Habeas Corpus!

John Nerbonne

Taal- en spraaktechnologie 2009

j.nerbonne@rug.nl

Using Jennifer Spenader’s slides as basis for lecture
on corpora.

Habeas Corpus

- *Habeas corpus*: Latin ‘you shall have the body’, explained in Wikipedia
  - a legal action […] through which a person can seek relief from unlawful detention. It protects the individual from […] being harmed by the judicial system.

Goals

- Learn what corpora are, how they are made
- Understand the motivation for the use of corpora in language technology (and linguistics)
- Learn what (types of) corpora are available
- Understand how the size and type of corpus affects what kind of questions you can answer and what type of applications you can make
- Get a feeling for the importance of size for different tasks

Corpora

What’s a *corpus*?

- A representative collection of texts formatted in a consistent way

Corpora as surveys

"Corpora are just like opinion polls: you take **a representative** and **sufficiently large** sample of **a well-defined population**"

(Lars Borin)

How do you build a corpus?

- Decide
  - what type of language you want to study/have information about
  - what criteria you will use to choose texts
  - what information you will include besides the texts themselves
  - what format the corpus will be in

A representative collection?

‘Representative’ often depend on one’s research questions, but often researchers wish to characterize a whole language, such as English or Dutch, but how do we sample a whole language?

- representative of what a normal person reads, in those amounts? Or representative of what he *hears*?
- representative of what is published?
- representative of what is purchased?

SUC

- Stockholm-Umeå Corpus of Written Swedish (SUC)
  - 1 million words, divided into **500 excerpts** of 2000 words each
  - Swedish text from the 1990’s
  - Balanced (modeled on the Brown corpus)
  - SGML format, TEI (text encoding initiative) compatable, POS tagged

Example of SUC <s> unit

```
<s id=ab07c-003>
<w>Året<ana><ps>NN<m>NEU SIN DEF NOM<b>år</w>
<w>var<ana><ps>VB<m>PRT AKT<b>vara</w>
<date>
<w>1932<ana><ps>RG<m>NOM<b>1932</w>
</date>
<d>.<ana><ps>MAD<b>.</d>
```

Criteria

- How was SUC sampled :
  1. Balanced: Different text types and different styles
     - Consulted publication catalogues
  2. Mirrors what people read:
     - Consulted library statistics about what was loaned
  3. Only original Swedish texts
     - so cheated on 2, since many people read translated texts
  4. All texts have to be distributable for non-commercial research, e.g. copyright has to be obtained
     - very difficult step, influences what can be done with the corpus when you are finished

Criteria, cont.

5. Corpus only contains published texts
   - no personal letters, diaries, etc.

6. As long as it doesn’t conflict with criteria 1-5, text choice should match the Brown and Lund-Oslo-Bergen (LOB) corpora as much as possible
   - this allows the best comparison between Swedish and English

7. Machine readable form, if possible…
   - Actually in 1990 most publishers only had paper copies so this created a lot of extra work (this was before good OCR!)

Categories in Brown-corpus

A. PRESS: REPORTAGE (44 texts)

B. PRESS: EDITORIAL (27 texts)

C. PRESS: REVIEWS (17 texts)

D. RELIGION (17 texts)

E. SKILL AND HOBBIES (36 texts)

F. POPULAR LORE

(48 texts)

G. BELLES-LETTRES (75 texts)

H. MISCELLANEOUS: GOVERNMENT HOUSE ORGANS (30 texts)

J. LEARNED (80 texts)

K: FICTION: GENERAL (29 texts)

L: FICTION: MYSTERY (24 texts)

Genres, cont.

M: FICTION: SCIENCE (6 texts)

N: FICTION: ADVENTURE (29 texts)

P.FICTION: ROMANCE (29 texts)

R. HUMOR (9 texts)

Different types of corpora

- **Balanced corpora**
  - SUC
  - Brown
  - BNC (British National Corpus)
    - 100 million words
- **Parallel corpora**
  - original text and a translation, linked with another,
  - EUROPARL
    - 27 million words
    - 11 languages
    - French, Italian, Spanish, Portugese, English, German, Dutch, Danish, Swedish, Greek and Finnish

Different corpora types, cont.

- Comparable corpora
  - ICE, International Corpus of English

“to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus.”

