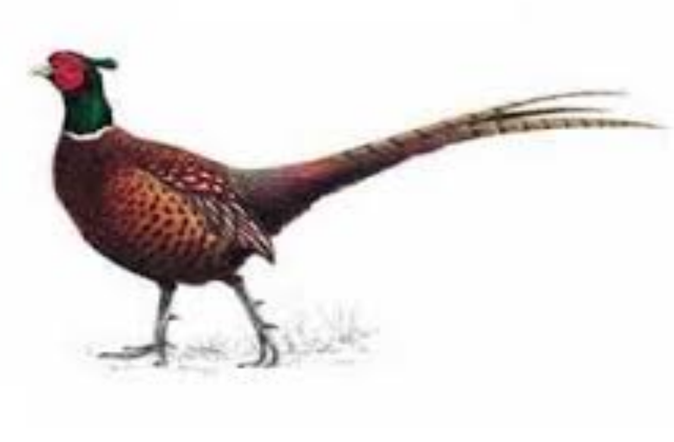
MEANING BANKING
AND THE LONG TAIL

JOHAN BOS
UNIVERSITY OF GRONINGEN



OUTLINE

1. The 80-20 rule
2. An anecdote (verb phrase ellipsis)
3. Meaning Banking (the GMB and the PMB)
4. Inspecting the tail of the GMB
5. The atoms of meaning

THE 80-20 RULE

NLP researchers (including computational linguists) follow the 80-20 rule.

OUTLINE

1. The 80-20 rule
2. An anecdote (verb phrase ellipsis)
3. Meaning Banking (the GMB and the PMB)
4. Inspecting the tail of the GMB
5. The atoms of meaning

VP ELLIPSIS

John$_i$ loves his$_i$ mother. Bill$_j$ does [...] too.

VP ELLIPSIS

John$_i$ loves his$_i$ mother. Bill$_j$ does […] too.

John$_i$ loves his$_i$ mother. Bill$_j$ does [love his$_i$ mother] too. *strict*

John$_i$ loves his$_i$ mother. Bill$_j$ does [love his$_j$ mother] too. *sloppy*

VP ELLIPSIS

John revised his paper before the teacher did […], and Bill did […] too.

Embedded VPE (Dalrymple et al. 1991)

**Mary revised her paper.
Jane did not […], although the teacher did […].**

**Joe first played tennis and then he went out for dinner.
Mark did […] too.**

**An American flag was hanging in front of each window,
and a Canadian flag was […] too.**

VP ELLIPSIS

John revised his paper before the teacher did […],
and Bill did […] too.

Embedded VPE
(Dalrymple et al. 1991)

Mary revised her paper.
Jane did not […], although the teacher did […].

Cascaded VPE

Joe first played tennis and then he went out for dinner.
Mark did […] too.

An American flag was hanging in front of each window,
and a Canadian flag was […] too.

VP ELLIPSIS

John revised his paper before the teacher did […],
and Bill did […] too.

Embedded VPE
(Dalrymple et al. 1991)

Mary revised her paper.
Jane did not […], although the teacher did […].

Cascaded VPE

Joe first played tennis and then he went out for dinner.
Mark did […] too.

Split antecedent VPE
(Prüst 1992, Asher 1993)

An American flag was hanging in front of each window,
and a Canadian flag was […] too.

VP ELLIPSIS

John revised his paper before the teacher did […],
and Bill did […] too.

Embedded VPE
(Dalrymple et al. 1991)

Mary revised her paper.
Jane did not […], although the teacher did […].

Cascaded VPE

Joe first played tennis and then he went out for dinner.
Mark did […] too.

Split antecedent VPE
(Prüst 1992, Asher 1993)

An American flag was hanging in front of each window,
and a Canadian flag was […] too.

