Corpus Statistics

COSI 140 – Natural Language Annotation for Machine Learning

James Pustejovsky

January 26, 2016
Brandeis University

Corpora, Types, and Tokens

We now have available large corpora of machine readable texts in many languages.

One good source: Project Gutenberg (http://www.promo.net/pg/)

We can analyze a corpus into a set of:

- word tokens (instances of words), and
- word types or terms (distinct words)

So, "The boys went to the park" contains 6 tokens and 5 types.

Zipf's Law

George Kingsley Zipf (1902-1950) observed that for many frequency distributions, the *n*-th largest frequency is proportional to a negative power of the rank order *n.*

Let t range over the set of unique events. Let f(t) be the frequency of t and let r(t) be its rank. Then:

$\forall t\ r(t) \approx c * f(t)^{-b}$ for some constants b and c.

Zipf's law

- Observation: Frequency decreases non-linearly with rank.

$$f(w) = \frac{C}{r(w)^a}$$

a constant, determined from data, roughly the frequency of the most frequent word

a constant, determined from data

- Suppose a = 1, and C = 60,000.
- The model predicts:
  - 2nd most frequent word will be C/2 = 30,000
  - 3rd most frequent: C/3 = 20,000
  - 20th most frequent = C/20 = 3000
- So frequency decreases very rapidly (exponentially) as rank increases.

Things to note

- The law doesn’t predict frequency ties
  - there are no ties among ranks

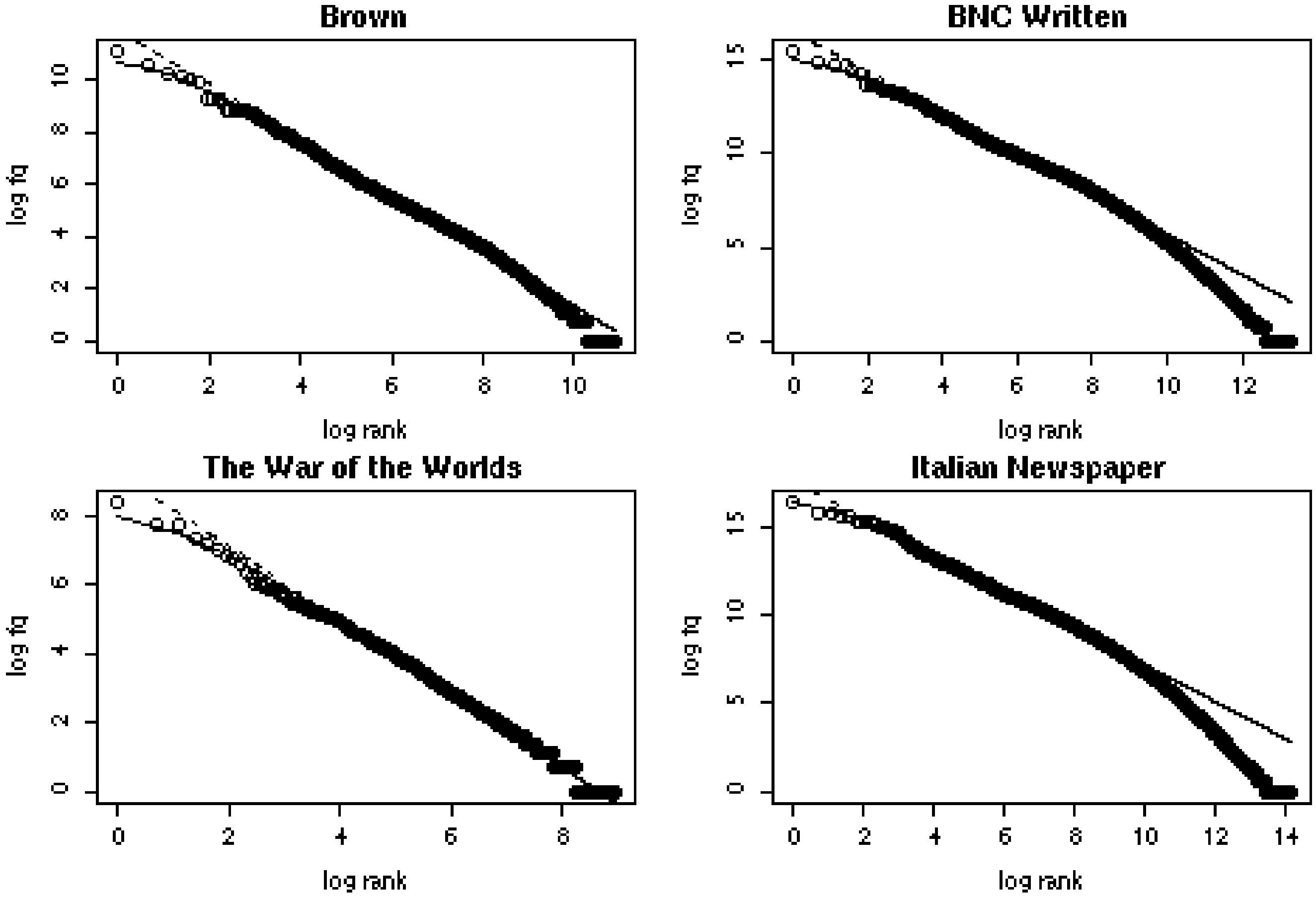
- The law is a power law: frequency is a function of negative power of rank

- Taking the log of both sides gives us a linear function:

$$\log f(w) = \log C - a \log r(w)$$

  - Basically a straight line plot.

Log-log plot for data from Baroni 2007



Some observations

- Empirical work has shown that the law doesn’t perfectly predict frequencies:
  - at the bottom ranks (low frequencies), actual frequency drops more rapidly than predicted
  - at the top ranks (high frequencies), the model predicts higher frequencies than actually attested

Mandelbrot’s law

- Mandelbrot proposed a version of Zipf’s law as follows:

$$f(w) = \frac{C}{(r(w) + b)^{a}}$$

  - (Note: Zipf’s original law is Mandelbrot’s law with b = 0)

- If *b* is a small value, it will make frequency of items ranked at the top (rank 1, 2, etc) significantly smaller, but won’t affect the lower ranks.

Comparison

- Let C = 60,000, a = 1 and b = 1

- Then, for a word of rank 1:
    - Zipf's law predicts f(w) = 60,000/1 = 60,000
    - Mandelbrot's law predicts f(w) = 60,000/(1+1) = 30,000

- For a word of rank 1000:
    - Zipf predicts: f(w) = 60,000/1000 = 60
    - Mandelbrot: f(w) = 60,000/1001 = 59.94

- So differences are bigger at the top than at the bottom.

Linear version of Mandelbrot

$$\log f(w) = \log C - a\log(r(w) - b)$$

- Note: this is no longer a linear curve, so should fit our data better.

Consequences of the law

- Data sparseness: no matter how big your corpus, most of the words in it will be of very low frequency.

- You can’t exhaust the vocabulary of a language: new words will crop up as corpus size increases.
  - implication: you can’t compare vocabulary richness of corpora of different sizes

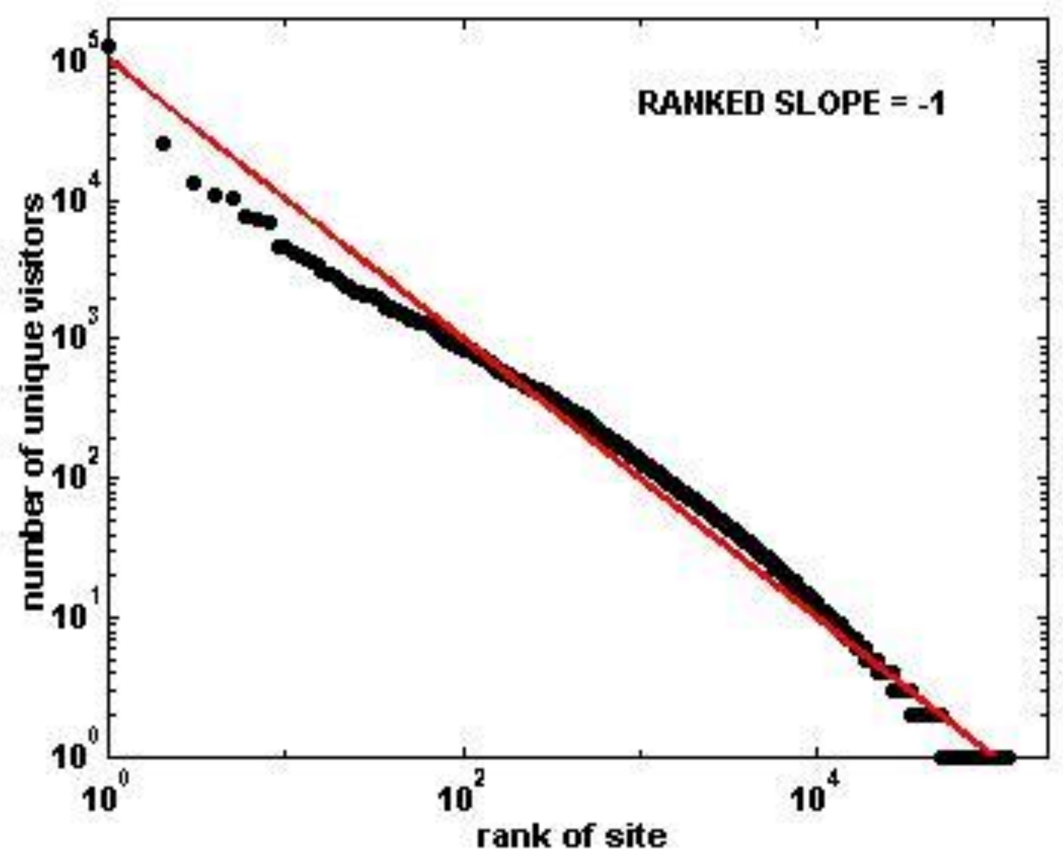
Explanation for Zipfian distributions

- Zipf’s own explanation (“least effort” principle):
  - Speaker’s goal is to minimise effort by using a few distinct words as frequently as possible
  - Hearer’s goal is to maximise clarity by having as large a vocabulary as possible

