman and Slav Petrov

The Google Books Ngram Corpus

[Michel et al. '11]

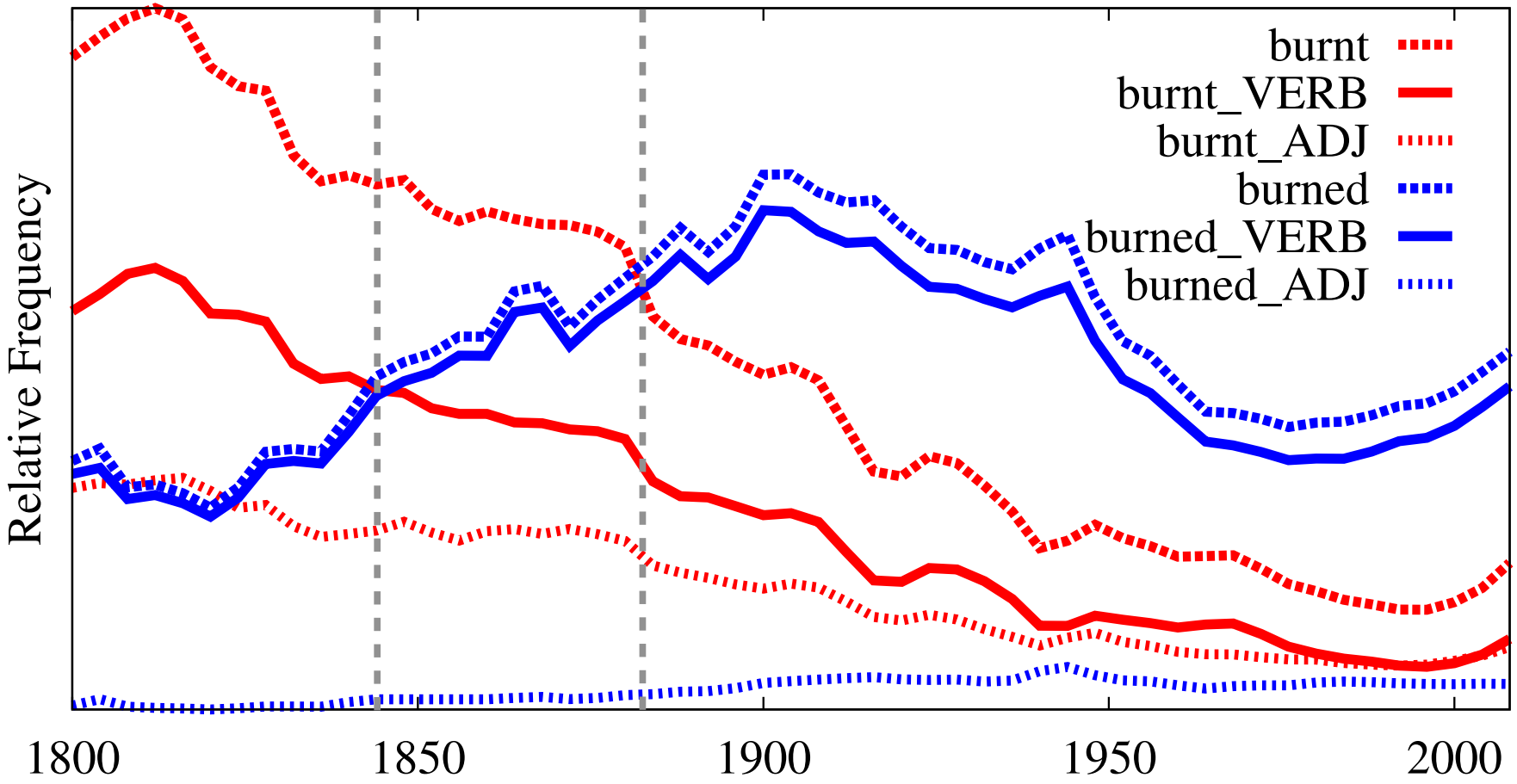
- Allows quantitative analysis of ngram frequencies over time
- Live at: *http://books.google.com/ngrams*



- Limitation: Words can have different syntactic (and semantic) interpretations that is not visible in the raw surface strings!

burned vs. burnt

- Syntactic disambiguation shows that the regularization of the past tense of the verb 'to burn' started much earlier than the raw frequencies indicate.