VPE with scope ambiguity
(Dalrymple et al. 1991)

VERB PHRASE ELLIPSIS (VPE) IN THE WSJ CORPUS

Bos & Spenader (2011): An annotated corpus for the analysis of VP ellipsis. *Language Resource and Evaluation* 45 (4): 463-494

Corpus size: >1 million words

VPE: 487

Sloppy/strict ambiguity: 9 (all of which were sloppy)

VERB PHRASE ELLIPSIS (VPE) IN THE WSJ CORPUS

Bos & Spenader (2011): An annotated corpus for the analysis of VP ellipsis. *Language Resource and Evaluation* 45 (4): 463-494

Corpus size: **>1 million words**

VPE: **487**

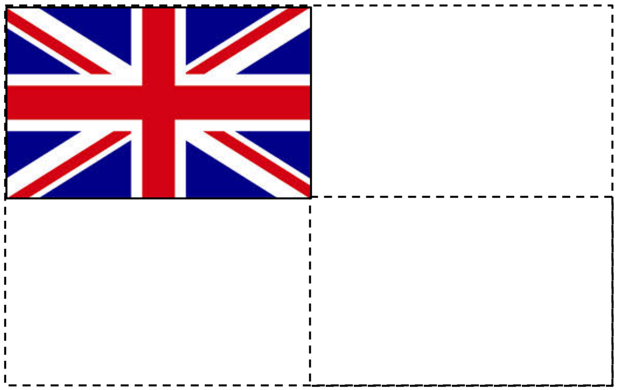
Sloppy/strict ambiguity: **9 (all of which were sloppy)**

No **embedded VPE**
No **cascaded VPE**
No **split antecedents VPE**
No **scope ambiguities with VPE**

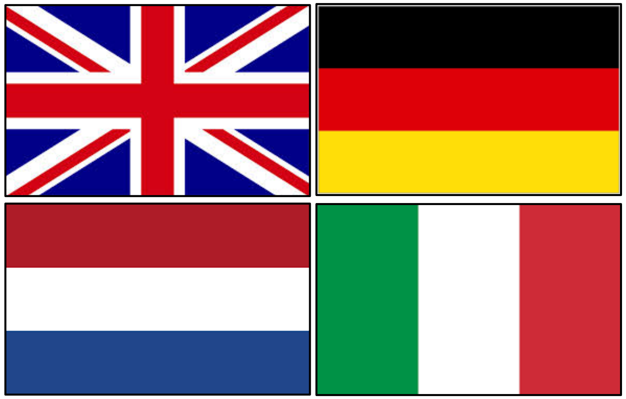
OUTLINE

1. The 80-20 rule
2. An anecdote (verb phrase ellipsis)
3. Meaning Banking (the GMB and the PMB)
4. Inspecting the tail of the GMB
5. The atoms of meaning

MEANING BANKING

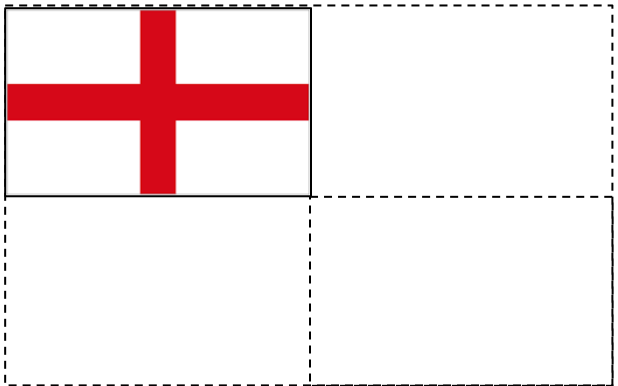


Groningen MEANING BANK

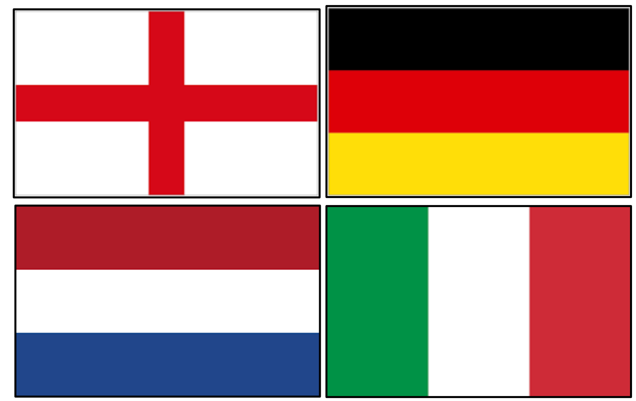


Parallel MEANING BANK

MEANING BANKING



Groningen Meaning Bank



Parallel Meaning Bank

THE PARALLEL MEANING BANK

I don’t see anything.