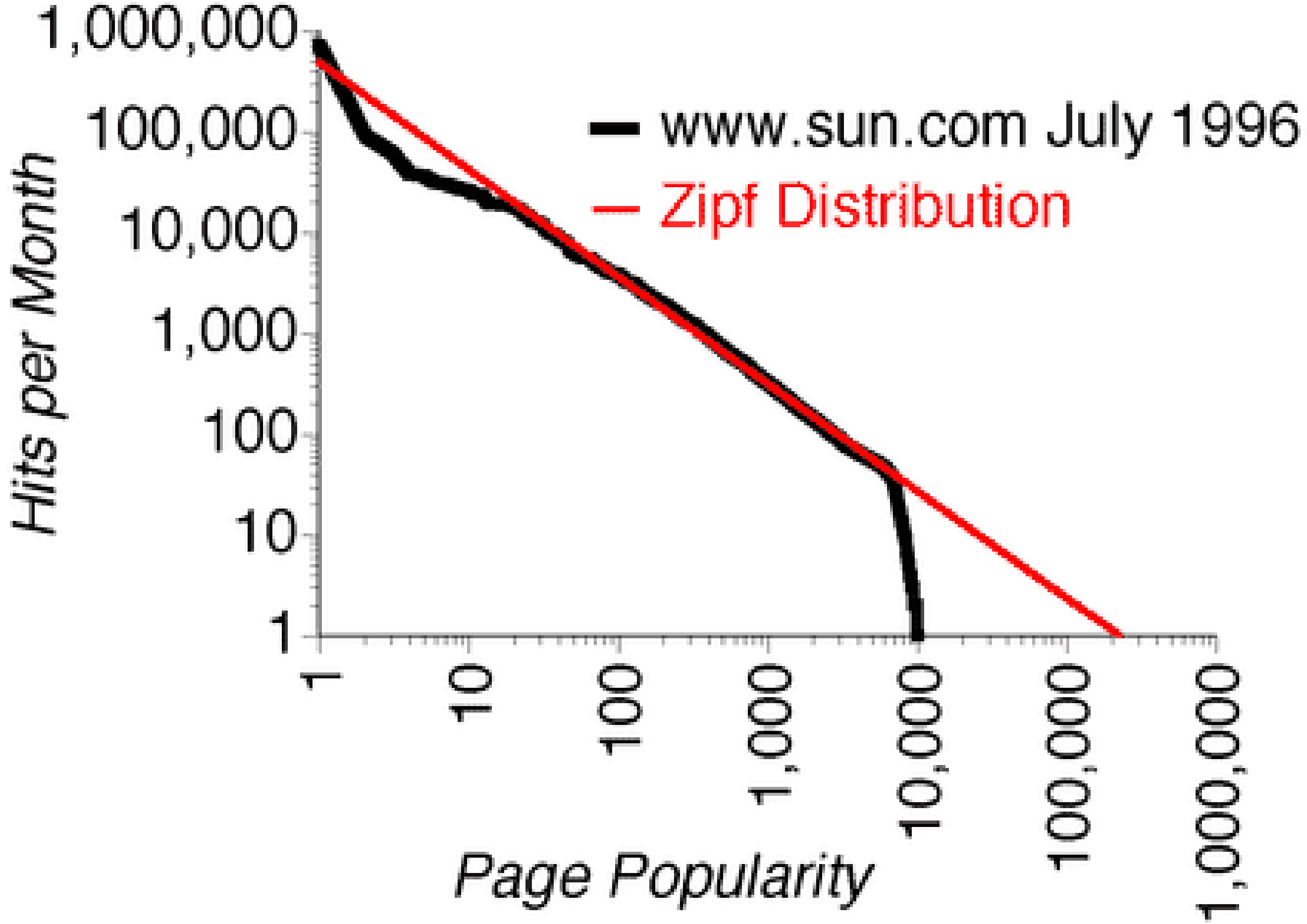
Zipf's Law Applies to Lots of Things

- frequency of accesses to web pages
- sizes of settlements
- income distribution amongst individuals
- size of earthquakes
- words in the English language

Zipf and Web Requests



Zipf and Web Requests



Zipf and Cities

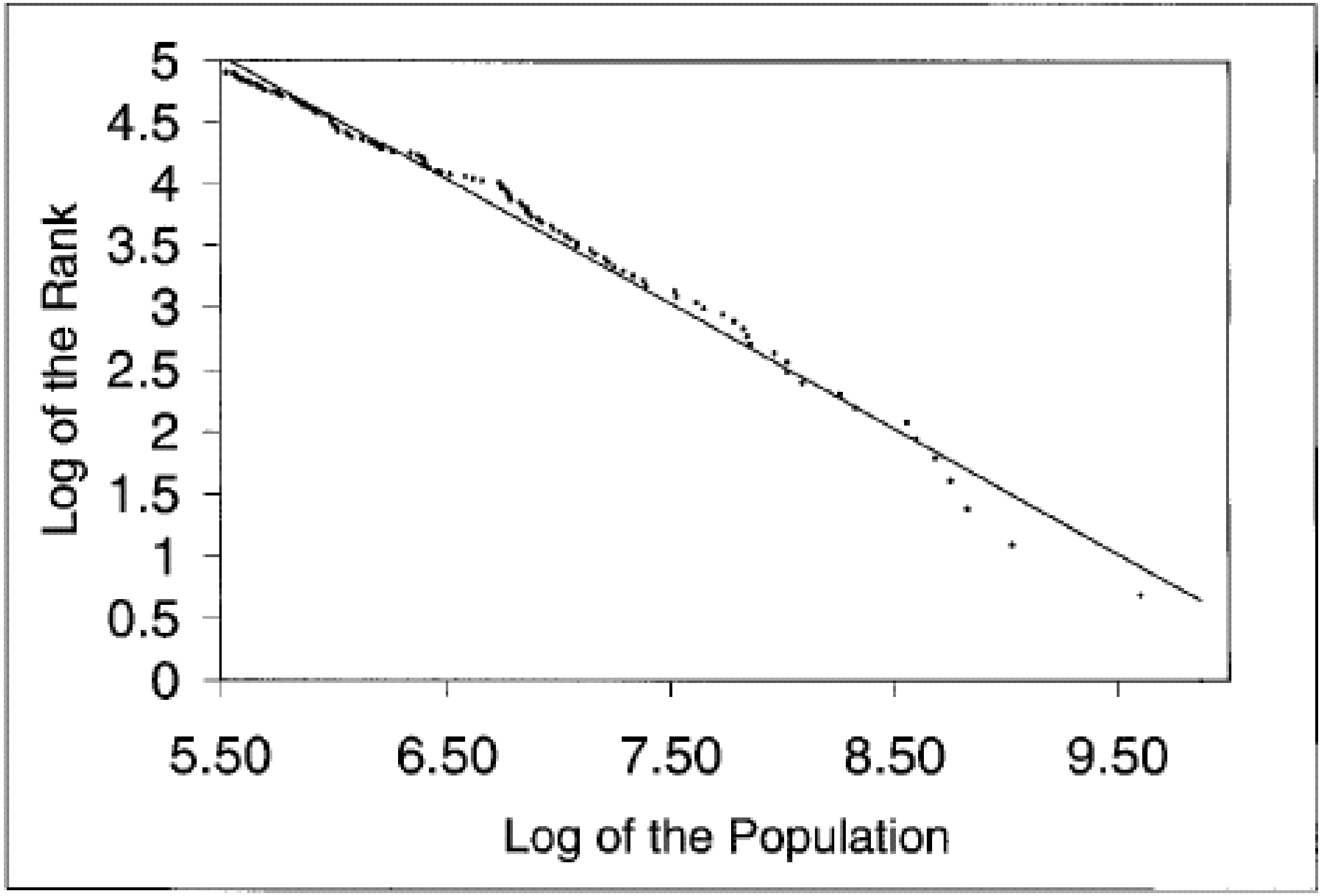


FIGURE I
Log Size versus Log Rank of the 135 largest U. S. Metropolitan Areas in 1991
Source: Statistical Abstract of the United States [1993].

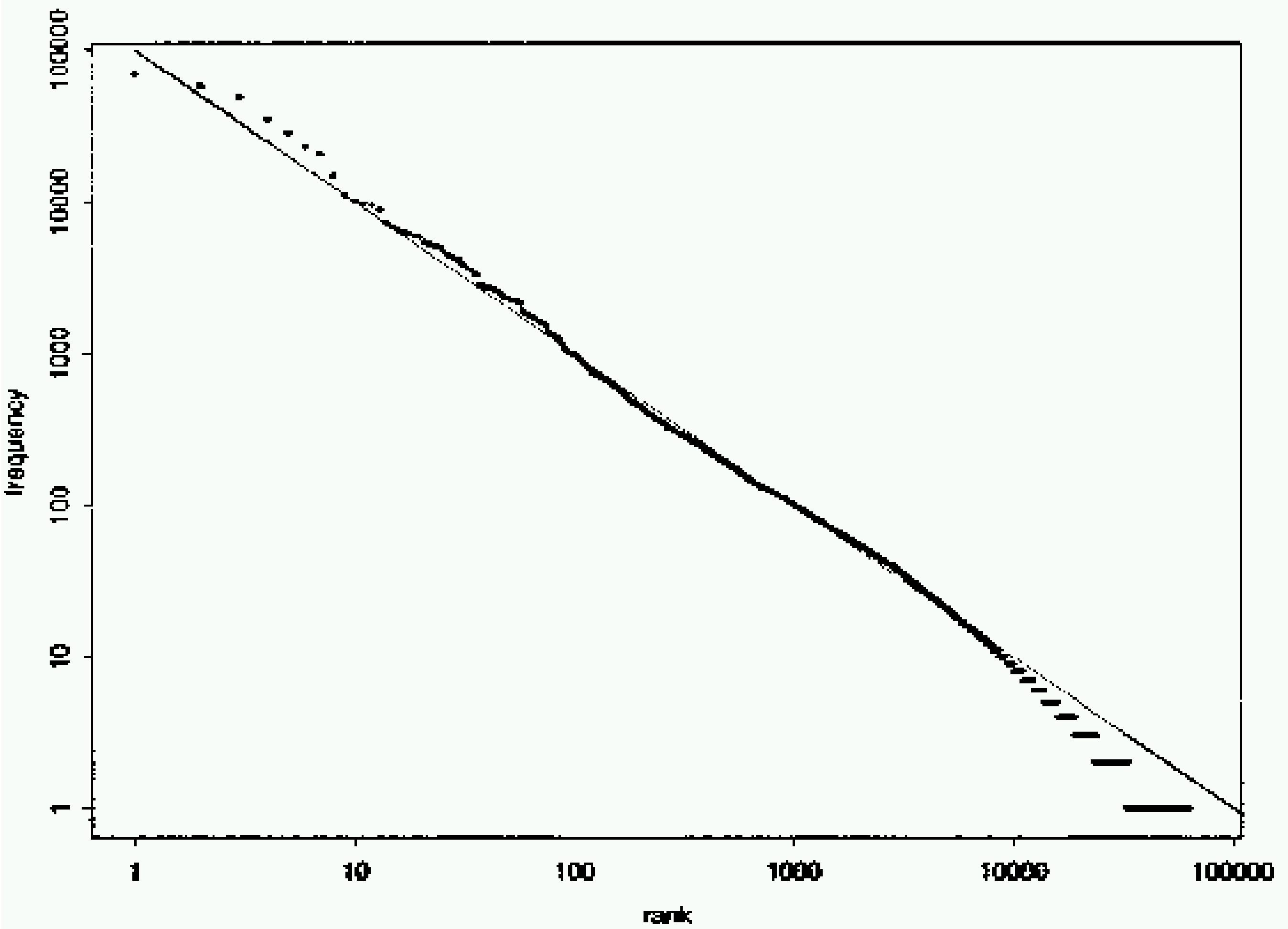
Applying Zipf's Law to Language

Applying Zipf's law to word frequencies, in a large enough corpus:

$\forall t\ r(t) \approx c * f(t)^{-b}$ for some constants b and c. In English texts, b is usually about 1 and c is about N/10, where *N* is the number of words in the collection.

