

# 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<sub>i</sub> loves his<sub>i</sub> mother. Bill<sub>j</sub> does [...] too.



# VP ELLIPSIS

John<sub>i</sub> loves his<sub>i</sub> mother. Bill<sub>j</sub> does [...] too.

John<sub>i</sub> loves his<sub>i</sub> mother. Bill<sub>j</sub> does [love his<sub>i</sub> mother] too. *strict*

John<sub>i</sub> loves his<sub>i</sub> mother. Bill<sub>j</sub> does [love his<sub>j</sub> 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**

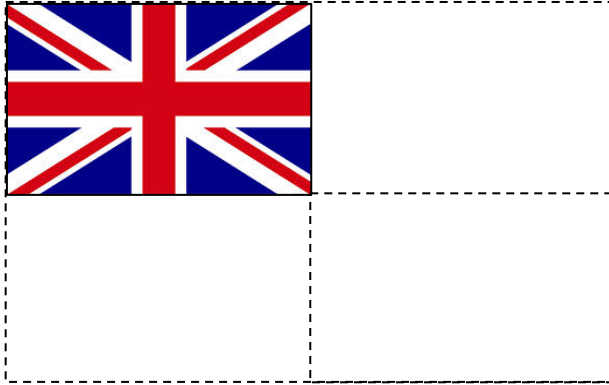


# OUTLINE

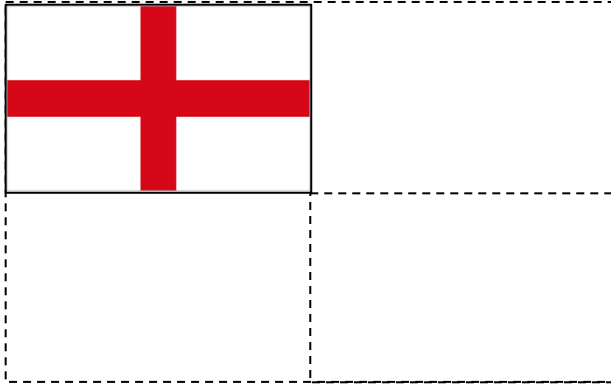
1. The 80-20 rule
2. An anecdote (verb phrase ellipsis)
3. Meaning Banking (the GMB and the PMB)
4. Inspecting the tail of the GMB
5. The atoms of meaning



# MEANING BANKING



# MEANING BANKING

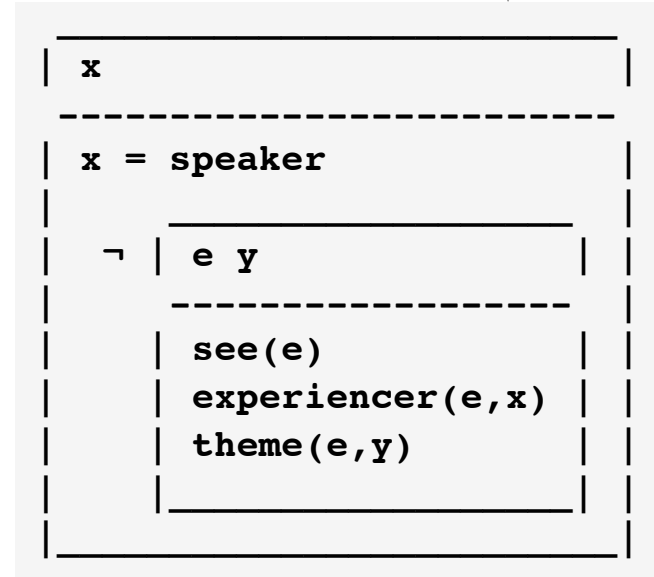


# THE PARALLEL MEANING BANK

I don't see anything.