Ich sehe nichts.

```
 _________________________
| x                       |
 -------------------------
| x = speaker             |
|     __________________  |
|  ¬ | e y              | |
|     ------------------  |
|    | see(e)           | |
|    | experiencer(e,x) | |
|    | theme(e,y)       | |
|    |__________________| |
|_________________________|
```

Non vedo niente.

Ik zie niets.

THE PARALLEL MEANING BANK 11,5M WORD TOKENS

Dutch Parallel Corpus

TATOEBA BETA project

INTERSECT

qtleap

CORPUS

LONWEB VOLUNTEERS program

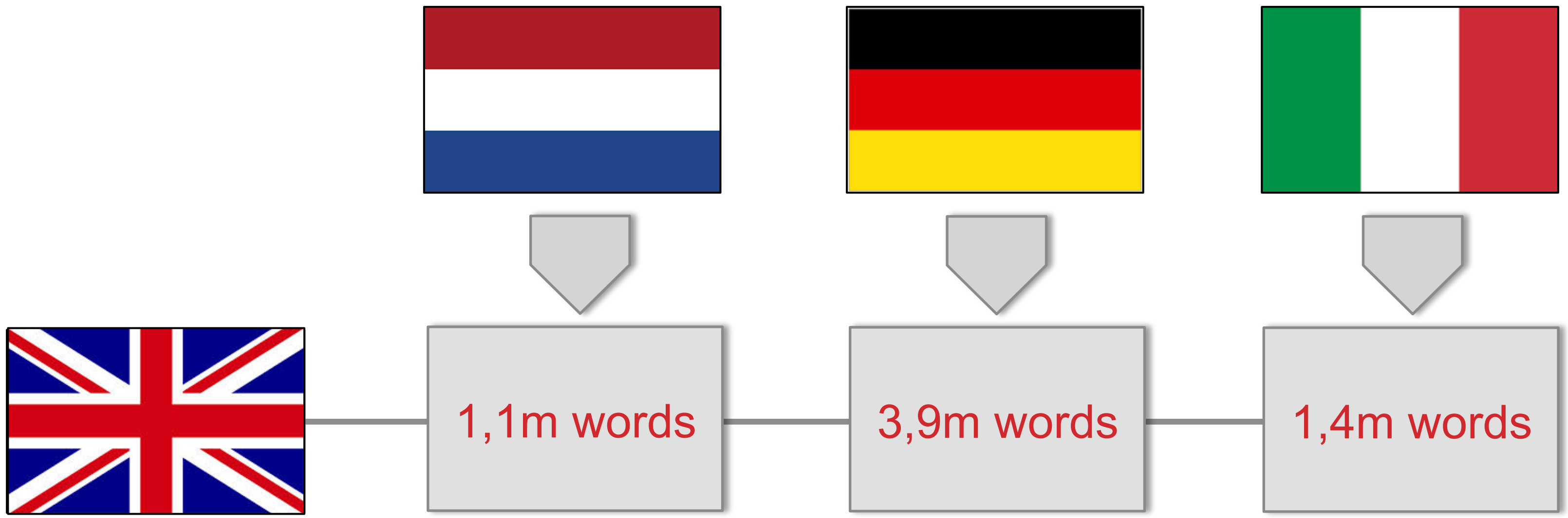
WWW.LONWEB.ORG

QA@CLEF-2004

PASCAL2

Pattern Analysis, Statistical Modelling and Computational Learning

TED



THE PARALLEL MEANING BANK

ENGLISH AS PIVOT LANGUAGE (5 MILLION WORDS)

(CA. 10,000 DOCUMENTS FOR ALL FOUR LANGUAGES)

METHOD

Provide gold standard for about 10% of the corpus

=> crowd-sourcing for common phenomena

=> expert annotators for harder stuff

Produce silver standard for the rest.