English: http://web.archive.org/web/20000818062828/http://hobart.cs.umass.edu/~allan/cs646-f97/char_of_text.html

Visualizing Zipf's Law



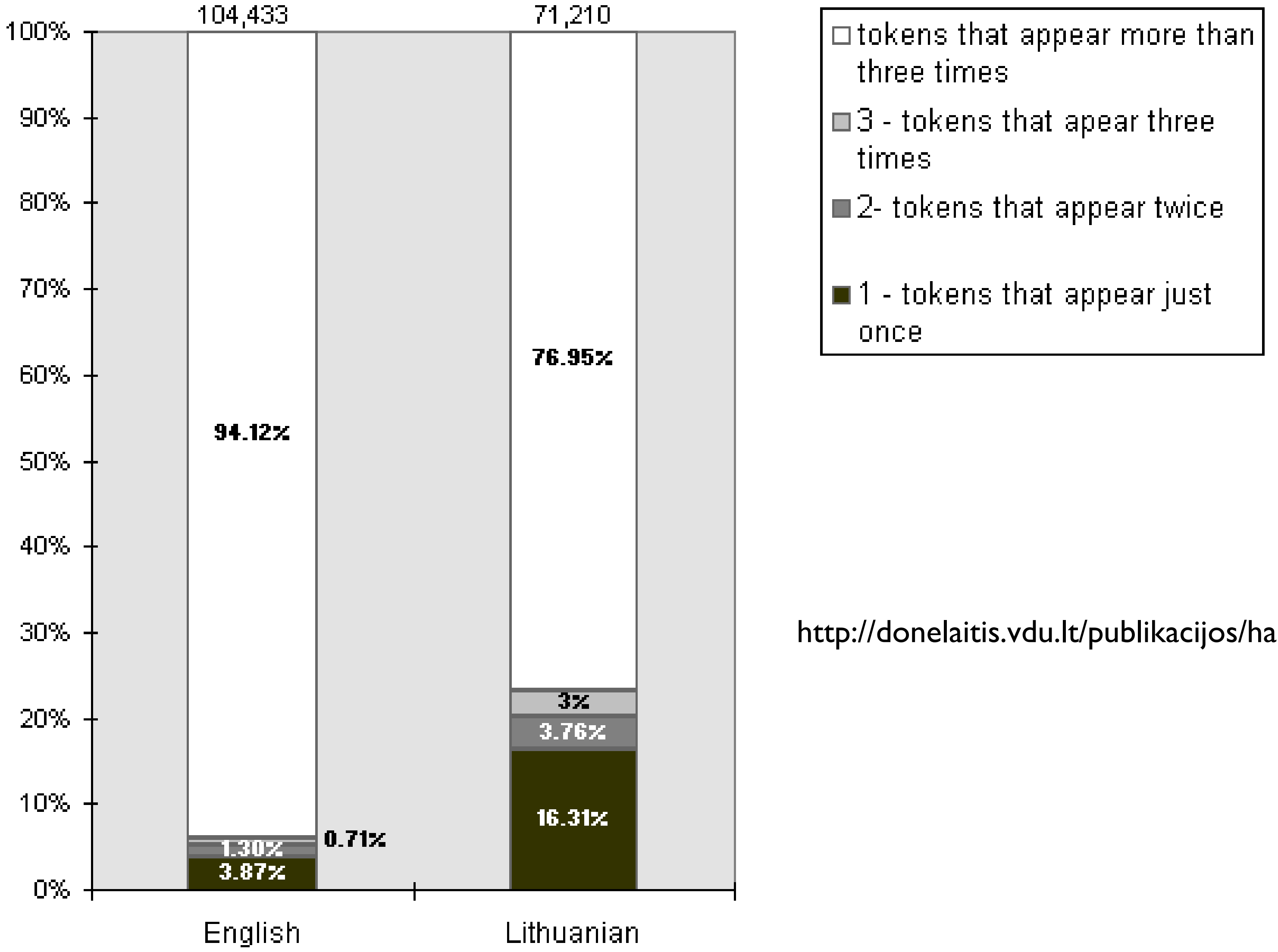
Word frequencies in the Brown corpus

From Judith A. Molka-Danielsen

Hapax Legomenon

From: Greek : *hapax*, once + *legomenon*, neuter sing. passive participle of *legein*, to count, say.

Orwell's 1984



http://donelaitis.vdu.lt/publikacijos/hapax.htm

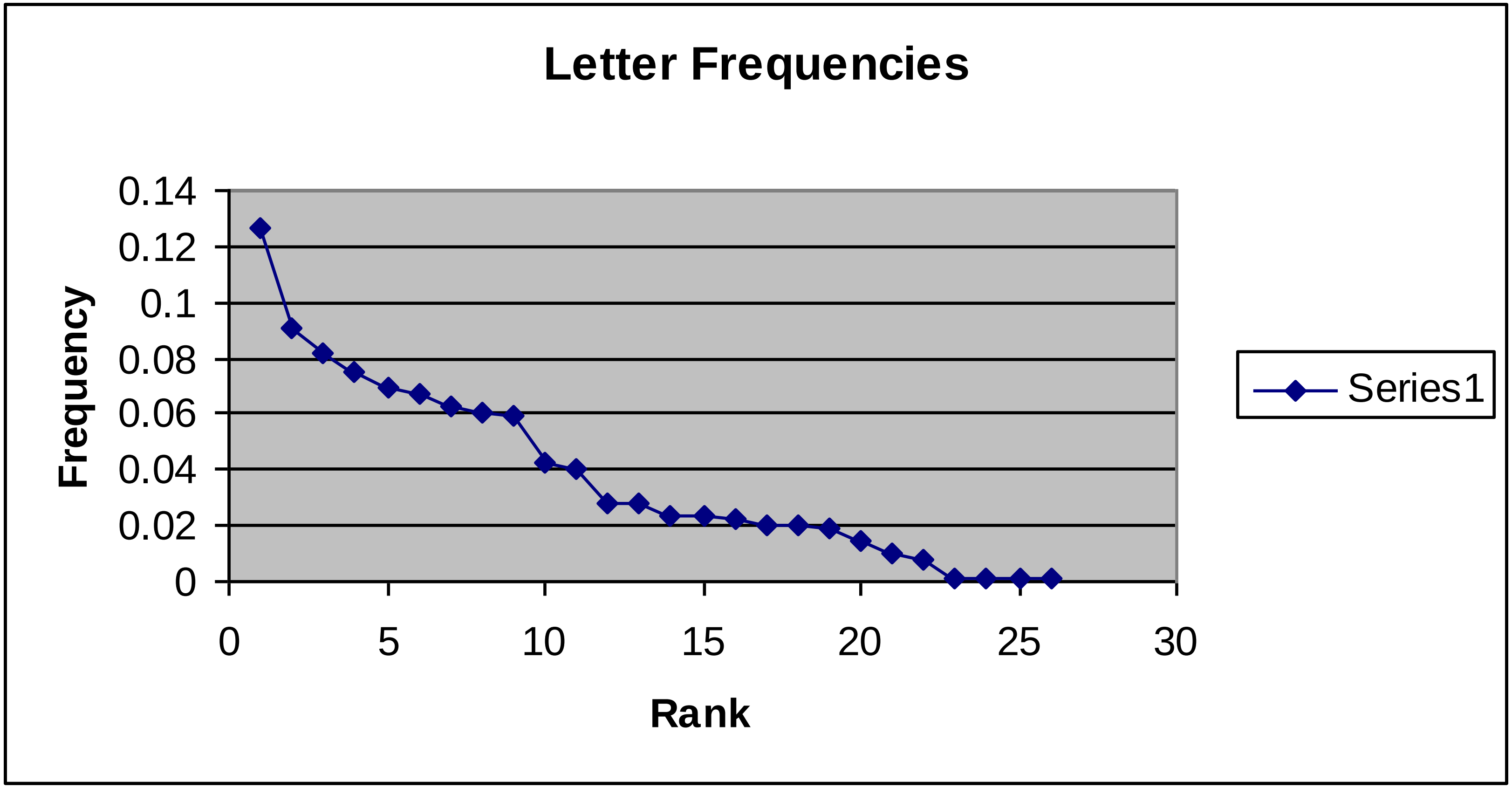
Eng: 104,433 tokens, 8,957 types. Lit: 71,210 tokens, 17,939 types

It’s Not Just English

Russian:

http://www.sewanee.edu/Phy_Students/123_Spring01/schnejm0/PROJECT.html

Letter Frequencies in English



Letter Frequencies – Additional Observations

•Frequencies vary across texts and across languages:

http://www.bckelk.uklinux.net/words/etaoin.html

•*Etaoin Shrdlu* and frequencies in the dictionary:

http://rinkworks.com/words/letterfreq.shtml

•Simon Singh's applet for computing letter frequencies:

http://www.simonsingh.net/The_Black_Chamber/frequencyanalysis.html

Redundancy in Text - Words

The stranger came early in February, one wintry day, ----- a biting wind and a driving snow, the last ----- of the year, over the down, walking from Bramblehurst ----- station, and carrying a little black portmanteau in his ----- gloved hand. He was wrapped up from head to -----, and the brim of his soft felt hat hid ----- inch of his face but the shiny tip of ----- nose; the snow had piled itself against his shoulders ----- chest, and added a white crest to the burden ----- carried. He staggered into the "Coach and Horses" more ----- than alive, and flung his portmanteau down. "A fire," ----- cried, "in the name of human charity! A room ----- a fire!" He stamped and shook the snow from ----- himself in the bar, and followed Mrs. Hall into ----- guest parlour to strike his bargain. And with that ----- introduction, that and a couple of sovereigns flung upon ----- table, he took up his quarters in the inn.

Redundancy in Text - Letters

Her visit-r, she saw as -he opened t-e door, was s-ated in the -rmchair be-ore the fir-, dozing it w-uld seem, wi-h his banda-ed head dro-ping on one -ide. The onl- light in th- room was th- red glow fr-m the fire—w-ich lit his -yes like ad-erse railw-y signals, b-t left his d-wncast fac- in darknes---and the sca-ty vestige- of the day t-at came in t-rough the o-en door. Eve-ything was -uddy, shado-y, and indis-inct to her, -he more so s-nce she had -ust been li-hting the b-r lamp, and h-r eyes were -azzled.

Redundancy in Text - Letters

Aft-r Mr-. Hall -ad l-ft t-e ro-m, he –ema-ned –tan-ing -n fr-nt o- the -ire, -lar-ng, s- Mr. H-nfr-y pu-s it, -t th- clo-k-me-din-. Mr. H-nfr-y no- onl- too- off -he h-nds -f th- clo-k, an- the -ace, -ut e-tra-ted -he w-rks; -nd h- tri-d to -ork -n as -low -nd q-iet -nd u-ass-min- a ma-ner -s po-sibl-. He w-rke- with -he l-mp c-ose -o hi-, and -he g-een –had- thr-w a b-ill-ant -ight -pon -is h-nds, -nd u-on t-e fr-me a-d wh-els, -nd l-ft t-e re-t of -he r-om s-ado-y. Wh-n he –ook-d up, -olo-red –atc-es s-am -n hi- eye-.

Order Doesn’t Seem to Matter

Aoccdrnig to rscheearch at an Elingsh uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, olny taht the frist and lsat ltteres are at the rghit pcleas. The rset can be a toatl mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae we do not raed ervey lteter by ilstef, but the wrod as a wlohe.

http://joi.ito.com/archives/2003/09/14/ordering_of_letters_dont_matter.html

Chatbots Exploit Redundancy

Let’s look at some data on the inputs to ALICE:

http://www.alicebot.org/articles/wallace/zipf.html

Why Do We Want to Predict Words?