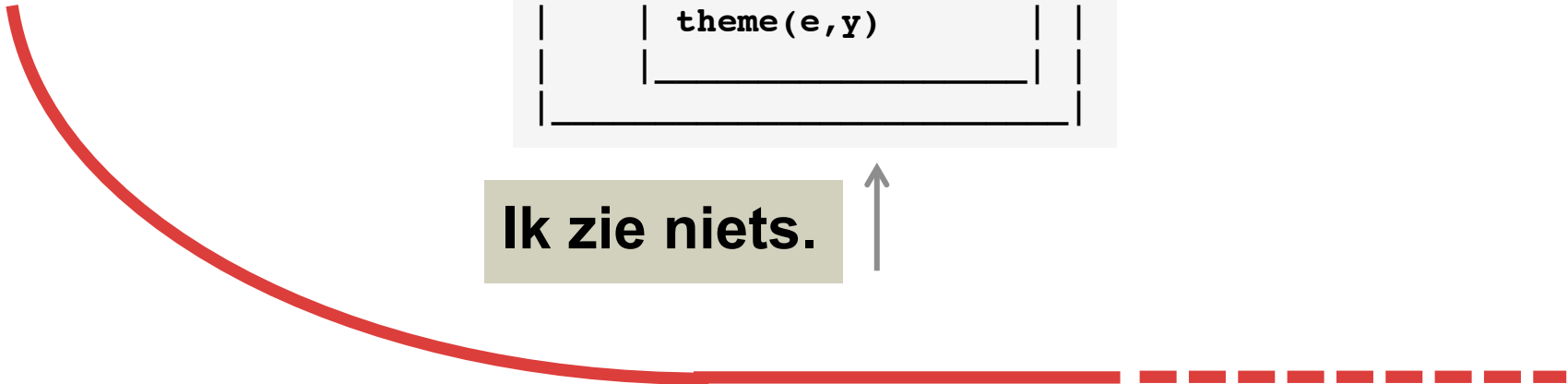


Ich sehe nichts.



Non vedo niente.

Ik zie niets.