(www.ilc.cnr.it/EAGLES96corpustypnode21.html)

More corpus types

- Monitor corpora
  - Bank of England corpus
  - continously changing
  - useful for diachronic studies
  - sometimes old texts are replaced with new ones…

Special corpora

- Childes (CHILd Language Data Exchange System)
- Enron Email dataset
- Ling-Spam
- Nigeria-letter spam corpus
- Map-task corpus

Dutch corpora

- Clef corpus
  - 78 million words (news texts)
- Twente Nieuws Corpus
  - around 700 million words now
- CGN (Corpus Gesproken Nederlands)
  - 10 million words, approx 1000 hours of speech

Corpora with annotation

- Treebanks
  - a corpus with linguistic annotation beyond the word level
  - usually manually checked syntactic annotation
  - automatically parsed corpora that have not been manually checked are not considered treebanks

    - WHY?: Because treebanks are made to be “gold standards” for other applications.

Penn Treebank

- English treebank built at the University of Pennsylvania http://www.cis.upenn.edu/~treebank

- skeletal parse
  - 2.6 million words
    - POS tagged
    - Wall Street Journal texts (mostly)
  - manually parsed
  - also includes grammatical functions and semantic relations

PTB02: wsj : raw

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.

wsj : tagged

```
[ Pierre/NNP Vinken/NNP ]
,/,
[ 61/CD years/NNS ]
old/JJ ,/, will/MD join/VB
[ the/DT board/NN ]
as/IN
[ a/DT nonexecutive/JJ director/NN Nov./NNP 29/CD ]
./.

[ Mr./NNP Vinken/NNP ]
is/VBZ
[ chairman/NN ]
of/IN
[ Elsevier/NNP N.V./NNP ]
,/,
[ the/DT Dutch/NNP publishing/VBG group/NN ]
./.
```

wsj : parsed

“Skeletal Parsing”

```
( (S (NP-SBJ (NP Pierre Vinken)
             ,
             (ADJP (NP 61 years)
         old)
             ,)
     (VP will
         (VP join
             (NP the board)
             (PP-CLR as
           (NP a nonexecutive director))
       (NP-TMP Nov. 29)))
     .))
```

wsj : parsed cont.

```
( (S (NP-SBJ Mr. Vinken)
     (VP is
         (NP-PRD (NP chairman)
        (PP of
           (NP (NP Elsevier N.V.)
              ,
              (NP the Dutch publishing group)))))
     .))
```

wsj : combined

```
( (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP-CLR (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director) ))
        (NP-TMP (NNP Nov.) (CD 29) )))
    (. .) ))
```

Childes

```
@Begin
@Participants: ABE Abel Target_Child, JEA Jeanet Mother, GER
   Gerard Investigator
@Sex of ABE:   male
@Birth of ABE: 31-OCT-1990
@Date:  30-SEP-1992
@Age of ABE:   1;10.30
@Location:     Amsterdam, The Netherlands
@Coder: Gerard Bol, Paulien
@Time duration:       9:30-10:20
@Situation:    drinking coffee at the table, where Abel sits
   in his
        highchair.
@comment:      ABE often babbles, which is indicated by <xx>
   or <xxx>.
```

Childes: Dutch “Abel”

```
*JEA:	ga je Josse bellen Abel?
*JEA:	met de telefoon.
*JEA:	nou, je bent even bezig met de koek.
*JEA:	neem me niet kwalijk.
*ABE:	nou koek.
*GER:	ja, lekker, he?
%com:	ABE laughs and GER laughs.
*GER:	com: GER scrapes his throat and ABE imitates GER.
*GER:	ja, <ik heb ook> [//] ik moet af en toe even mijn keel
   schrapen.
*ABE:	koekje.
*GER:	koek lekker.
*ABE:	koeke koeke.
*GER:	hmm.
*ABE:	xxx.
*ABE:	xxx.
*ABE:	koekje.
*JEA:	ikke koeke.
*JEA:	ja, hij kan inderdaad nu &ik zeggen.
```

CLEF: adb

```
<DOC>
<DOCNO>AD19940103-0001</DOCNO>
<DOCID>AD19940103-0001</DOCID>
<HEAD>
<PA>1</PA>
<DAT>19940103</DAT>
<UI>ad</UI>
<COP>Algemeen Dagblad</COP>
<RES>Nee</RES>
<LEW>256</LEW>
</HEAD>
```

CLEF: adb

```
<BODY>
<TI>
<P>Jaarwisseling eist vier levens</P>
</TI>
<LE>
<P>AMERONGEN - In de nieuwjaarsnacht zijn vier doden gevallen. Ook
   hebben tientallen mensen ernstige verminkingen opgelopen door
   vuurwerk. De politie noemde de jaarwisseling "rustig'.</P>
</LE>
<TE>
<P>Feestvierders hebben vooral politiemensen regelmatig op de korrel
   genomen. De agenten moesten waarschuwingsschoten lossen om zichzelf
   te ontzetten of de doorgang van hulpdiensten te bewerkstelligen. In
   Amersfoort werd een agent in zijn buik geschoten; bij is geopereerd
   en zijn toestand is nu redelijk. In Leeuwarden, Rotterdam, Amerongen
   en Groningen vielen doden bij de nieuwjaarsviering. In de grote
   steden was het - met uitzondering van Rotterdam - rustiger dan
   andere jaren. Dat was vooral opvallend in Den Haag, waar "slechts'
   35 aanhoudingen werden verricht. </P>
```