- Chatbots
- Speech recognition
- Handwriting recognition/OCR
- Spelling correction
- Augmentative communication

Predicting a Word Sequence

The probability of "The cat is on the mat" is

P(the cat is on the mat) = P(the | <s>) □
P(cat | <s> the) □
P(is | <s> the cat) □
P(on | <s> the cat is) □
P(the | <s> the cat is on) □
P(mat | <s> the cat is on the) □
P(</s> | <s> the cat is on the mat)

where the tags <s> and </s> indicate beginning and end of the sentence.

But that is not a practical solution. Instead taking only two previous tokens,

P(the cat is on the mat) = P(the | <s>) □
P(cat | <s> the) □
P(is | the cat) □
P(on | cat is) □
P(the | is on) □
P(mat | on the) □
P(</s> | the mat)

N-grams

Approximating reality (let V be the number of words in the lexicon and T be the number of tokens in a training corpus):

$P(w_k = W) = c(W) / T$ word frequencies

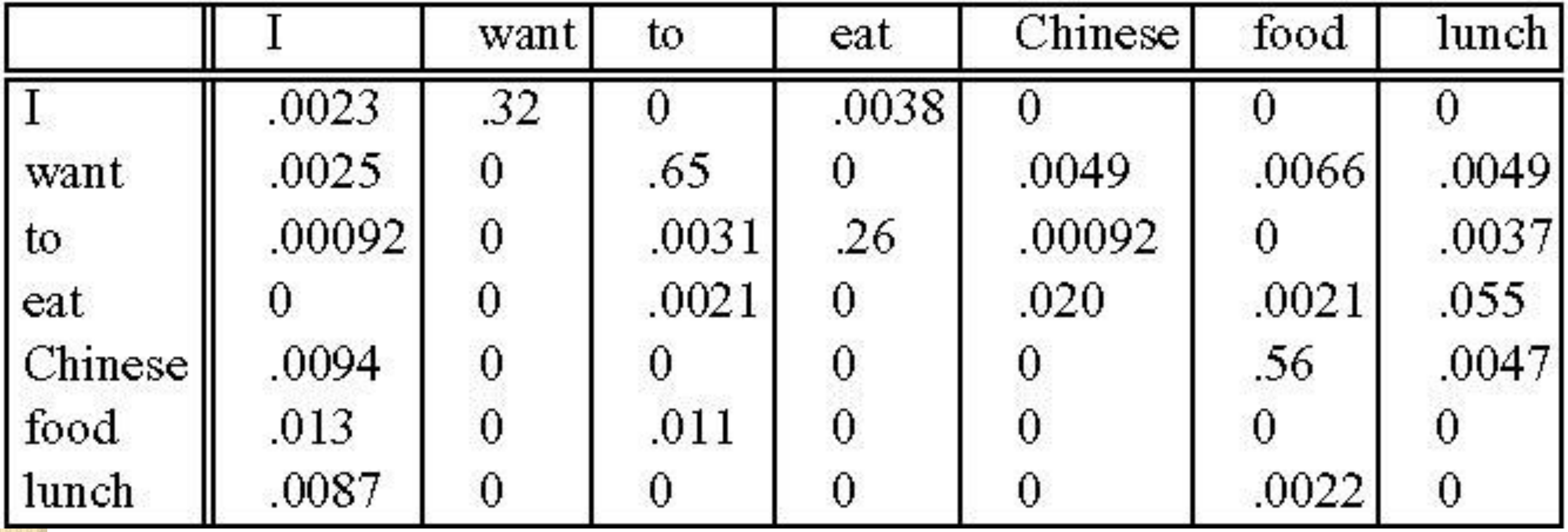
$P(w_k = W_1 \mid w_{k-1} = W_0) = c(W_0W_1)/c(W_0)$ bigrams

…

Abbreviating $P(w_k = W_1 \mid w_{k-1} = W_0)$ to $P(W_1|W_0)$.
For example P(rabbit | the).

$P(W_n|W_{n-2}W_{n-1}) = c(W_{n-2}W_{n-1}W_n)/c(W_{n-2}W_{n-1})$ trigrams

Bigram Example



	I	want	to	eat	Chinese	food	lunch
I	.0023	.32	0	.0038	0	0	0
want	.0025	0	.65	0	.0049	.0066	.0049
to	.00092	0	.0031	.26	.00092	0	.0037
eat	0	0	.0021	0	.020	.0021	.055
Chinese	.0094	0	0	0	0	.56	.0047
food	.013	0	.011	0	0	0	0
lunch	.0087	0	0	0	0	.0022	0

Smoothing

What does it mean if a word (or an N-gram) has a frequency of 0 in our data?

Examples:

•In the restaurant corpus, *to want* doesn' t occur. But it could: *I ' m going to want to eat lunch at 1*.

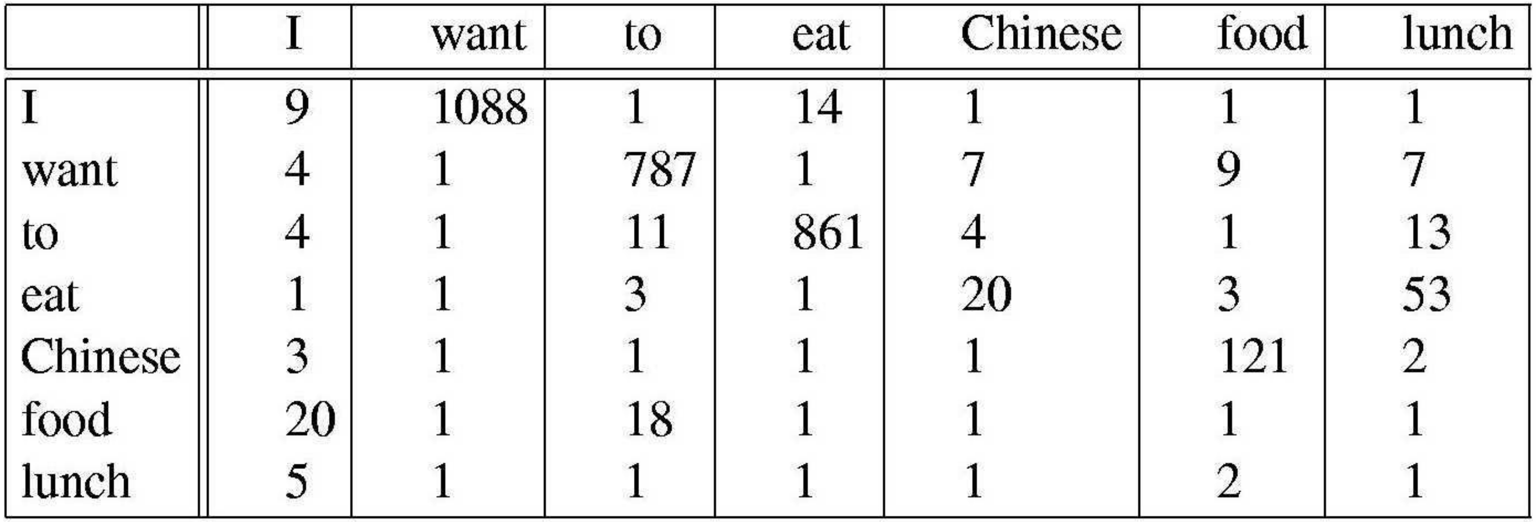
•The words *knit*, *purl*, *quilt*, and *bobcat* are missing from our list of the top 10,000 words in a newswire corpus.

•In Alice' s Adventures in Wonderland, the words *half* and *sister* both occur, but the bigram *half sister* does not.

But this does not mean that the probability of encountering *half sister* in some new text is 0.

Add-One Smoothing

First, we simply add 1 to all the counts, so we get:

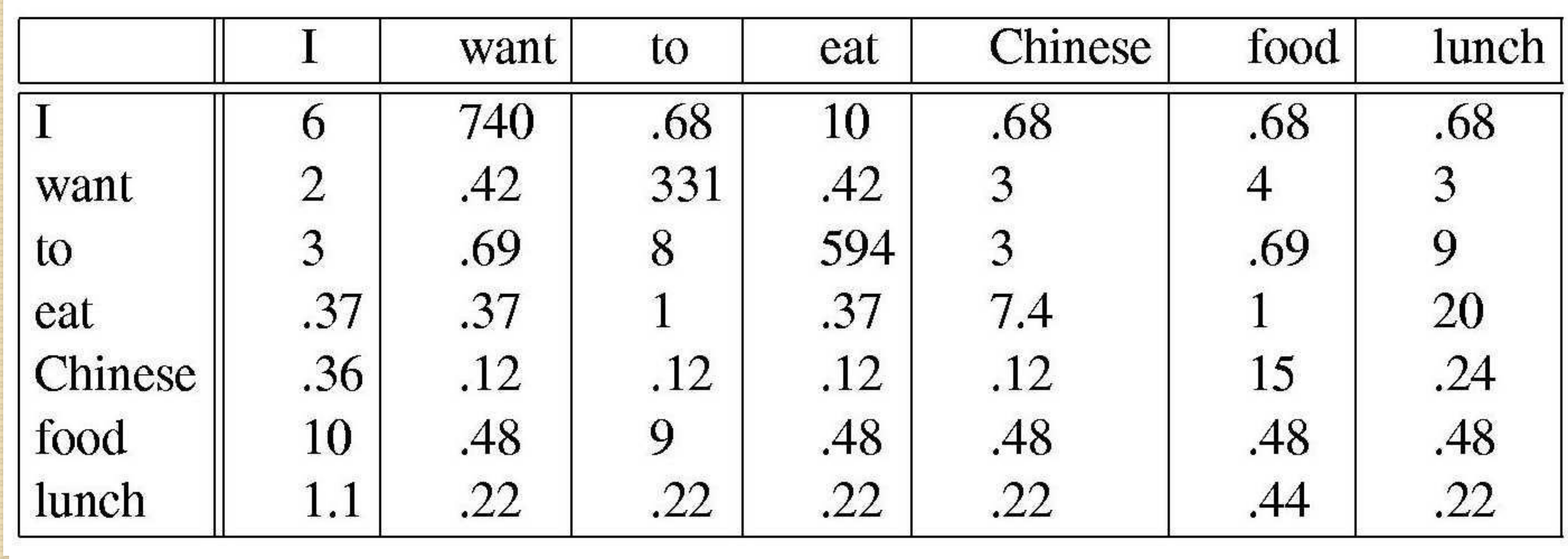


	I	want	to	eat	Chinese	food	lunch
I	9	1088	1	14	1	1	1
want	4	1	787	1	7	9	7
to	4	1	11	861	4	1	13
eat	1	1	3	1	20	3	53
Chinese	3	1	1	1	1	121	2
food	20	1	18	1	1	1	1
lunch	5	1	1	1	1	2	1

Add-One Smoothing, cont.

But now we can’ t compute probabilities simply by dividing by N, the number of words in the corpus, since we have, effectively, added words. So we need to normalize each count:

$c_i^* = (c_i + 1) \times N/(N+V)$



	I	want	to	eat	Chinese	food	lunch
I	6	740	.68	10	.68	.68	.68
want	2	.42	331	.42	3	4	3
to	3	.69	8	594	3	.69	9
eat	.37	.37	1	.37	7.4	1	20
Chinese	.36	.12	.12	.12	.12	15	.24
food	10	.48	9	.48	.48	.48	.48
lunch	1.1	.22	.22	.22	.22	.44	.22

Too Much Probability Moved to Empty Cells

Compare:

Count (*want to*) went from 787 to 331.