=> automatically generated with models learned from gold standard

=> hand-corrected for selected phenomena

BIN/BOXER

```
Mr. Johnson was travelling to San Franacie Bay. He went to New York and he smoked.

 ______________________________________________________________________________
|                                                                              |
|      ______________________________      ______________________      ____________   |
|     |x1 e1 x2                      |    |x1 e2 x3              |    |x1 e3      |   |
|  k1:|..............................| k2:|......................| k3:|...........|   |
|     |named(x1,mr.~johnson,per)     |    |male(x1)              |    |male(x1)   |   |
|     |travel(e1)                    |    |go(e2)                |    |smoke(e3)  |   |
|     |agent(e1,x1)                  |    |agent(e2,x1)          |    |agent(e3,x1)|  |
|     |named(x2,san~franacie~bay,geo)|    |named(x3,new~york,geo)|    |___________|   |
|     |to(e1,x2)                     |    |to(e2,x3)             |                    |
|     |______________________________|    |______________________|                    |
|______________________________________________________________________________|
|                                                                              |
|continuation(k1,k2)                                                           |
|continuation(k2,k3)                                                           |
|parallel(k2,k3)                                                               |
|______________________________________________________________________________|
```

OUTLINE

1. The 80-20 rule
2. An anecdote (verb phrase ellipsis)
3. Meaning Banking (the GMB and the PMB)
4. Inspecting the tail of the GMB
5. The atoms of meaning

