

## Principles

- Take the designer's perspective
- Introduce some corpora
- Describe them
- Aim for critical appreciation

## Reference or monitor?

- Brown or Collins Bank of English
- Experiment or lexicography (or diachronic variation)
- AP Newswire somewhere in between

## Copyright

- Can't ignore it
- Tension between science and commerce

## Getting the corpus

- Sample frame
- Stratified
- Solicited
- Generated under (semi-) controlled conditions
- By-product of paper publication
- Test-suites
- Opportunistic
- Externally dictated. Often commercial and/or secret

## Sample frame

- Manageable subset of holdings of (say) EU library
- Maybe stratified into classes like belles lettres, legal cowboy fiction (e.g. Brown and its successors)

## Solicited

- Henry Thompson posted a couple of French texts to the Net.
- Got back 40 or so self-selected English translations of same text for comparison
- Potential weakness: are they representative?

## Generated under (semi-)controlled conditions

- Undirected (tell me what you did at the weekend, Johnny)
- Wizard of Oz (e.g. Andernach's fake Reservation Agent)
- Production experiments (tell the story you see)
- Map Task (relatively natural)
- Translations commissioned from Heriot-Watt students (compulsion!?)

## By-products

- Newspapers (have editorial control)
- Typesetting tapes of dictionaries (better now that publishers have computer systems)
- Usenet news groups (too nerdy?)

## Commercial or secret corpora

- We have had projects to use machine learning to help indexers add keywords to technical abstracts.
- Governments and their agencies are serious consumers of information retrieval and indexing technology

## Planning the science

- Have a scientific question in mind
- Borrow somebody else's corpus design
- Sketch an experiment. Eg. How much text needed to get 100 relative clauses?

## Planning the science: Map Task

- Instruction Giver/Instruction Follower
- Familiarity
- Eye-Contact
- Order of presentation of landmarks
- Phonetic reduction (planned for)
- Not enough clefts

## Planning the science: be lazy

- DCIEM: Map task with drugged-up Canadian soldiers
- Japanese Map Task

## Does size matter?

- Yes. The larger the better
- It might be a pain if it's too big
- You can't expect to post-edit 100,000,000 words of BNC

## Annotation

- Choose what to annotate
- Choose external form for annotations
- Almost all annotation is opinion. Even spelling “errors”.
  - Use explicit guidelines
  - Use multiple annotators
  - Specify how to resolve conflicts
- Keep everything. Never change input text
- London-Lund sound recordings

## No Markup: Leverhulme Corpus of Essays

C6. Whilst physical ailments do play a major part in the area of prematurity, it is only one of the factors which affects the child, the parents and the infant - caregiver interactions. Whilst the problems caused by prematurity are not insurmountable, direct immediate steps have to be taken so that the effects of prematurity do not continue into later life. A premature baby is one that is defined as having been born before 37 weeks of gestation and weighing less than 2.5 kg. The low weight is a problem because the infant is less able to maintain homeostasis and thus needs to be kept in an incubator. The infant is also relatively physiologically immature as myelination of the neurons has not yet fully taken place. The lack of weight means its appearance is somewhat scrawly and if born at less than 32 weeks of gestational age, will look withered. This can cause problems in the infant - caretaker interaction and the caretaker's attitudes as will be discussed later. Physiological immaturity means there is a general lack of bodily control, the infant tends to lie sprawled in a 'frog-like' manner and when it does move the movements are often sudden + jerky. The child is more prone to physiological and respiratory problems, jaundice, digestive illnesses and has a weaker immune system. State regulation is less well controlled + more difficult to define, especially if the infant spends more of its time asleep.

Who typed this? Did the student? If the typist, who made the errors?

## Light Markup: ECI Corpus: Dickens

```
<!-- ECI Corpus File:  Three English Novels (eng01)          -->
<!--      Component:  A Christmas Carol (a)                -->
<!-- Copyright 1994 ACL.                                   -->
<!-- Material as supplied to ECI Public Domain.           -->
<div0 type=file id=eng01.eci n=xmas.tei>
<omit desc="front matter" resp=eci>
<head>
A CHRISTMAS CAROL.
</head>

<div1 type=stave id=s1>
<head>
MARLEY'S GHOST
</head>
<pb n=1>
<p>
Marley was dead: to begin with. There is no doubt whatever
about that. The register of his burial was signed by the
clergyman, the clerk, the undertaker, and the chief mourner.
Scrooge signed it. And Scrooge's name was good upon 'Change,
for anything he chose to put his hand to. Old Marley was as
dead as a door-nail.
</p>
...
```

Note header.