P(*want to*) went from 787/N (.65) to 331/(N+V) (.28)

Although the events with count = 0 are not impossible, most of them still wouldn’ t occur even in a much larger sample.

How likely is it, if we were to read more text, that the next word would cause us to see a new N-gram that we hadn’ t already seen?

Use Count of Things Seen Once

Key Concept. Things Seen Once: Use the count of things you' ve seen once to help estimate the count of things you' ve never seen.

Compute the probability that the next N-gram is a new one by counting the number of times we saw N-grams for the first time in the training corpus and dividing by the total number of events in the corpus =

T/(N + T) (T = # types; N = # tokens)

Now, to compute the probability of any particular novel N-gram, divide that total probability mass by the number of unseen N-grams:

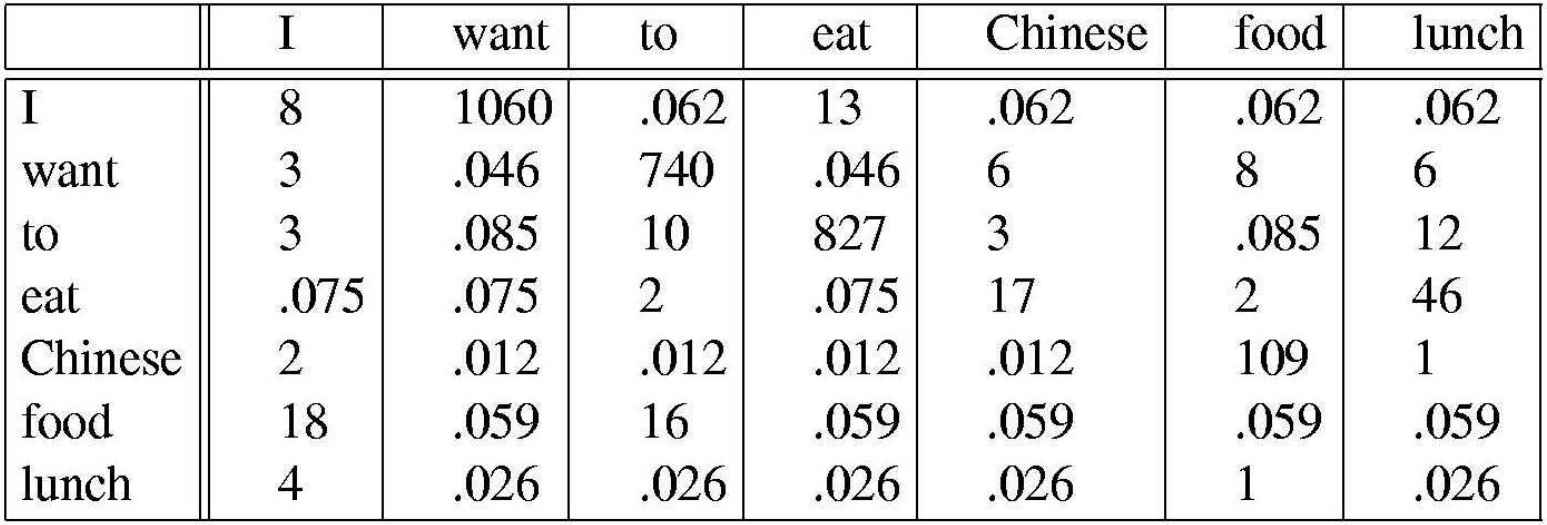
$$p_i^* = \frac{T}{Z(N+T)}$$ (Z = # of N-grams with count = 0)

Two More Issues

But we just added probability mass. It has to come from somewhere, so we need a way to discount the counts of the N-grams that did occur in the training text.

If we're using N-grams and N>1, then we want to condition the probability of a new N-gram $w_1\ w_2\ \ldots\ w_n$, by the probability of seeing $w_1\ w_2\ \ldots\ w_{n-1}$.

The Revised (Smoothed) Bigram Table



	I	want	to	eat	Chinese	food	lunch
I	8	1060	.062	13	.062	.062	.062
want	3	.046	740	.046	6	8	6
to	3	.085	10	827	3	.085	12
eat	.075	.075	2	.075	17	2	46
Chinese	2	.012	.012	.012	.012	109	1
food	18	.059	16	.059	.059	.059	.059
lunch	4	.026	.026	.026	.026	1	.026

Gathering linguistic evidence by corpus annotation

- Collections of written and spoken texts (CORPORA) useful
  - As sources of examples (more confidence that one hasn’ t forgotten some crucial data)
  - To gather statistics
  - To evaluate one’ s system (especially if ANNOTATED)
  - To train machine learning algorithms (SUPERVISED and UNSUPERVISED)

Issues in corpus construction & analysis

- Corpus construction as a scientific experiment:
  - Ensuring the corpus is an appropriate SAMPLE
  - Ensuring the annotation is done RELIABLY
    - Addressing the problem of AMBIGUITY and OVERLAP
- Corpus construction as resource building:
  - Finding the appropriate MARKUP METHOD
    - Makes REUSE & EXCHANGE easy
    - As corpora grow larger, push towards ensuring they are going to be a resource of general use

Corpus contents

- Language type
  - Text:
    - Edited: articles, books, newswires
    - Spontaneous: Usenet
  - Speech:
    - Spontaneous: Switchboard
    - Task-oriented: ATIS, MapTask
- Genre
  - Fiction, non-fiction

Some well-known corpora

Corpus	**# Tokens**	**Comments**
Brown	1 000 000	Tagged, balanced
Susanne	120 000	Parsed subset of Brown
LOB	1 000 000	UK's response to Brown
Penn Treebank	2 000 000	Parsed
MapTask	150 000	Spoken dialogue, parsed, dialogue acts
British National Corpus (BNC)	100 000 000	POS tagged

Different measures of `corpus size'

- Word TOKEN count N: how big is the corpus?
- Word TYPE count: how many different words are there?
  - What is the size V of the vocabulary?
- Word type FREQUENCIES

Levels of corpus analysis

- Simple TRANSCRIPTION
- Many cases of annotation to test a specific hypothesis
- Part-of-speech tagging (e.g., Brown Corpus, BNC)
- Special tokens: names, citations
- Syntactic structures ('Treebank') (E.g., Lancaster/IBM Treebank, Penn Treebank)
- Word sense (e.g., SEMCOR)
- Dialogue acts (e.g., MAPTASK, TRAINS)
- \`Coreference' : MUC, Lancaster UCREL, GNOME

Transcription, or: what counts as a 'word' ?

- Tokenization
  - $22.50
  - George W. Bush
- Normalization
  - The / the / THE
  - Calif. / California

Markup formats

- Inline annotation of tokens (e.g., Brown)
  - John/PN left/VBP ./.
- Tabular format (e.g., Suzanne)

A12:0210	John	John	PN
A12:0211	Left	Leave	VBP
A12:0212	.	Period	PUNC

- General markup formats:
  - SGML: <W C= ‘PN’ >John <W C= ‘VBP’ >left <W C= ‘.’ >.
  - XML

Example 1: The Brown Corpus (of Standard American English)

- The first modern computer-readable corpus (Francis and Kucera, 1961)
- 500 texts, each 2,000 words long
- From American books, newspapers and magazines
- 15 genres: science fiction, romance fiction, press reportage, scientific writing
- Part of Speech (POS) tagged: 87 classes

POS Tagging in the Brown corpus

Television/NN has/HVZ yet/RB to/TO work/VB out/RP a/AT living/RBG arrangement/NN with/IN jazz/NN ,/, which/VDT comes/VBZ to/IN the/AT medium/NN more/QL as/CS an/AT uneasy/JJ guest/NN than/CS as/CS a/AT relaxed/VBN member/NN of/IN the/AT family/NN ./.

Ambiguity in POS tagging

The	AT		
man	NN	VB	
still	NN	VB	RB
saw	NN	VBD	
her	PPO	PP$	

Example II: Beyond Tagging
The Penn Treebank

- One of the first syntactically annotated corpora
- Contents (Treebank II): about 3M words
    - Brown corpus (Treebank I)
    - 1 million words from Wall Street Journal Corpus (Treebank II)
    - ATIS corpus
- More info:
    - Marcus, Santorini, and Marcinkiewicz, 1993
    - http://www.cis.upenn.edu/~treebank

The Penn Treebank (Treebank I format – ‘skeletal’)

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

Reliability

- Crucial requirement for the corpus to be of any use, is to make sure that annotation is RELIABLE (I.e., two different annotators are likely to mark in the same way)
- E.g., make sure they can agree on part-of-speech tag
  - … we walk in SNAKING lines (JJ? VBG?)
- Or on attachment
- Agreement more difficult the more complex the judgments asked of the annotators
  - E.g., on givenness status
- Often a detailed ANNOTATION MANUAL required
- Task must also have to be simplified

Coding Instructions

- In order to achieve a reliable coding, it is necessary to tell the annotators what to do in case of problems
- Example I: the Gundel Zacharski and Hedberg coding protocol for givenness status
- Example II: the Poesio & Vieira coding instructions for definite type