# THE PARALLEL MEANING BANK

## 11,5M WORD TOKENS



INTERSECT

qt leap

CRPUS



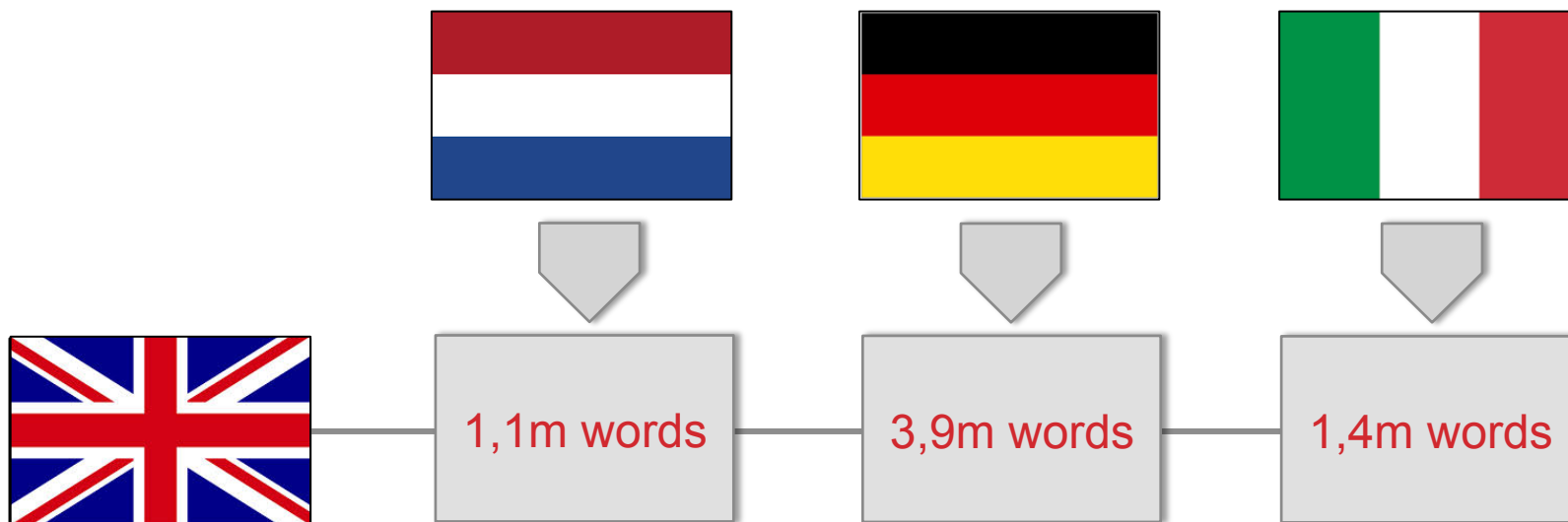
LONWEB  
VOLUNTEERS  
program

WWW.LONWEB.ORG

QA@CLEF-2004



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**

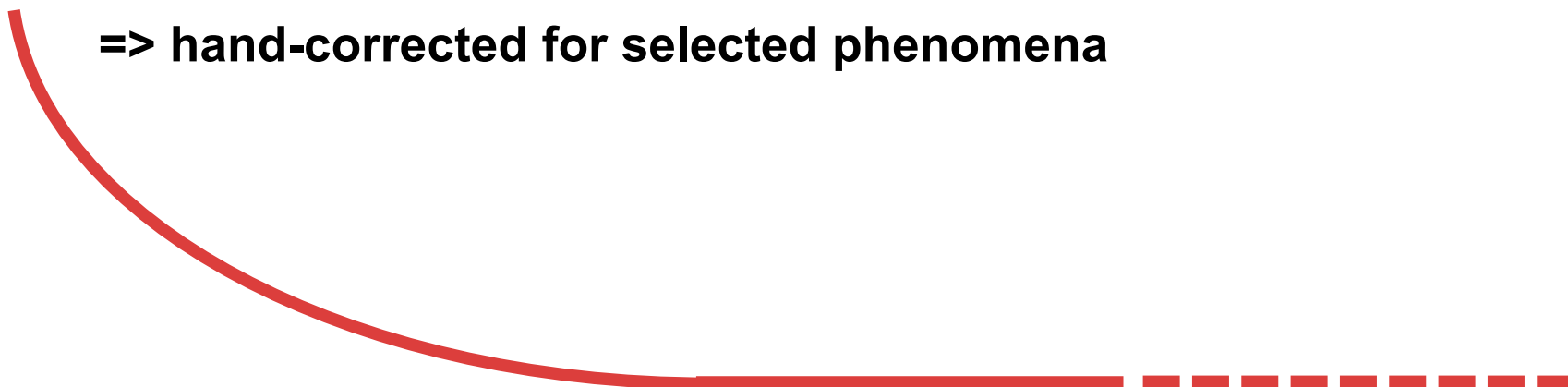
**=> crowd-sourcing for common phenomena**

**=> 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.

<u>x1 e1 x2</u>	<u>x1 e2 x3</u>	<u>x1 e3</u>
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

k0 :

```

x2 x3 x5 p6 e7 19 110
thousand(x2)
prisoner(x3)
named(x5, kenya, loc)
in(x3, x5)
of(x2, x3)
p6:
  k11 :
    x13 e14 x16 x18
    lunch(x13)
    skip(e14)
    Agent(e14, x2)
    Theme(e14, x13)
    named(x16, new_year, org)
    of(x18, x16)
    day(x18)
    on(e14, x18)
  k19 :
    p21
    p21:
      x24 e26 19 127 x28 e29 x28
      food(x24)
      send(e26)
      Theme(e26, x24)
      now(19)
      e26 < 127
      19 < 127
      thing(x28)
      suffer(e29)
      Experiencer(e29, x28)
      famine(x28)
      from(e29, x28)
      to(e26, x28)
    so(k11, k19)
  volunteer(e7)
  Agent(e7, x2)
  Theme(e7, p6)
  now(19)
  e7 < 110
  110 < 19
  reportedly(e7)

```

k30 :

```

x3 x32 e34 19 135
thing(x3)
meal(x32)
skip(e34)
Agent(e34, x3)
Theme(e34, x32)
now(19)
e34 < 135
135 < 19

```

k36 :

```

x39 x40 x42 x5 x44 e45 19 146
president(x40)
with(x39, x40)
named(x39, mwai_kibaki, per)
national(x42)
disaster(x42)
named(x5, kenya, loc)
of(x44, x5)
famine(x44)
declare(e45)
Agent(e45, x39)
Theme(e45, x44)
Result(e45, x42)
now(19)
e45 < 146
146 < 19

```

k47 :

```

x39 x48 e50 19 151 x52 e53 x54 x55
president(x48)
with(x39, x48)
named(x39, mwai_kibaki, per)
ask(e50)
Agent(e50, x39)
now(19)
e50 < 151
151 < 19
more(e53)
Patient(e53, x52)
dollar(x52)
|x52 = 150000000
relief(x55)
in(x54, x55)
aid(x54)
in(x52, x54)
for(e50, x52)

```

k56 :

```

x58 x5 x60 x61 x62 x63 x64 x66 x67 x68 e69 19 x70 e71
named(x58, inadequate, nam)
rainfall(x58)
named(x5, kenya, loc)
of(x60, x5)
rainy(x61)
of(x60, x61)
season(x60)
during(x58, x60)
crop(x64)
of(x63, x64)
failure(x63)
x63 < x62
depletion(x66)
livestock(x68)
of(x67, x68)
herd(x67)
of(x66, x67)
x66 < x62
cause(e69)
Cause(e69, x58)
Theme(e69, x62)
now(19)
x70 = 19
e71 > x70
e69 > c e71