Alpino parsed CLEF

```
<alpino_ds version="1.2">
–
<node begin="0" cat="top" end="5" id="0" rel="top">
–
<node begin="0" cat="smain" end="4" id="1" rel="--">
<node begin="0" end="1" frame="tmp_noun(de,count,sg)" gen="de" id="2" lcat="np" num="sg" pos="noun"
    rel="su" root="jaarwisseling" sense="jaarwisseling" special="tmp" word="Jaarwisseling"/>
<node begin="1" end="2" frame="verb(hebben,sg3,transitive)" id="3" infl="sg3" lcat="smain" pos="verb" rel="hd"
    root="eis" sc="transitive" sense="eis" word="eist"/>
–
<node begin="2" cat="np" end="4" id="4" rel="obj1">
<node begin="2" end="3" frame="number(hoofd(pl_num))" id="5" infl="hoofd(pl_num)" lcat="detp" pos="num"
    rel="det" root="vier" sense="vier" word="vier"/>
<node begin="3" end="4" frame="tmp_noun(het,count,pl)" gen="het" id="6" lcat="np" num="pl" pos="noun"
    rel="hd" root="leven" sense="leven" special="tmp" word="levens"/>
</node>
</node>
<node begin="4" end="5" frame="punct(punt)" id="7" lcat="punct" pos="punct" rel="--" root="." sense="."
    special="punt" word="."/>
</node>
<sentence>Jaarwisseling eist vier levens .</sentence>
<comments>
  </comments>
</alpino_ds>
```

Corpus linguistics

- When you use corpus data to draw linguistic conclusions
  - requires first determining whether corpus data can answer one’s research question
- All language areas have researchers using corpus data:
  - lexicography, syntax, semantics, pragmatics, child/second language acquisition, language contact, …

Should corpora be “loathed” (Pullum)?

- Before Chomsky, use of corpora was widespread (Brew & Moens, 2002:11)
  - “Käding conducted a heroic feat of social engineering by organizing 5000 Prussian analysts to count letter occurrences in 11 million words of text, using this as the basis of a treatise on spelling rules. It is worth considering the logistics of doing this in 1897. It now takes a matter of minutes to obtain similar data from the large corpora of text which are available to us.”

Chomsky

- You can learn nothing from corpora
  - the frequency of a word tells you nothing about the word itself, only about the concept it is used for
  - corpora represent performance, not competence. The goal of linguistic research should be to determine what the underlying competence of a speaker is. Since corpora contain errors which native speakers recognize, the corpora do not reflect competence
  - introspection (aka `armchair linguistics’) is the only reliable research method

Three reasons to want corpora

- For some sorts of language, native speaker introspections are unavailable
- Bresnan: introspection underestimates variation, and appears to reflect probability
- Abney: ambiguity is so widespread that characterizing competence leaves little chance of modeling (or mimicking) communication

Where introspection fails:

- Studying child language, language contact, second language learning, …
- Our intuitions are not always reliable
  - Is the verb '***quake***' transitive or intransitive, e.g.
    - He quaked.
    - He quaked the table.
  - Cobuild and Longman: intransitive
  - Oxford: transitive
- Atkins & Levin (1995): 50,000,000 words: ***quake*** is transitive.

Against Reliance on Introspection

- Joan Bresnan (2007):
  - “Linguistic intuitions of grammaticality are deeply flawed because (1) they seriously underestimate the space of grammatical possibility by ignoring the effects of multiple conflicting formal, semantic and contextual constraints, and (2) they may reflect probability instead of grammaticality.”

- Bresnan’s Ph.D. thesis (1972) was supervised by Chomsky.