## Dense Markup: BNC

```

</W><W TYPE=AVP>on </W><W TYPE=PRP>with </W>
<W TYPE=ATO>the </W> <W TYPE=ORD>second </W><W TYPE=NNO>half </W>
<W TYPE=PRF>of </W><W TYPE=ATO>the</W><W TYPE=NN1>meeting</W>
<C TYPE=PUN>.</C></S>
<S N=008> <W TYPE=AVO>So </W><W TYPE=NPO>Brenda</W><C TYPE=PUN>.</C></S>
</U>
<U WHO=D90PS002>
<S N=009> <W TYPE=VVB>Thank </W><W TYPE=PNP>you</W><C TYPE=PUN>.</C></S>
<S N=010><PAUSE> <W TYPE=AVO>Well </W><W TYPE=DTO>some </W>
<W TYPE=PRF>of </W><W TYPE=PNP>you </W><W TYPE=VHB>have </W>
<W TYPE=VVN>brought</W>
<W TYPE=DTO>some </W><W TYPE=UNC>erm </W><W TYPE=AJO>interesting</W>
<W TYPE=NN2>items </W><W TYPE=AVP>along</W><C TYPE=PUN>.</C></S>
<S N=011> <W TYPE=VHB>Have</W><W TYPE=XXO>n't </W><W TYPE=VHN>had </W>
<W TYPE=NN1>time </W><W TYPE=TOO>to </W><W TYPE=VVI>look </W><W TYPE=PRP>at
</W><W TYPE=PNP>them </W><W TYPE=DTO>all</W><C TYPE=PUN>.</C></S>
<S N=012> <W TYPE=UNC>erm </W><W TYPE=PNP>I</W><W TYPE=VBB>'m </W>
<W TYPE=XXO>not </W><W TYPE=VVG>going </W><W TYPE=TOO>to </W>
<W TYPE=VVI>keep </W><W TYPE=PNP>you
</W><W TYPE=AVO>very </W><W TYPE=AVO>long </W><W TYPE=CJS>because</W>
<W TYPE=PNP>I</W><W TYPE=VHB>'ve </W><W TYPE=AVO>nearly
</W><W TYPE=VVN>finished </W><W TYPE=VVG>talking </W>
<W TYPE=AVO>so </W><W TYPE=UNC>erm </W><W TYPE=AVQ-CJS>when
</W><W TYPE=PNP>I</W><W TYPE=VHB>'ve </W><W TYPE=VVN>finished </W>
<W TYPE=AVO>perhaps </W><W TYPE=PNP>you </W><W TYPE=VMO>woul

```

Actually not that dense, but looks it.

## Tagged Brown Corpus from Treebank

=====

```
[ The/DT Fulton/NNP County/NNP Grand/NNP Jury/NNP ]  
said/VBD  
[ Friday/NNP ]  
  
[ an/DT investigation/NN ]  
of/IN  
[ Atlanta/NNP 's/POS recent/JJ primary/JJ election/NN ]  
produced/VBD ''/''  
[ no/DT evidence/NN ]  
''/'' that/IN  
[ any/DT irregularities/NNS ]  
took/VBD  
[ place/NN ]  
./.
```

=====

Can annotators agree on parts of speech?

## Susanne

A01:0010a	-	YB	<minbrk>	-	[Oh.Oh]
A01:0010b	-	AT	The the	[O[S[Nns:s.	
A01:0010c	-	NP1s	Fulton Fulton	[Nns.	
A01:0010d	-	NN1cb	County county	.Nns]	
A01:0010e	-	JJ	Grand grand	.	
A01:0010f	-	NN1c	Jury jury	.Nns:s]	
A01:0010g	-	VVDv	said say	[Vd.Vd]	
A01:0010h	-	NPD1	Friday Friday	[Nns:t.Nns:t]	
A01:0010i	-	AT1	an an	[Fn:o[Ns:s.	
A01:0010j	-	NN1n	investigation investigation	.	
A01:0020a	-	IO	of of	[Po.	
A01:0020b	-	NP1t	Atlanta Atlanta	[Ns[G[Nns.Nns]	
A01:0020c	-	GG	+<apos>s	- .G]	
A01:0020d	-	JJ	recent recent	.	
A01:0020e	-	JJ	primary primary	.	
A01:0020f	-	NN1n	election election	.Ns]Po]Ns:s]	
A01:0020g	-	VVDv	produced produce	[Vd.Vd]	
A01:0020h	-	YIL	<ldquo> -	.	
A01:0020i	-	ATn	+no no	[Ns:o.	
A01:0020j	-	NN1u	evidence evidence	.	