```

k72 :

```

x39 x73 p75 e76 19 177 x39 x16 x79
president(x73)
with(x39, x73)
named(x39, kibaki, per)
p75:
  x82 x83 x84 x85 e86 x88 19 189
  nearly(x82)
  |x82 = 1
  |x83 = 10
  Kenyan(x83)
  in(x82, x83)
  famine(x85)
  of(x84, x85)
  relief(x84)
  need(e86)
  Pivot(e86, x82)
  Theme(e86, x84)
  next(x88)
  |x88 = 6
  month(x88)
  for(e86, x88)
  now(19)
  e86 < 189
  19 < 189
  say(e76)
  Cause(e76, x39)
  Topic(e76, p75)
  now(19)
  e76 < 177
  177 < 19
  male(x39)
  of(x16, x39)
  named(x16, new_year, org)
  of(x79, x16)
  address(x79)
  in(e76, x79)

```

k90 :

```

x92 x93 x94 x95 e96 x97 x98 x99 x44 19 x100 e101 e102
|x93 = 20
people(x93)
x93 < x92
thousand(x94)
livestock(x95)
of(x94, x95)
x94 < x92
die(e96)
Patient(e96, x92)
result(x97)
x99 < x98
x44 < x98
drought(x99)
famine(x44)
of(x97, x98)
as(e96, x97)
now(19)
x100 > 19
e101 > x100
e96 > c e101
least(e102)
Patient(e102, e101)
at(e102)

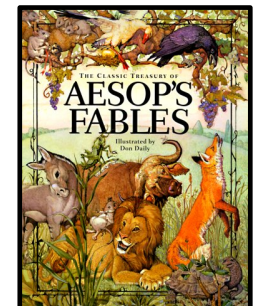
```

continuation(k37, k47)  
parallel(k37, k47)  
continuation(k47, k56)  
continuation(k56, k72)  
continuation(k72, k90)

continuation(k0, k30)  
after(k30, k36)

# 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
		<b>11,440</b>	<b>64,830</b>	<b>1,418,567</b>	<b>5.7</b>	<b>21.9</b>





# 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	



# 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%



# 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 words are taken as  
the smallest grammatical units.**

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



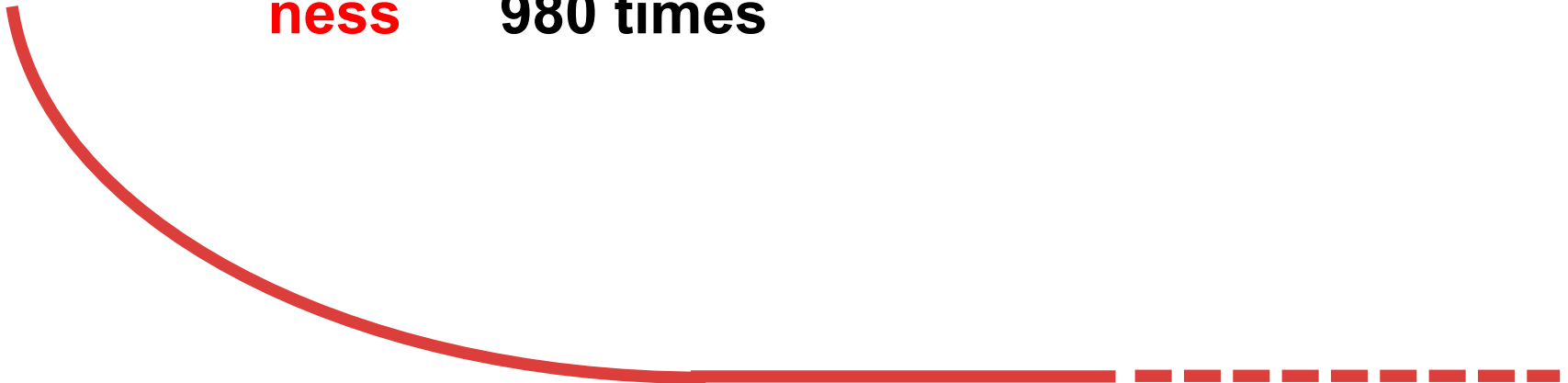
# MORPHEMES

Consider: *unhappiness*.

This word does not occur in the GMB. Shocking!

But its morphemes do:

<b>un-</b>	<b>3,990 times</b>
<b>happy</b>	<b>24 times</b>
<b>ness</b>	<b>980 times</b>





# 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**