Bresnan’s Arguments

- *Who gave you that wonderful watch?*
- *Who gave that wonderful watch to you?*

- Double Object ‘*gave you that watch*’
- Prepositional Dative ‘*gave that watch to you*’

- Dozens of studies, 1965-2005
  - Very well-studied on basis of intuition

(1) Underestimating Grammatical Possibilities, Background

- Intuition-based research reports several non-alternating verbs where examples are attested in corpora
  - Research report: * I dragged John the box
  - Attested: … while Sumomo dragged him a can of beer

  - Report: * Susan whispered Rachel the news
  - Attested: She came back and whispered me the price
- Both follow pattern: double NP construction 10 times more common when 1st NP is pronoun
  - But ignored in many research papers

Bresnan (2): Intuitive judgments reflect probabilty

- Corpus studies: dbl. obj. variant preferred depending on verb & semantic class, increasing when NP object (first element) is
  - pronominal (vs. full)
  - definite (vs. indefinite)
  - human (vs. non-human)
  - highly accessible (vs. previously unmentioned)
- Model predicts variant w. acc. 94%
- Hypothesis: intuitive well-formedness (Chomskyan) judgments mirror corpus probabilities

Judgments

- Subjects asked to rate pairs examples taken from dialogues (and modified), given context:

> 25, 26 yrs. ago, my brother showed up in the front yard pulling a trailer. And in this trailer he had a pony, which I didn't know he was bringing. So over the weekend I had to go out and find some wood and put up some kind of structure to house that pony …

- …because he brought the pony to my children
- …because he brought my children the pony

Experiment (cont.), Results

- Subjects had to give two scores totalling 100 to the the two options (e.g. 95 – 5, 33 -67)

- **Results**: subjects tend to favor the options predicted by the model derived from the corpus

- **Conclusion**: intuitive judgments reflect (complex) probabilities implicit in corpora

Experiment 2

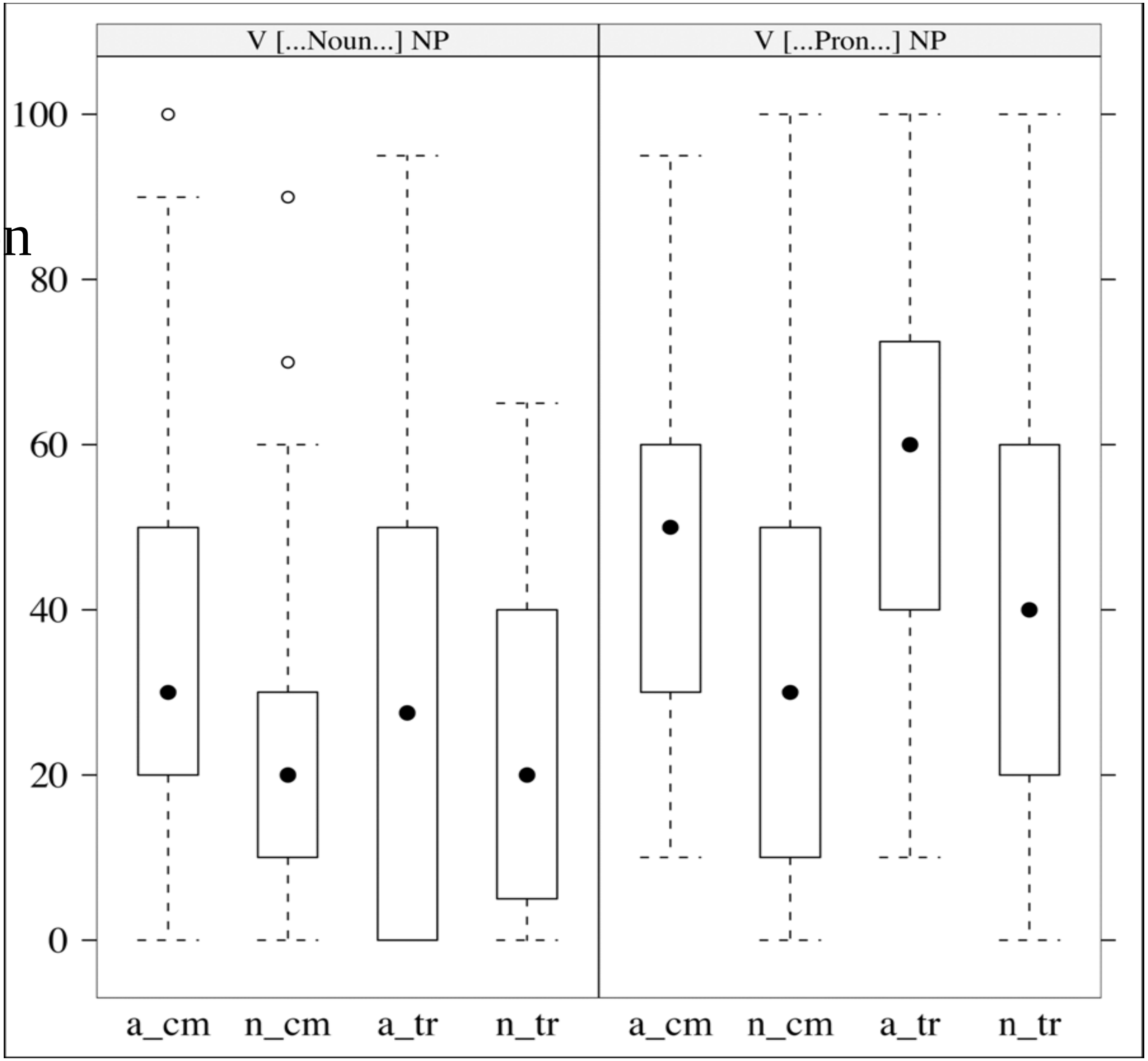
- Begin w. allegedly non-alternating verbs
  - * I pushed him the pot
- Find context in which alternative occurs
  - Money in the pot is dead money. It doesn’t belong to anyone until the hand is over and
  - …the dealer pushed the pot to someone
- & modify it to create “ill-formed” exmpl
  - …the dealer pushed someone the pot
- Hypothesis: even allegedly non-alternating verbs will show improved acceptability.

