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.  \textit{strict}

John$_i$ loves his$_i$ mother. Bill$_j$ does [love his$_j$ mother] too. \textit{sloppy}
VP ELLIPSIS

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

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

Embedded VPE
(Dalrymple et al. 1991)

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.

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. 

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.

*Embedded VPE*  
(Dalrymple et al. 1991)

*Cascaded VPE*

*Split antecedent VPE*  
(Prüst 1992, Asher 1993)
VP ELLIPSIS

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

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

Joe first played tennis and then he went out for dinner.

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

Embedded VPE
(Dalrymple et al. 1991)

Cascaded VPE

Split antecedent VPE
(Prüst 1992, Asher 1993)

VPE with scope ambiguity
(Dalrymple et al. 1991)
VERB PHRASE ELLIPSIS (VPE) IN THE WSJ CORPUS


Corpus size: >1 million words

VPE: 487

Sloppy/strict ambiguity: 9 (all of which were sloppy)
VERB PHRASE ELLIPSIS (VPE) IN THE WSJ CORPUS


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

[Flag images of the United Kingdom and Germany, and further flag images indicating the banking context in different countries, with logos for Groningen MEANING BANK and Parallel MEANING BANK.]
<table>
<thead>
<tr>
<th>Country</th>
<th>Flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>🇬🇧</td>
</tr>
<tr>
<td>Netherlands</td>
<td>🇳🇱</td>
</tr>
<tr>
<td>Germany</td>
<td>🇩🇪</td>
</tr>
<tr>
<td>Italy</td>
<td>🇮🇹</td>
</tr>
</tbody>
</table>
I don’t see anything.

Ich sehe nichts.

Non vedo niente.

Ik zie niets.
THE PARALLEL MEANING BANK
11,5M WORD TOKENS
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
Mr. Johnson was travelling to San Franacie Bay. He went to New York and he smoked.

```plaintext
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”)
# GRONINGEN MEANING BANK: CORPUS SIZE

<table>
<thead>
<tr>
<th>genre</th>
<th>texts</th>
<th>sentences</th>
<th>words</th>
<th>s/t</th>
<th>w/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice of America</td>
<td>newswire</td>
<td>9,207</td>
<td>57,174</td>
<td>1,238,576</td>
<td>6.2</td>
</tr>
<tr>
<td>CIA world factbook</td>
<td>almanac</td>
<td>514</td>
<td>4,436</td>
<td>112,516</td>
<td>8.6</td>
</tr>
<tr>
<td>Aesop’s Fables</td>
<td>narrative</td>
<td>224</td>
<td>949</td>
<td>23,105</td>
<td>4.2</td>
</tr>
<tr>
<td>jokes</td>
<td>humor</td>
<td>122</td>
<td>443</td>
<td>7,531</td>
<td>3.6</td>
</tr>
<tr>
<td>MASC</td>
<td></td>
<td>35</td>
<td>291</td>
<td>6,985</td>
<td>8.3</td>
</tr>
<tr>
<td>RTE</td>
<td></td>
<td>1,338</td>
<td>1,537</td>
<td>29,854</td>
<td>1.1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>11,440</td>
<td>64,830</td>
<td>1,418,567</td>
<td>5.7</td>
</tr>
</tbody>
</table>
WORDS IN THE GMB TAIL = TOKENS THAT OCCUR ONCE

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Types</th>
<th>Head</th>
<th>Tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,982</td>
<td>840</td>
<td>266</td>
<td>574</td>
</tr>
<tr>
<td>13,718</td>
<td>3,396</td>
<td>1425</td>
<td>1,971</td>
</tr>
<tr>
<td>142,344</td>
<td>13,011</td>
<td>6,980</td>
<td>6,031</td>
</tr>
<tr>
<td>1,354,149</td>
<td>39,423</td>
<td>23,170</td>
<td>16,253</td>
</tr>
</tbody>
</table>
WORDS IN THE GMB
TAIL = TOKENS THAT OCCUR ONCE

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Types</th>
<th>Head</th>
<th>Tail</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,982</td>
<td>840</td>
<td>266</td>
<td>574</td>
<td>68%</td>
</tr>
<tr>
<td>13,718</td>
<td>3,396</td>
<td>1425</td>
<td>1,971</td>
<td>58%</td>
</tr>
<tr>
<td>142,344</td>
<td>13,011</td>
<td>6,980</td>
<td>6,031</td>
<td>46%</td>
</tr>
<tr>
<td>1,354,149</td>
<td>39,423</td>
<td>23,170</td>
<td>16,253</td>
<td>41%</td>
</tr>
</tbody>
</table>
**CHARACTERS IN THE GMB**
**TAIL = TOKENS THAT OCCUR ONCE**

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Types</th>
<th>Head</th>
<th>Tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>844</td>
<td>49</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>11,355</td>
<td>68</td>
<td>65</td>
<td>3</td>
</tr>
<tr>
<td>77,713</td>
<td>81</td>
<td>76</td>
<td>5</td>
</tr>
<tr>
<td>810,481</td>
<td>86</td>
<td>82</td>
<td>4</td>
</tr>
<tr>
<td>7,711,817</td>
<td>228</td>
<td>202</td>
<td>26</td>
</tr>
</tbody>
</table>
### Characters in the GMB
**Tail = Tokens that occur once**

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Types</th>
<th>Head</th>
<th>Tail</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>844</td>
<td>49</td>
<td>37</td>
<td>12</td>
<td>32%</td>
</tr>
<tr>
<td>11,355</td>
<td>68</td>
<td>65</td>
<td>3</td>
<td>4%</td>
</tr>
<tr>
<td>77,713</td>
<td>81</td>
<td>76</td>
<td>5</td>
<td>6%</td>
</tr>
<tr>
<td>810,481</td>
<td>86</td>
<td>82</td>
<td>4</td>
<td>5%</td>
</tr>
<tr>
<td>7,711,817</td>
<td>228</td>
<td>202</td>
<td>26</td>
<td>11%</td>
</tr>
</tbody>
</table>
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
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:

- 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