Notation!

## Penn Treebank 1

```
( (S
  (NP The Fulton County Grand Jury)
  (VP said
    (NP Friday)
    (SBAR 0
      (S
        (NP an investigation
          (PP of
            (NP
              (NP Atlanta)
              's
              recent primary
              election)))
          (VP produced
            (NP ‘‘
              no evidence
              ’’
              (SBAR that
                (S (NP any irregularities)
                  (VP took
                    (NP place))))))))))
  .)
```

Need to be able to be vague when constituency unclear.

## Semcor

```

<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done rdf=group pos=NNP lemma=group wnsn=1 lexs=1:03:00::
pn=group>
Fulton_County_Grand_Jury</wf>
<wf cmd=done pos=VB lemma=say wnsn=1 lexs=2:32:00::>said</wf>
<wf cmd=done pos=NN lemma=friday wnsn=1 lexs=1:28:00::>Friday</wf>
<wf cmd=ignore pos=DT>an</wf>
<wf cmd=done pos=NN lemma=investigation wnsn=1 lexs=1:09:00::>
investigation</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=NN lemma=atlanta wnsn=1 lexs=1:15:00::>Atlanta</wf>
<wf cmd=ignore pos=POS>'s</wf>
<wf cmd=done pos=JJ lemma=recent wnsn=2 lexs=5:00:00:late:00>recent</wf>
<wf cmd=done pos=NN lemma=primary_election wnsn=1 lexs=1:04:00::>
primary_election</wf>
<wf cmd=done pos=VB lemma=produce wnsn=6 lexs=2:39:01::>produced</wf>
<punc>'</punc>
<wf cmd=ignore pos=DT>no</wf>
<wf cmd=done pos=NN lemma=evidence wnsn=1 lexs=1:09:00::>evidence</wf>
<punc>'</punc>

```

Word sense might not attach to words only.

## SGML

- Good because:
  - International Standard
  - Good tools
  - Convenient for programs (esp. `nsgml`)
  - Widely used in real text processing
  - Sophisticated about multilinguality
- Thought bad because:
  - verbose (both language and documentation)
  - complex (both language and documentation)
- Solution:
  - Good visualization tools for humans
  - Pluggable pipelines for software

## Ad-hoc corpora

Manchester Guardian, Feb. 2002, processed by Charniak's parser.

```
(S1 (S (NP (NP (NP (NNP Britain) (POS 's))
  (JJ first)
  (JJ known)
  (NN case))
  (PP (IN of)
    (NP (NP (JJ common) (NN assault))
      (PP (IN by) (NP (NN iguana-throwing))))))
  (VP (VBD went)
    (PP (TO to) (NP (NN court)))
    (NP (NN yesterday)
      (, ,)
      (PP (IN with)
        (NP (DT the) (JJ alleged) (JJ offensive) (NN weapon)))
      (S (NP (PRP himself))
        (VP (VBG watching)
          (ADVP (RB beadily))
          (PP (IN from)
            (NP (NP (DT a) (NN tank))
              (PP (IN by) (NP (DT the) (NN dock))))))))
    (. .)))
```

```

(S1 (S (NP (NP (DT The) (NN reptile))
  (, ,)
  (VP (VBN known)
    (PP (PP (IN as) (NP (NNP Igwig)))
      (, ,)
      (CC and)
      (ADVP (NP (RB almost) (DT a) (NN metre)) (RB long))))
    (, ,))
  (VP (VBD curled)
    (PRT (RP up))
    (PP (IN below) (NP (NNS magistrates)))
    (PP (IN at)
      (NP (NP (NNP Newport) (, ,) (NNP Isle))
        (PP (IN of) (NP (NNP Wight)))))
      (, ,)
      (SBAR (IN as)
        (S (NP (NP (PRP$ his) (NN owner) (NNP Susan) (NNP Wallace))
          (, ,)
          (VP (ADVP (RB locally))
            (VBN nicknamed)
            (NP (DT the) (NNP Lizard) (NNP Lady)))
          (, ,))
        (VP (VBD denied)
          (NP (NP (DT the) (NN attack))
            (CC and)
            (NP (NP (CD two) (NNS charges))
              (PP (IN of) (NP (JJ animal) (NNP cruelty))))))
          (, .)))

```