Experiment 2, cont.

- Method as in Exp. 1, relative well-formedness

a alternating
n non-alternating

cm – communication
"whisper"
tr – transfer, "push"



Note: alternating on left (NP), non-alternating on right (pro)

Experiment 2, cont.

- Subjects rated “non-alternating” verbs with pronoun better that alternating verbs with full NPs!
  - “Non-alternating” examples claimed to be ill-formed.
- Conclusion: Intuitive judgments unreliable because they abstract away from context and important additional acceptability factors

More on datives

- Bresnan (2007), “Typology”: more examples claimed to be ill-formed, attested on the web
  - The movie gave me the creeps
  - * The movie gave the creeps to me

- But
  - That life-size prop will give the creeps to just about anyone.
  - Stories like these must give the creeps to people whose idea of heaven is a world without religion

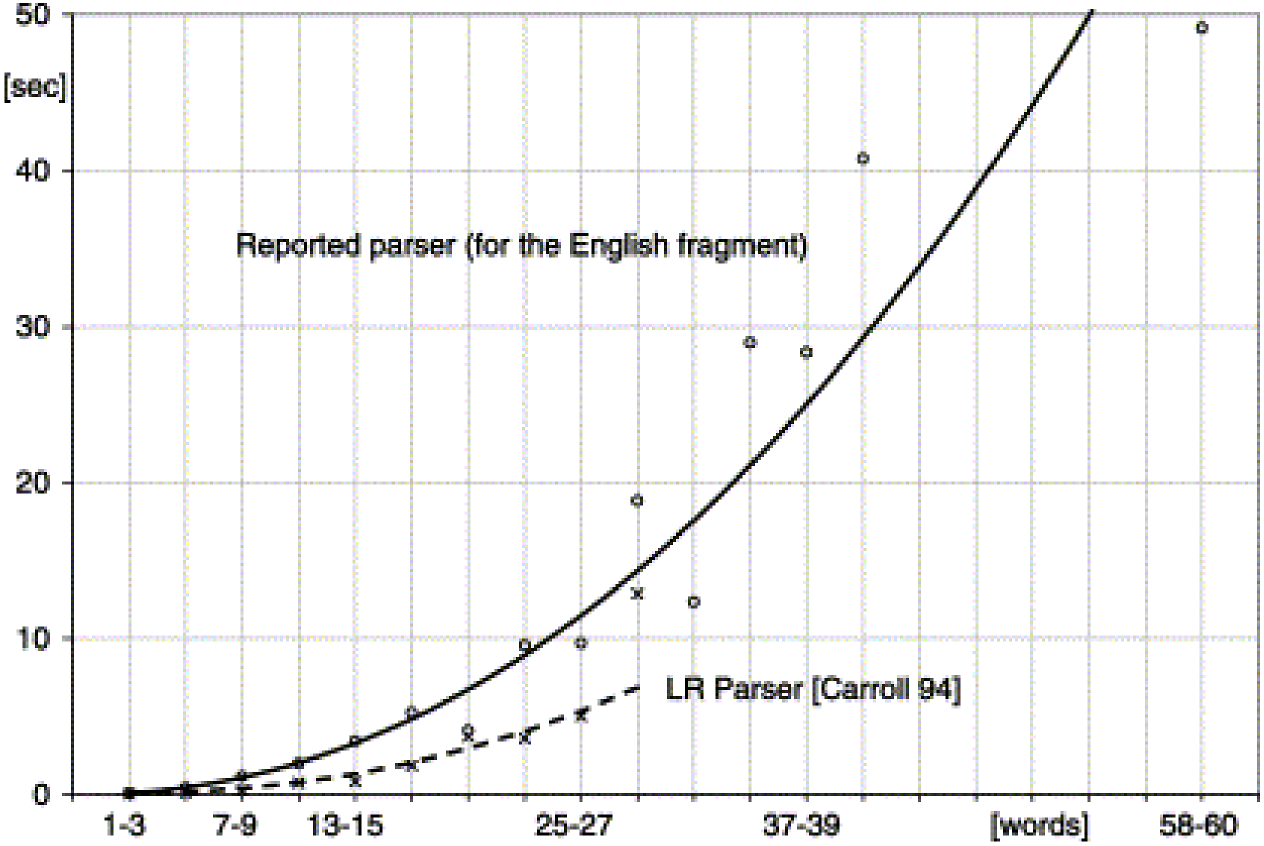
Corpora & Disambiguation

- Disambiguation: choosing which structure or meaning is meant
  - Marie bought a car in Haren
  - Marie bought (a car in Haren)
  - Marie bought a car. The act of buying took place in Haren. (Compare buying stocks, buying vacation property at lawyer’s office.)
- Focus of lots of work in CL, Psycholing.

Abney’s contribution

- If we focus only on competence, we do not model disambiguation.
  - **Cognitively**, disambiguation is very interesting
  - **Practically**, disambiguation is needed in many applications – interfaces, information retrieval and extraction, translation, …
- The problem is substantial – many sentences are ambiguous
  - Attachment ambiguities multiply in the worst case, leading to exponential growth of number of readings in sentence length

Parsing times as function of sentence length (optimized)



Komagata, *Computer Speech & Language*, 2004

How big a corpus is needed?

- Depends on the task
  - if your studying *do*-support OR want to train a part-of-speech tagger
    - small corpus is probably fine
  - But for some tasks Very Large Corpora are needed
  - Clef corpus: 78 million words…
    - not a single example of *zich* or *zichzelf* used with *kammen*, e.g.
      - Hij kamt zichzelf.

Why not just use the internet?

Linguists for the responsible use of the internet

(www.unc.edu/ lajanda/responsible.html)

Advantages of internet data:

- The internet contains billions of electronically searchable words, more than any corpus.