Approach

- Train supervised taggers and parsers for each language.
- Sentence break, tag and parse the entire collection.
- Extract ngrams from all sentences.
- Aggregate ngram counts.
- Discard rare ngrams.

- Tagger: Conditional Random Field with prefix, suffix and cluster features.
- Parser: Deterministic Shift-Reduce Parser with standard feature set.

Universal Part-of-Speech [Petrov, Das & McDonald '12]

- Provide a language independent and intuitive interface:

VERB	**PRON**
NOUN	**DET**
ADJ	**ADP**
ADV	**PRT**
NUM	**CONJ**
X	**.**

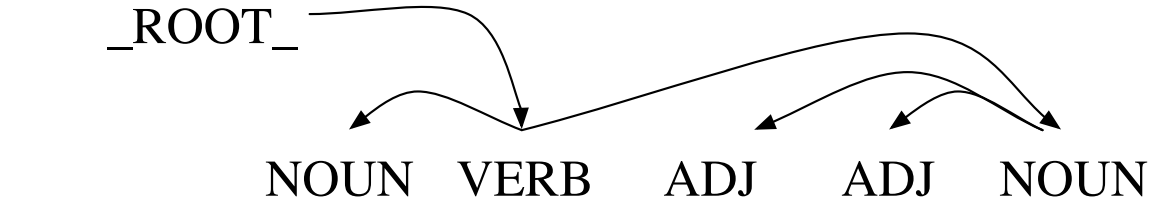
- Manually map language-specific tags to universal tags:
  - English (Penn Treebank):

NN	→	**NOUN**	PRP	→	**PRON**
NNS	→	**NOUN**	PRP$	→	**PRON**
NNP	→	**NOUN**	WP	→	**PRON**
NNPS	→	**NOUN**	WP$	→	**PRON**

  - German (Negra Treebank):

NE	→	**NOUN**	PDAT	→	**PRON**
NN	→	**NOUN**	PDS	→	**PRON**
			PIAT	→	**PRON**
			PIDAT	→	**PRON**
			PIS	→	**PRON**
			PPER	→	**PRON**
			...	→	**...**

(Annotated) Ngrams



Raw Ngrams

John	*short*
John has	*...*
...	*short black hair*

Annotated Ngrams

START John	*John_NOUN*	*hair=>short*	*hair=>short_ADJ*
...	*John has_VERB*	*hair=>black*	*...*
*hair _END_*	*John _VERB_ short*	*_NOUN_<=has*	*_ROOT_=>has*

Domain Adaptation

- Use word cluster features to handle lexical sparsity from (i) domain shift, and (ii) OCR errors:
- Exchange Algorithm Clustering [Uszkoreit & Brants' 08]:
  - Cluster 17: *with, wtb, witr, withthe, withits, voith, withl, vyith, wiht, fwith, wdth, witlt, wiTh, wth, wi.h, witih, ...*
  - Cluster 17: *best, beft, ...*

- Accuracy assessment:

	POS Tags		Dependencies	
Domain	base	adapted	base	adapted
Newswire	97.9	97.9	90.1	90.1
Brown	96.8	97.5	84.7	87.1
Questions	94.2	97.5	85.3	91.2
Historical	91.6	93.3	-	-