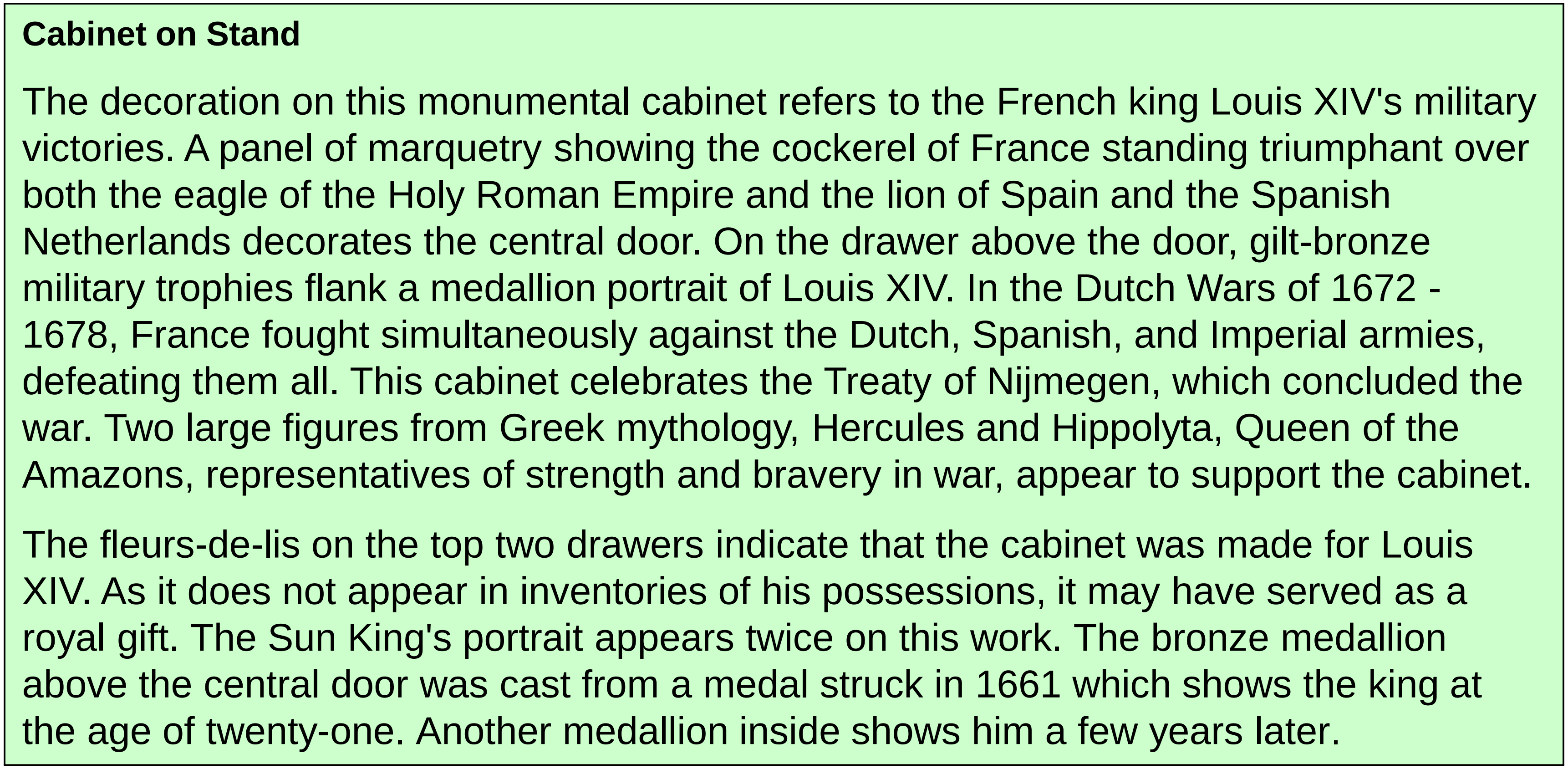
A measure of agreement: the K statistic

- Carletta, 1996: in order for the statistics extracted from an annotation to be reproducible, it is crucial to ensure that the coding distinctions are understandable to someone other than the person who developed the scheme
- Simply measuring the percentage of agreement does not take chance agreement into account
- The K statistic (Siegel and Castellan, 1988):
  - K=0: no agreement
  - .6 <= K < .8: tentative agreement
  - .8 <= K <= 1: OK agreement

Example III - Annotating referring expressions: the GNOME corpus

- Primary goal: studying the effect of salience on nominal expression generation
- Collected at the University of Edinburgh, HCRC
- 3 Genres (about 3000 NPs in each genre)
  - Descriptions of museum pages (including the ILEX/SOLE corpus)
  - ICONOCLAST corpus (500 pharmaceutical leaflets)
  - Tutorial dialogues from the SHERLOCK corpus

An example GNOME text



Cabinet on Stand

The decoration on this monumental cabinet refers to the French king Louis XIV's military victories. A panel of marquetry showing the cockerel of France standing triumphant over both the eagle of the Holy Roman Empire and the lion of Spain and the Spanish Netherlands decorates the central door. On the drawer above the door, gilt-bronze military trophies flank a medallion portrait of Louis XIV. In the Dutch Wars of 1672 - 1678, France fought simultaneously against the Dutch, Spanish, and Imperial armies, defeating them all. This cabinet celebrates the Treaty of Nijmegen, which concluded the war. Two large figures from Greek mythology, Hercules and Hippolyta, Queen of the Amazons, representatives of strength and bravery in war, appear to support the cabinet.

The fleurs-de-lis on the top two drawers indicate that the cabinet was made for Louis XIV. As it does not appear in inventories of his possessions, it may have served as a royal gift. The Sun King's portrait appears twice on this work. The bronze medallion above the central door was cast from a medal struck in 1661 which shows the king at the age of twenty-one. Another medallion inside shows him a few years later.

Annotating referring expressions: the GNOME corpus

- Syntactic features: grammatical function, agreement
- Semantic features:
  - Logical form type (term / quantifier / predicate)
  - `Structure': Mass / count, Atom / Set
  - Ontological status: abstract / concrete, animate
  - Genericity
  - 'Semantic' uniqueness (Loebner, 1985)
- Discourse features:
  - Deixis
  - Familiarity (discourse new / inferrable / discourse old) (using anaphoric annotation)
  - Is the entity the current CB (computed)

Agreement on NE attributes

NP Type	.9
Agreement	.9
Gramm Function	.85
Animacy	.81
Deix	.81

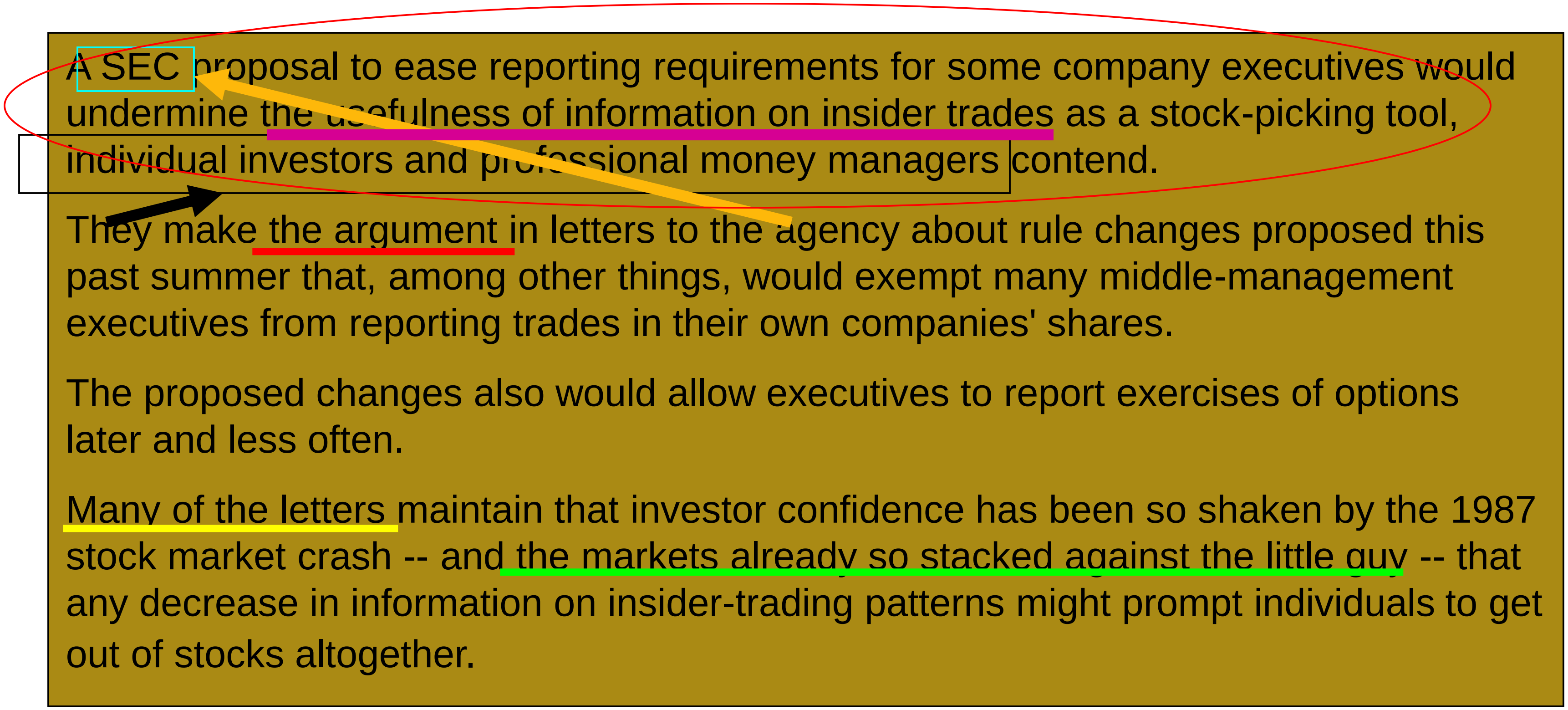
Some problems in classifying referring expressions

- Reference to kind / to specific instance
  - *the interiors of this coffer are lined with tortoise shell and brass or pewter*
- Objects which are difficult to analyze:
  - Abstract terms:
    - *… each decorated using a technique known as premiere partie marquetry, a pattern of brass and pewter on a tortoiseshell ground …*
  - Attributes:
    - *the age of four years*

Problematic attributes

Genericity	.89 (but only after many trials)
‘Loebner’ (functionality)	.82 (same)
CB	.6
Thematic role	.42
Topic	.375

The annotation of context dependence (`coreference’ and other things)



A SEC proposal to ease reporting requirements for some company executives would undermine the usefulness of information on insider trades as a stock-picking tool, individual investors and professional money managers contend.

They make the argument in letters to the agency about rule changes proposed this past summer that, among other things, would exempt many middle-management executives from reporting trades in their own companies' shares.

The proposed changes also would allow executives to report exercises of options later and less often.

Many of the letters maintain that investor confidence has been so shaken by the 1987 stock market crash -- and the markets already so stacked against the little guy -- that any decrease in information on insider-trading patterns might prompt individuals to get out of stocks altogether.

Issues in annotating context dependence

- Which markables?
  - Only anaphoric relations between entities realized as NPs?
  - Also when antecedent is not realized by NP?
  - Also when anaphoric expression not NP? (E.g., ellipsis)
- Only `anaphoric' ? Only `coreference' ?
- How many relations?
- Do you need the antecedent?

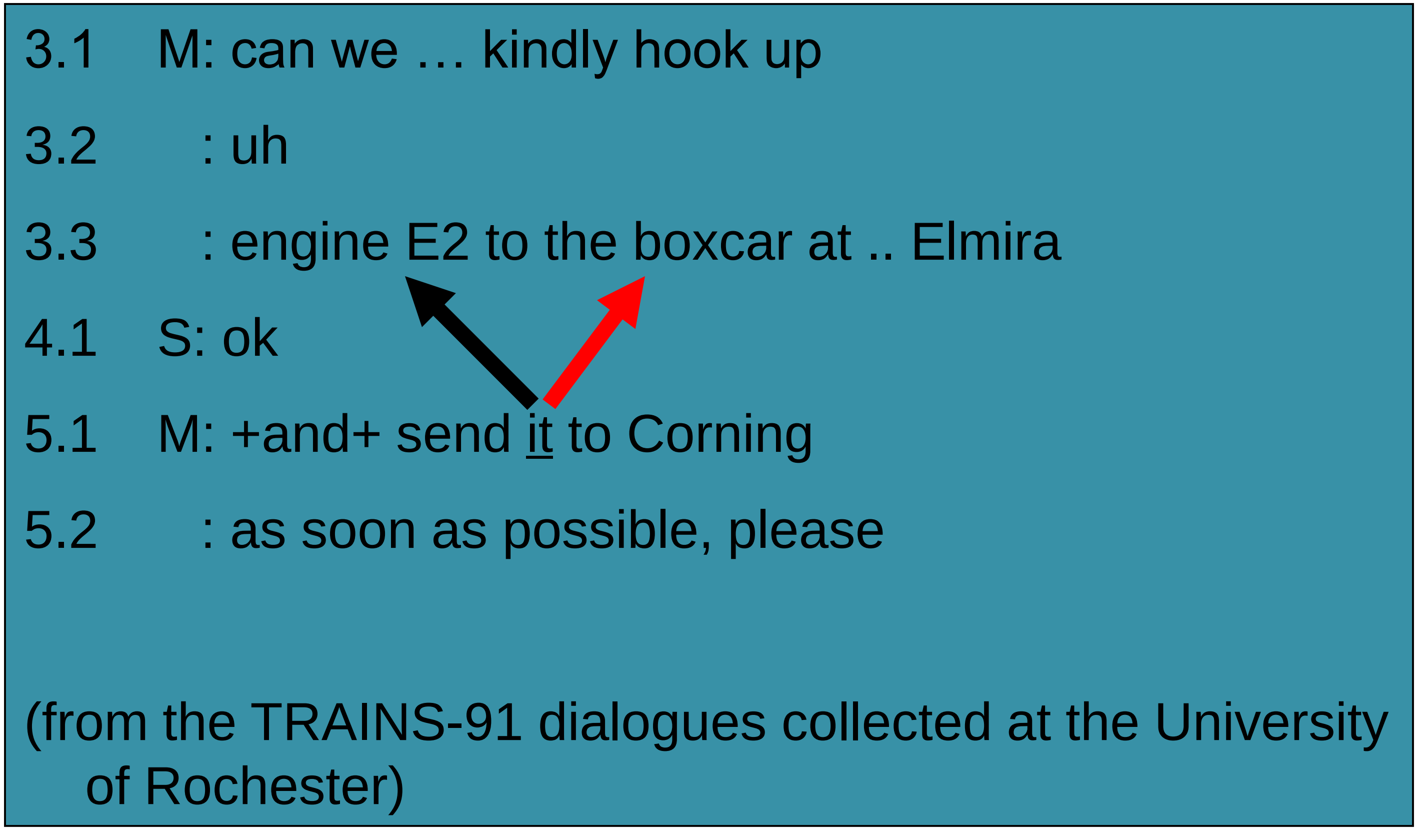
What is the annotation for?