- New data is constantly being added, which means that data includes current usage (unlike corpora, which quickly age and become outdated).
- Search engines can provide statistical data about the number of ”hits” for each search, making it possible to compare the relative frequency of linguistic forms.

Disadvantages

Disadvantages of internet data:

- There is no editorial oversight, no gatekeeper, which means that we have no control over what kinds of texts are included or whether they contain typographical errors, etc.
- Some internet pages contain material not written by native speakers of the language and might not represent authentic usage.
- Since it is difficult to determine the total number of words queried in an internet search, it may be impossible to determine the absolute frequency of a linguistic phenomenon.
- **It isn’t balanced: the reliability of an observation is not determinable**

Representative of the language means?

- How big does your corpus have to be to be “representative”?

– We want our corpus to include reliable information about the use of what we’re studying.

- But this is often very hard

Types and tokens

- **Token** : a single occurrence of a construction

  A word, but also a punctuation mark, a digit, or a date
- **Type :** a unique construction, a form
- **Type / token ratio** number of types divided by the number of tokens
- *the young man and the old man and the old cat*
- 11 tokens
- 6 types
- type/token ratio = 6/11 0.545
- Slide adapted from Begona Villada

How much is representative?

- “A representative corpus should include the majority of the types in the language as recorded in a comprehensive dictionary.”
- *S o how big would a corpus need to be to get the representativeness we want?*
  - *is a million words enough?*
  - *is 70 million words enough?*

  - The Brown Corpus contains slightly more than one million words (tokens!)
  - What does this mean in terms of **types**

High frequency types in the Brown Corpus

- The ten top-ranked word types in the Brown corpus are
  - *the, of, and, to, a, in, that, is, was, for*
  - These types are so frequent that they make up 23% of the total token count
    - 232,425 occurrences over 996,883 tokens in total
  - Every fifth word in the corpus is one of these words!
- *the* as the most frequently occurring type by itself accounts for nearly 7% of all token occurrences !
  - 69971 out of slightly over 1 million
  - Only 135 word types are needed to account for half the Brown Corpus!