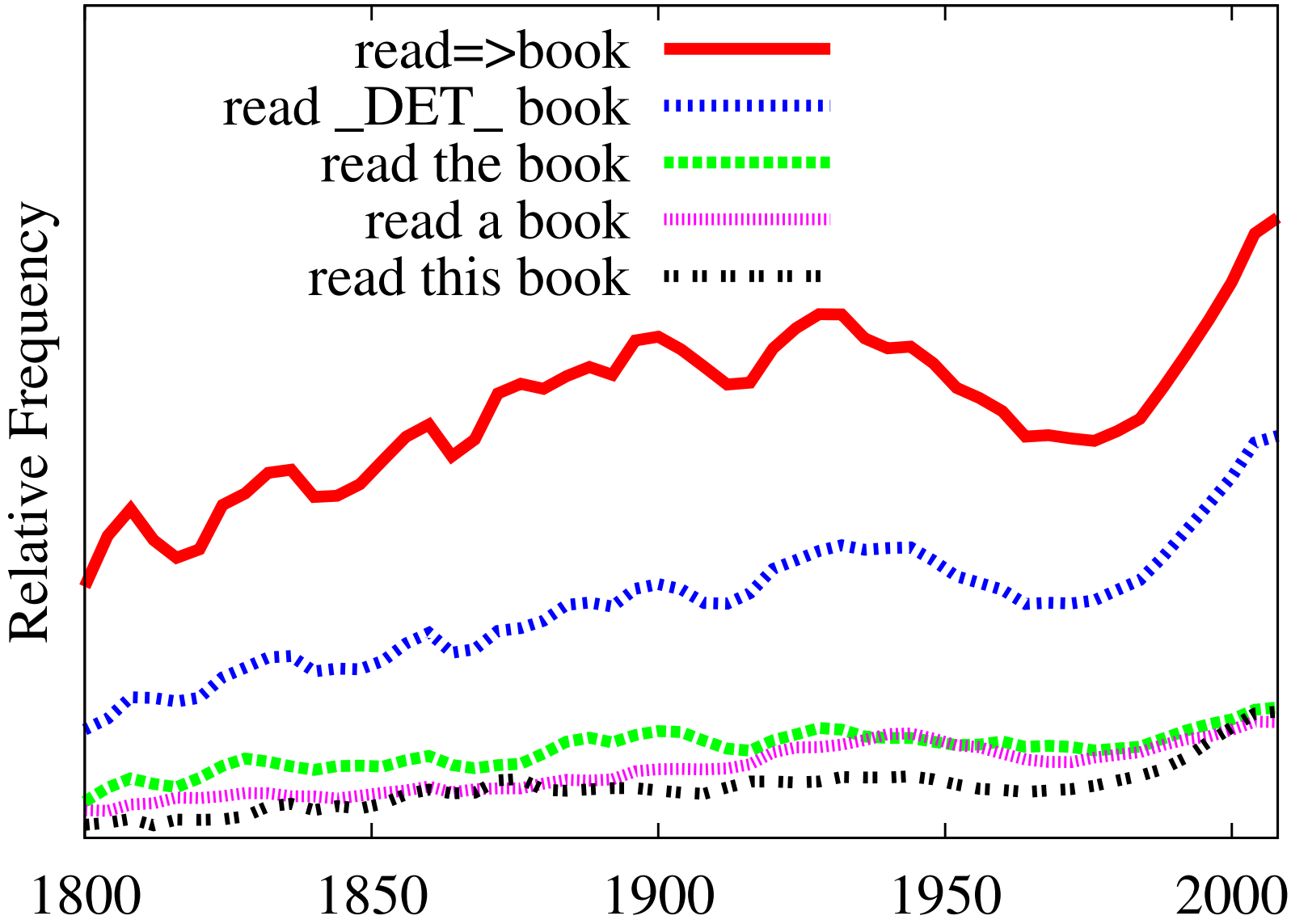
Corpus Statistics

Language	#Volumes	#Tokens
English	4,541,627	468,491,999,592
Spanish	854,649	83,967,471,303
French	792,118	102,174,681,393
German	657,991	64,784,628,286
Russian	591,310	67,137,666,353
Italian	305,763	40,288,810,817
Chinese	302,652	26,859,461,025
Hebrew	70,636	8,172,543,728