- For ‘higher level’ annotation, having a clear goal (scientific or engineering) is essential
- Uses of coreference annotation:
  - To study a certain discourse phenomenon (e.g., Centering theory)
  - To test an anaphora resolution system (e.g., a pronominal resolver)
  - For a particular application: information extraction (e.g., MUC), summarization, question-answering

Markables

- Only NPs?
  - Clitics?
    - *A:Adesso dammelo. [Now give-to me-it]*
  - Traces?
    - *A: _ Sta arrivando. [He/She is on her/his way]*
- All NPs?
  - Appositions:
    - *one of engines at Elmira, say engine E2*
    - *The Admiral's Head, that famous Portsmouth hostelry*
  - Predicative NPs:
    - *John is the president of the board*

Identifying antecedents: Ambiguous anaphoric expressions



3.1 M: can we … kindly hook up
3.2 : uh
3.3 : engine E2 to the boxcar at .. Elmira
4.1 S: ok
5.1 M: +and+ send it to Corning
5.2 : as soon as possible, please

(from the TRAINS-91 dialogues collected at the University of Rochester)

Disagreements on anaphora (Poesio and Vieira, 1998)

About 160 workers at *a factory* that made paper for the Kent filters were exposed to asbestos in the 1950s.

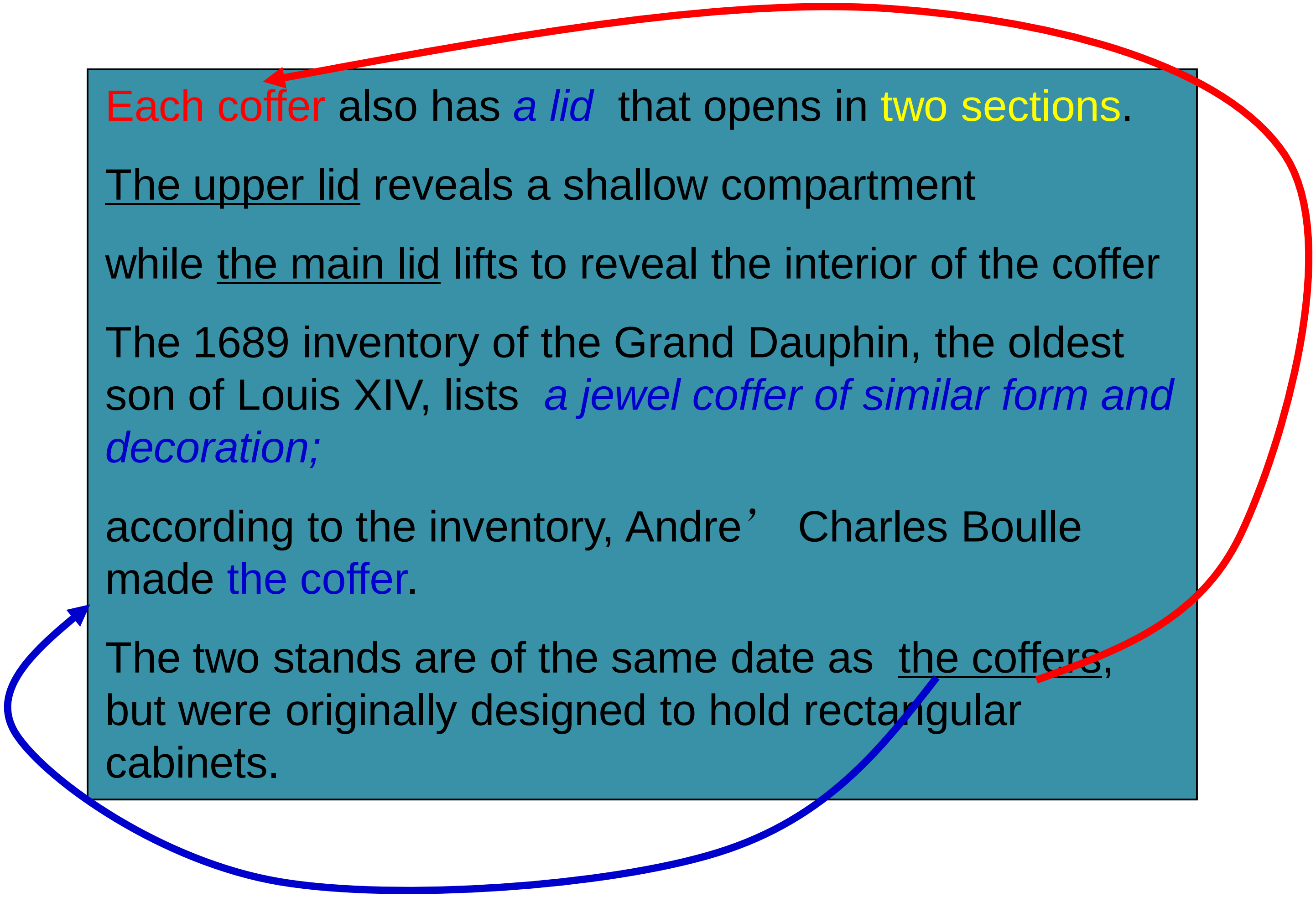
Areas of the factory were particularly dusty where the crocidolite was used.

Workers dumped large burlap sacks of the imported material into a huge bin, poured in cotton and acetate fibers and mechanically mixed the dry fibers in a process used to make filters.

Workers described "clouds of blue dust" that hung over *parts of the factory*,

even though exhaust fans ventilated the area.

Identifying antecedents: complex anaphoric relations



Each coffer also has *a lid* that opens in two sections.

The upper lid reveals a shallow compartment

while the main lid lifts to reveal the interior of the coffer

The 1689 inventory of the Grand Dauphin, the oldest son of Louis XIV, lists *a jewel coffer of similar form and decoration;*

according to the inventory, Andre’ Charles Boulle made the coffer.

The two stands are of the same date as the coffers, but were originally designed to hold rectangular cabinets.

Deictic references

FOLLOWER: Uh-huh. Curve round. To your right.
GIVER: Uh-huh.
FOLLOWER: Right.... Right underneath the diamond mine.
Where do I stop.
GIVER: Well....... Do. Have you got a graveyard?
Sort of in the middle of the page? ... On on a level to
the c-- ... er diamond mine.
FOLLOWER: No. I've got a fast running creek.
GIVER: A fast flowing river,... eh.
FOLLOWER: No. Where's that . Mmhmm,... eh. Canoes

The GNOME annotation manual: Markables

- ONLY ANAPHORIC RELATIONS BETWEEN NPs
- DETAILED INSTRUCTIONS FOR MARKABLES
  - ALL NPs are treated as markables, including predicative NPs and expletives (use attributes to identify non-referring expressions)

Achieving agreement (but not completeness) in GNOME

- RESTRICTING THE NUMBER OF RELATIONS
  - IDENT (*John … he, the car … the vehicle*)
  - ELEMENT (*Three boys … one (of them)*)
  - SUBSET (*The vases … two (of them) …*)
  - Generalized POSSession (*the car … the engine*)
  - OTHER (when no other connection with previous unit)

Limiting the amount of work

- Restrict the extent of the annotation:
  - ALWAYS MARK AT LEAST ONE ANTECEDENT FOR EACH EXPRESSION THAT IS ANAPHORIC IN SOME SENSE, BUT NO MORE THAN ONE IDENT AND ONE BRIDGE;
  - ALWAYS MARK THE RELATION WITH THE CLOSEST PREVIOUS ANTECEDENT OF EACH TYPE;
  - ALWAYS MARK AN IDENTITY RELATION IF THERE IS ONE; BUT MARK AT MOST ONE BRIDGING RELATION

Agreement results

- RESULTS (2 annotators, anaphoric relations for 200 NPs)
  - Only 4.8% disagreements
  - But 73.17% of relations marked by only one annotator
- The GNOME annotation scheme:
  - http://www.hcrc.ed.ac.uk/~poesio/GNOME/anno_manual_4.html

A standard markup format: SGML/XML

- Early annotations all used different markup methods
- SGML developed as a universal format
  - No need of special software to deal with the way info is marked up
- XML a simplified version
  - end tags required
  - standard format for attributes

XML Basics

```
<p>
    <s> And then John left . </s>
    <s> He did not say another word</s>
</p>
```

```
<utt speaker= “Fred”  date= “10-Feb-1998” >
    That is an ugly couch.
</utt>
```

Words in XML

```
<!DOCTYPE SYSTEM  “words.dtd” >
<words>
    <word id= “w1” >turn</word>
    <word id= “w2” >right</word>
    <word id= “w3” >for</word>
    <word id= “w4” >three</word>
    <word id= “w5” >centimetres</word>
    <word id= “w6” >okay</word>
</words>
```

The DTD (for the words level)

```
<!ELEMENT words (word*)>

<!ELEMENT word (#PCDATA)>

<!ATTLIST word id ID #REQUIRED>

<!ATTLIST word starttime CDATA #IMPLIED>

<!ATTLIST word endtime CDATA #IMPLIED>
```