Low frequency words in the Brown Corpus

- Looking at *least frequent words* we find that there is a massive amount of hapax legomena, types that only occur once in the corpus
  - 24,374 types only occur once
    - Words like : *salaries, parentheses*
  - The Brown corpus only has 53,076 distinct types
  - This means that half the types in the corpus are types that only occur once
  - 70% of the corpus is made up of types that occur less than four times.
  - However, these types only make up 5% of the total tokens (52,033 tokens, and there is a total of 996,883 tokens)

Which are tokens and which are types?

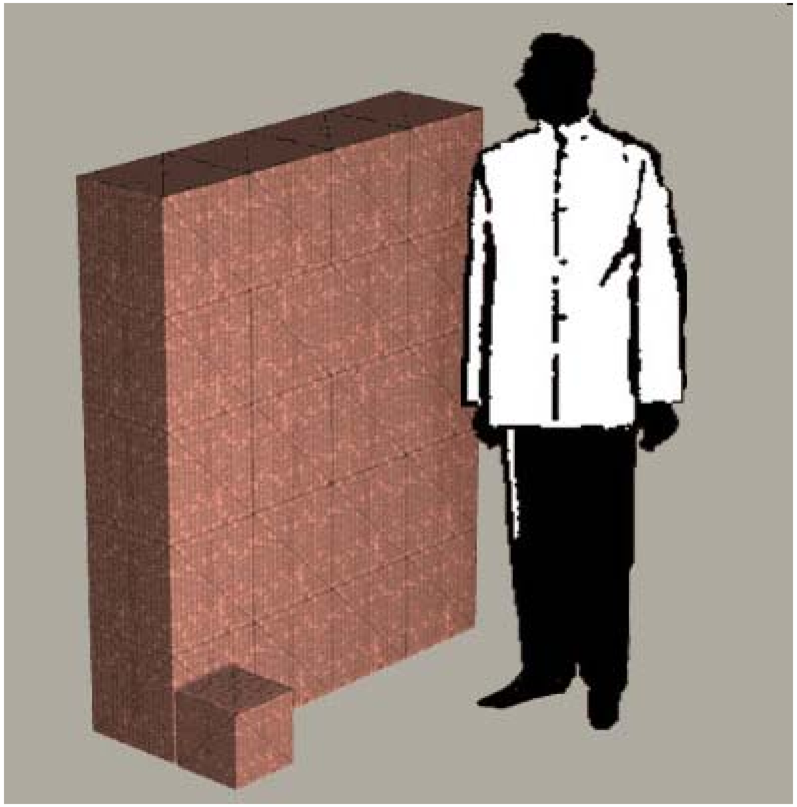
- in the Brown Corpus "the" is the most frequently occurring **word**, and all by itself accounts for nearly 7% of all **word** occurrences (69971 out of slightly over 1 million). True to Zipf's Law, the second-place **word** "of" accounts for slightly over 3.5% of **words** (36411 occurrences), followed by "and" (28852). Only 135 **vocabulary items** are needed to account for half the Brown Corpus.

What do these statistics tell us about the nature of language?