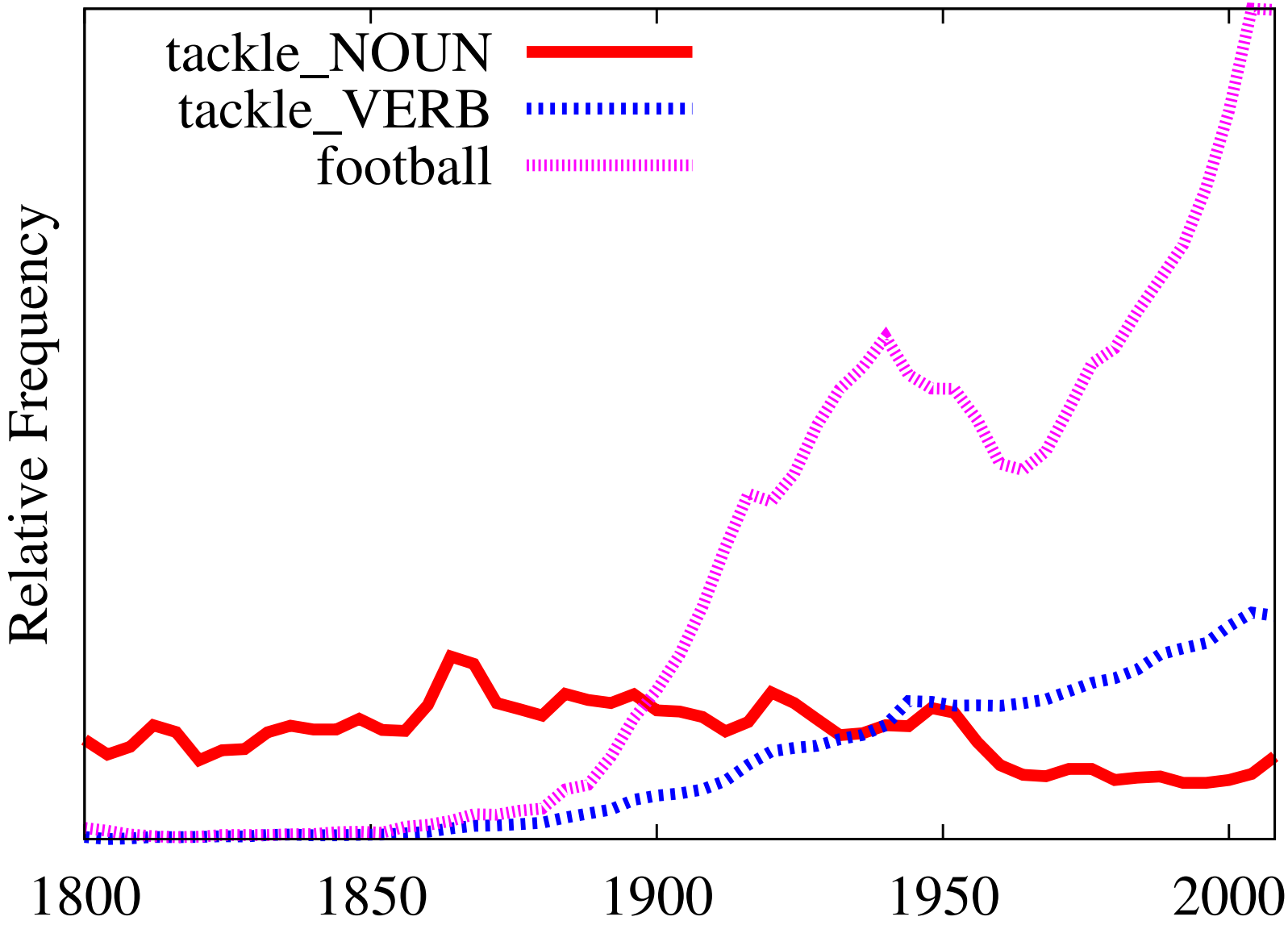
Final Accuracies

Language	POS Tags	Dependencies
English	97.9	90.1
Spanish	96.9	74.5
German	98.8	83.1
French	97.3	84.7
Italian	95.6	80.0
Russian	96.8	86.2
Chinese	92.6	73.2
Hebrew	91.3	76.2

Example Queries



The rise of ‘tackle’ and its relation to football



Is the world getting more quantitative?