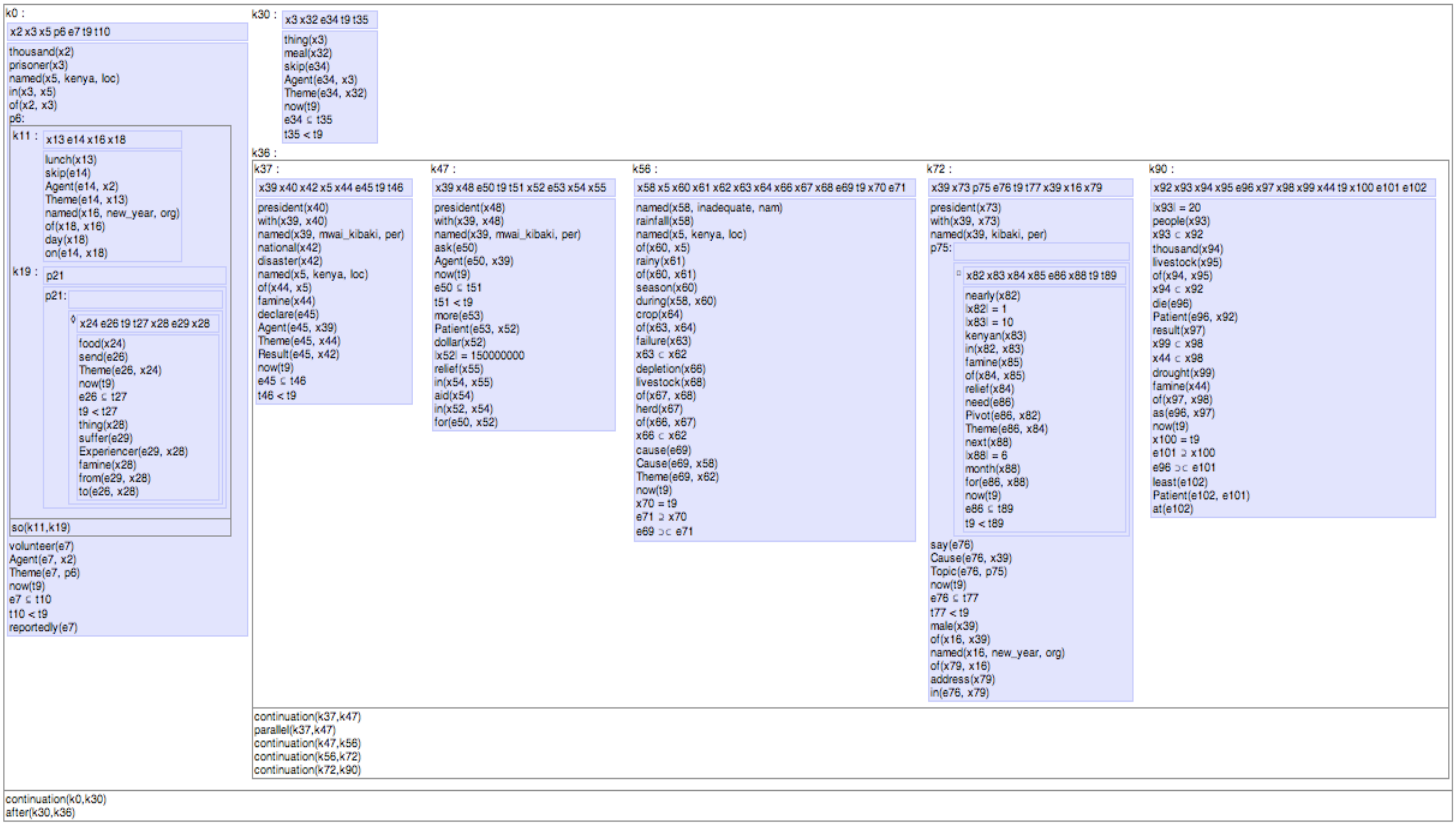
THE GRONINGEN MEANING BANK

Large (English) corpus of public domain texts

Annotated with meaning representations

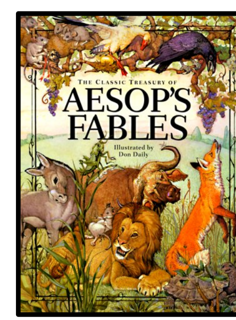
- generated by Boxer (semantic parser)
- corrected by humans (experts and “the crowd”)

10,000 MEANING REPRESENTATIONS



GRONINGEN MEANING BANK: CORPUS SIZE

	genre	texts	sentences	words	s/t	w/s
Voice of America	newswire	9,207	57,174	1,238,576	6.2	21.7
CIA world factbook	almanac	514	4,436	112,516	8.6	25.4
Aesop’s Fables	narrative	224	949	23,105	4.2	24.3
jokes	humor	122	443	7,531	3.6	17.0
MASC		35	291	6,985	8.3	24.0
RTE		1,338	1,537	29,854	1.1	19.4
		11,440	**64,830**	**1,418,567**	**5.7**	**21.9**



WORDS IN THE GMB
TAIL = TOKENS THAT OCCUR ONCE

Tokens	Types	Head	Tail	
1,982	840	266	574	
13,718	3,396	1425	1,971	
142,344	13,011	6,980	6,031	
1,354,149	39,423	23,170	16,253	

Groningen
MEANING
BANK

WORDS IN THE GMB
TAIL = TOKENS THAT OCCUR ONCE

Tokens	Types	Head	Tail	
1,982	840	266	574	68%
13,718	3,396	1425	1,971	58%
142,344	13,011	6,980	6,031	46%
1,354,149	39,423	23,170	16,253	41%

CHARACTERS IN THE GMB

TAIL = TOKENS THAT OCCUR ONCE

Tokens	Types	Head	Tail	
844	49	37	12	
11,355	68	65	3	
77,713	81	76	5	
810,481	86	82	4	
7,711,817	228	202	26	

CHARACTERS IN THE GMB
TAIL = TOKENS THAT OCCUR ONCE

Tokens	Types	Head	Tail	
844	49	37	12	32%
11,355	68	65	3	4%
77,713	81	76	5	6%
810,481	86	82	4	5%
7,711,817	228	202	26	11%

Groningen MEANING BANK

CAUGHT BY THE TAIL

OUTLINE

1. The 80-20 rule
2. An anecdote (verb phrase ellipsis)
3. Meaning Banking (the GMB and the PMB)
4. Inspecting the tail of the GMB
5. The atoms of meaning

ATOMS OF MEANING

Sentences have meaning.
This meaning has to come from somewhere.

In mainstream NLP, usually <u>words</u> are taken as the smallest grammatical units.

But words are not the atoms of meaning. Morphemes are.

MORPHEMES

Consider: ***unhappiness*****.**

This word does not occur in the GMB. Shocking!

But its morphemes do:

un-	**3,990 times**
happy	**24 times**
ness	**980 times**

WORD EMBEDDINGS AND MORE

Word embedding models promising (each word is associated with a vector)

Cao & Rei (2016):

- **present a model that learns morphology and word embeddings jointly**
- **Character-level models can predict good quality representations for unseen words**

CONCLUSIONS

- Rare phenomena … are very common!
- Discrepancy between frequency of semantic phenomena in theory (fantasy corpus) and practice (real world corpus)
- Meaning bank suffers (obviously) from the long-tail-problem
- Modelling morphemes rather than words might deal with (part of) the problem