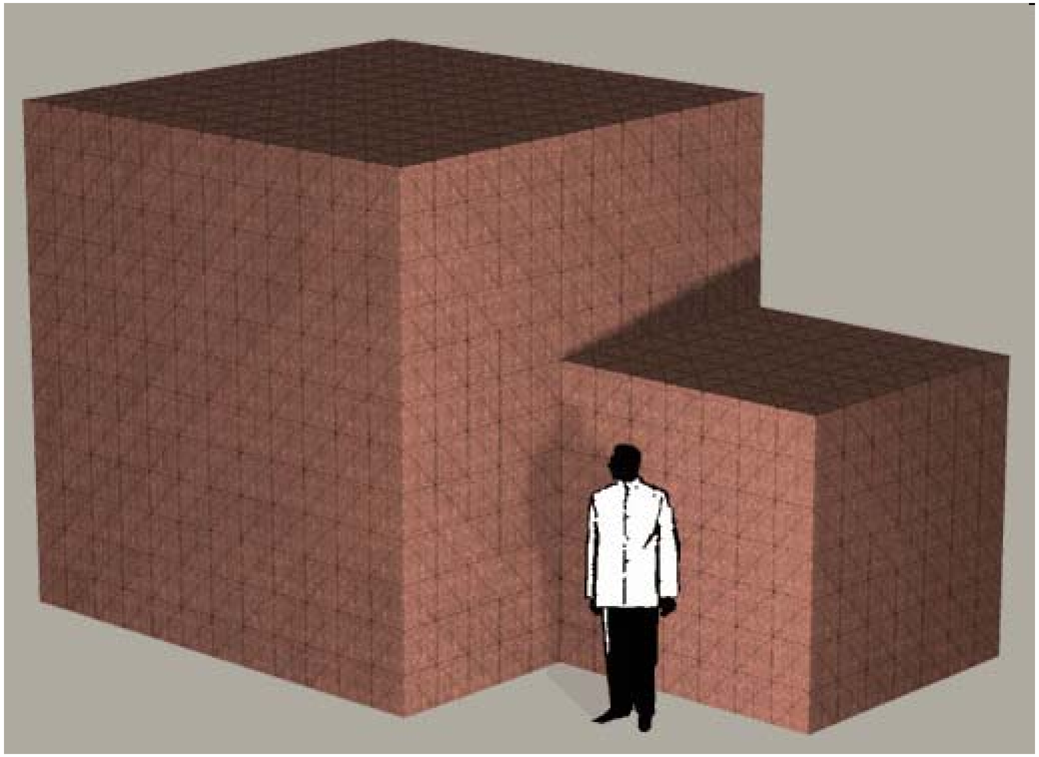
Size of corpora

- SUC & Brown = 1 million words
- Penn Tree bank: 2.6 million
- Michigan Corpus of Academic Spoken English = 1.8 million
- CGN: 10 million!
- BNC: 100 million words
- Twente Nieuws Corpus: 700 million
- Marcu & Echihabi (2002): 1 billion words
- For more corpora than you would ever want to know about:
  - www. lancs.ac.uk/postgrad/xiaoz/papers/corpus

1,003,776
One million, three thousand,
seven hundred and seventy-six
Pennies



100,016,640
One hundred million, sixteen thousand, six hundred and forty
Pennies



http://www.kokogiak.com/megapenny/eight.asp

Zich-zichzelf distribution

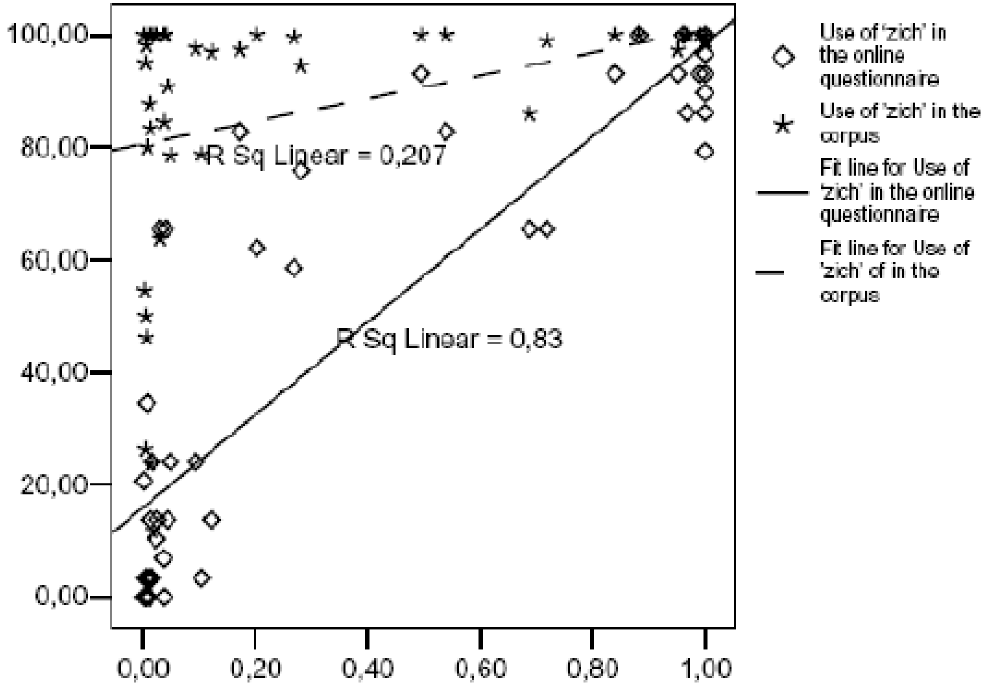
- Smits, E.J, P. Hendriks & J. Spenader (2007). Using very large parsed corpora and judgment data to classify verb reflexivity. In: António Branco (Ed.), Anaphora: Analysis, Algorithms and Applications. 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007, LNAI (Lecture Notes in Artifical Intelligence) #4410, Springer-Verlag Berlin Heidelberg, pp. 77-93.

(1) Jan wast zich/zichzelf.

(2) Jan schaamt zich/*zichzelf

(3) Jan wast zijn zoon.

Question can be answered with corpus data



Summary

- Corpora are important tools for
  - studying language
  - developing and testing language models
- They are more than just a collection of texts
  - because they are representative they are more useful