The GNOME example, again

Cabinet on Stand

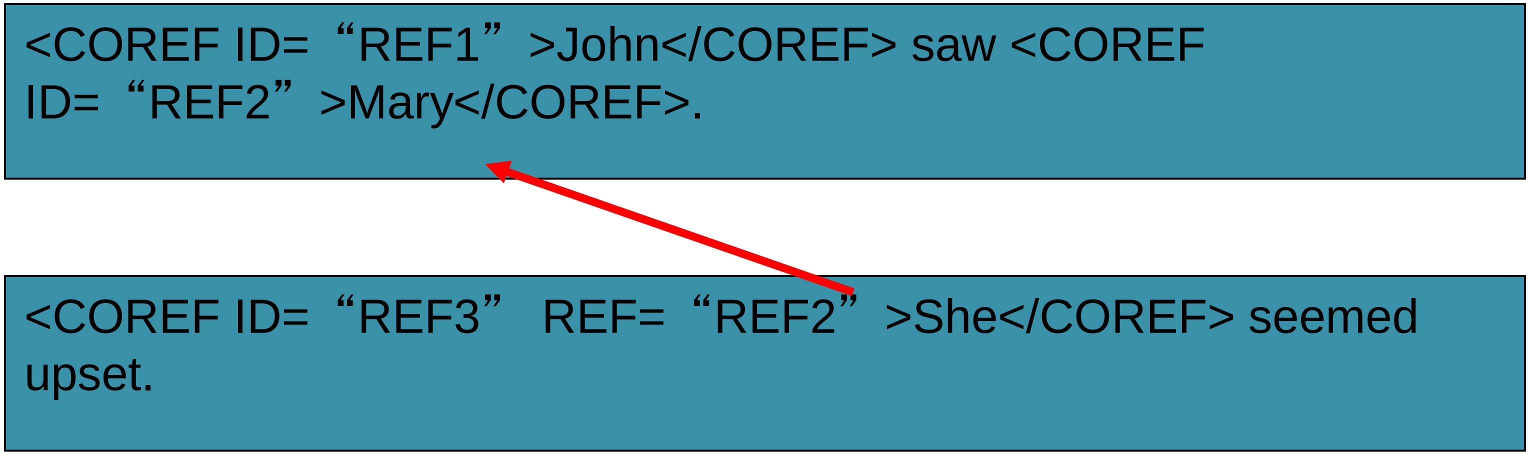
The decoration on this monumental cabinet refers to the French king Louis XIV's military victories. A panel of marquetry showing the cockerel of France standing triumphant over both the eagle of the Holy Roman Empire and the lion of Spain and the Spanish Netherlands decorates the central door. On the drawer above the door, gilt-bronze military trophies flank a medallion portrait of Louis XIV. In the Dutch Wars of 1672 - 1678, France fought simultaneously against the Dutch, Spanish, and Imperial armies, defeating them all. This cabinet celebrates the Treaty of Nijmegen, which concluded the war. Two large figures from Greek mythology, Hercules and Hippolyta, Queen of the Amazons, representatives of strength and bravery in war, appear to support the cabinet.

The fleurs-de-lis on the top two drawers indicate that the cabinet was made for Louis XIV. As it does not appear in inventories of his possessions, it may have served as a royal gift. The Sun King's portrait appears twice on this work. The bronze medallion above the central door was cast from a medal struck in 1661 which shows the king at the age of twenty-one. Another medallion inside shows him a few years later.

The GNOME NE annotation in XML format

```
<ne id="ne109"
cat="this-np" per="per3" num="sing" gen="neut  “ gf="np-mod"
lftype="term" onto="concrete  “ ani="inanimate"
structure="atom" count="count-yes" generic="generic-no  “deix="deix-
yes" reference="direct" loeb="disc-function" >  this  monumental
cabinet </ne>
```

Coreference in XML: MUC (Hirschman, 1997)



<COREF ID= “REF1” >John</COREF> saw <COREF ID= “REF2” >Mary</COREF>.

<COREF ID= “REF3” REF= “REF2” >She</COREF> seemed upset.

Problems with the MUC scheme

- Markup issues:
  - Only one type of anaphoric relation
  - No way of marking ambiguous cases
- Notion of 'coreference' used dubious (see van Deemter and Kibble, 2001)

The MATE/GNOME Markup Scheme

```
<NE ID= “ne07” >Scottish-born, Canadian based jeweller, Alison Bailey-Smith</NE>
<NE ID= “ne08” > <NE ID= “ne09” >Her</NE> materials</NE>
```

```
<ANTE  CURRENT= “ne09”  REL= “ident” >
    <ANCHOR ANTECEDENT= “ne07”  />
</ANTE>
```

Ambiguous anaphoric expressions in the MATE/GNOME scheme

3.3: <NE ID= “ne01” >*engine E2*</NE> to
<NE ID= “ne02” >*the boxcar at … Elmira*</NE>

5.1: and send <NE ID= “ne03” >*it*</NE> to
<NE ID= “ne04” >*Corning*</NE>

```
<ANTE  CURRENT= “ne03”  REL= “ident” >
    <ANCHOR ANTECEDENT= “ne01”  />
    <ANCHOR ANTECEDENT= “ne02”  />
</ANTE>
```

Marking bridging relations

We gave <NE ID= “ne01” >*each of* <NE ID= “ne02” > *the boys*</NE> </NE> <NE ID= “ne03” > *a shirt*</NE>, *but* <NE ID= “ne04” > *they*</NE> *didn’t fit.*

```
<ANTE  CURRENT= “ne04”  REL= “element-inv” >
      <ANCHOR ANTECEDENT= “ne03”  />
</ANTE>
```

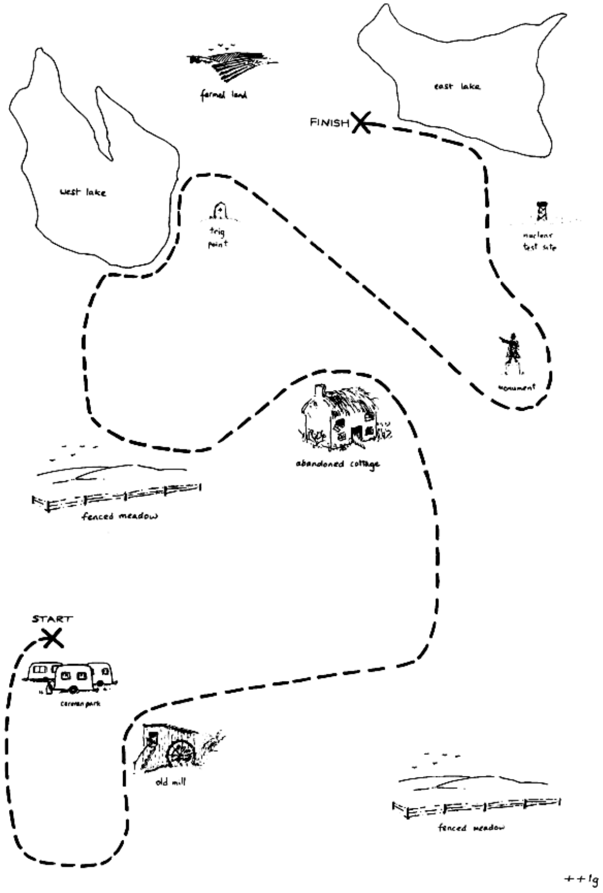
XML Standoff

- Typically will want to do multiple layers of annotation (e.g., transcription, markables, coreference)
- Want to be able to keep them independent so that
  - New levels of annotation can be added without disturbing existing ones
  - Editing one level of annotation has minimal knock-on effects on others
  - People can work on different levels at the same time without worrying about creating different versions

The HCRC MAPTASK corpus

- A collection of annotated spoken dialogues between subjects doing the Map Task
- Collected at the Universities of Edinburgh and Glasgow – 1983 first round, then in 1991
- 1991 corpus:
  - 128 dialogues, 64 eye contact, 64 No ec
  - About 15 hours of speech, 146,855 word tokens
- www.hcrc.ed.ac.uk/maptask

An example of map



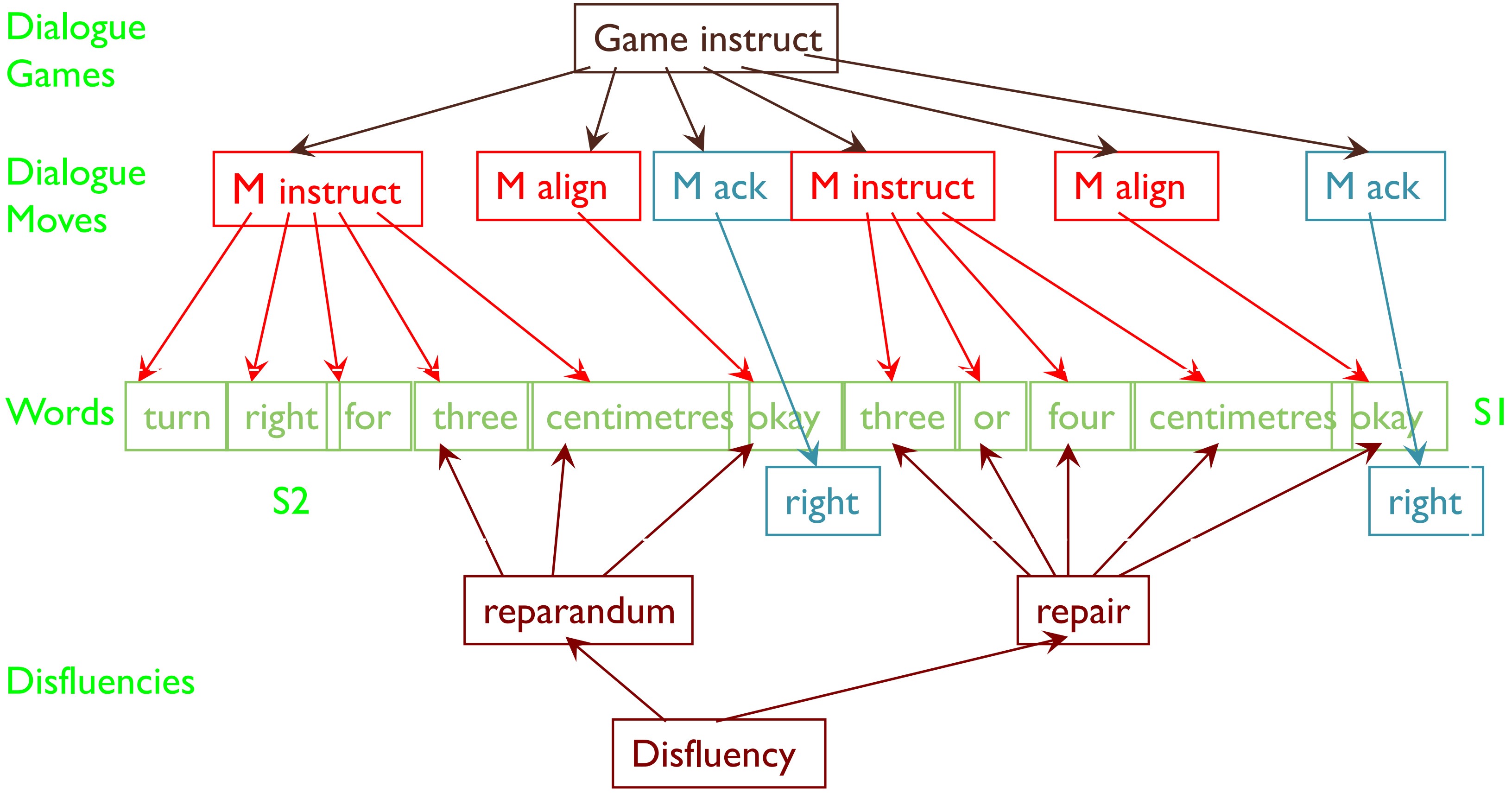
An example dialogue

GIVER: right, you got a map with an extinct volcano?
FOLLOWER: right yes i have, i'm just in front of that.
GIVER: right.
FOLLOWER: with the start.
GIVER: right, you've got a cross marked start?
FOLLOWER: yes.
GIVER: right, if you just want to come ... ... like down past the extinct volcano ... down to like to towards the bottom of the page.
FOLLOWER: right okay, just straight down directly south?
GIVER: uh-huh ... just straight down, uh south.
FOLLOWER: how far?

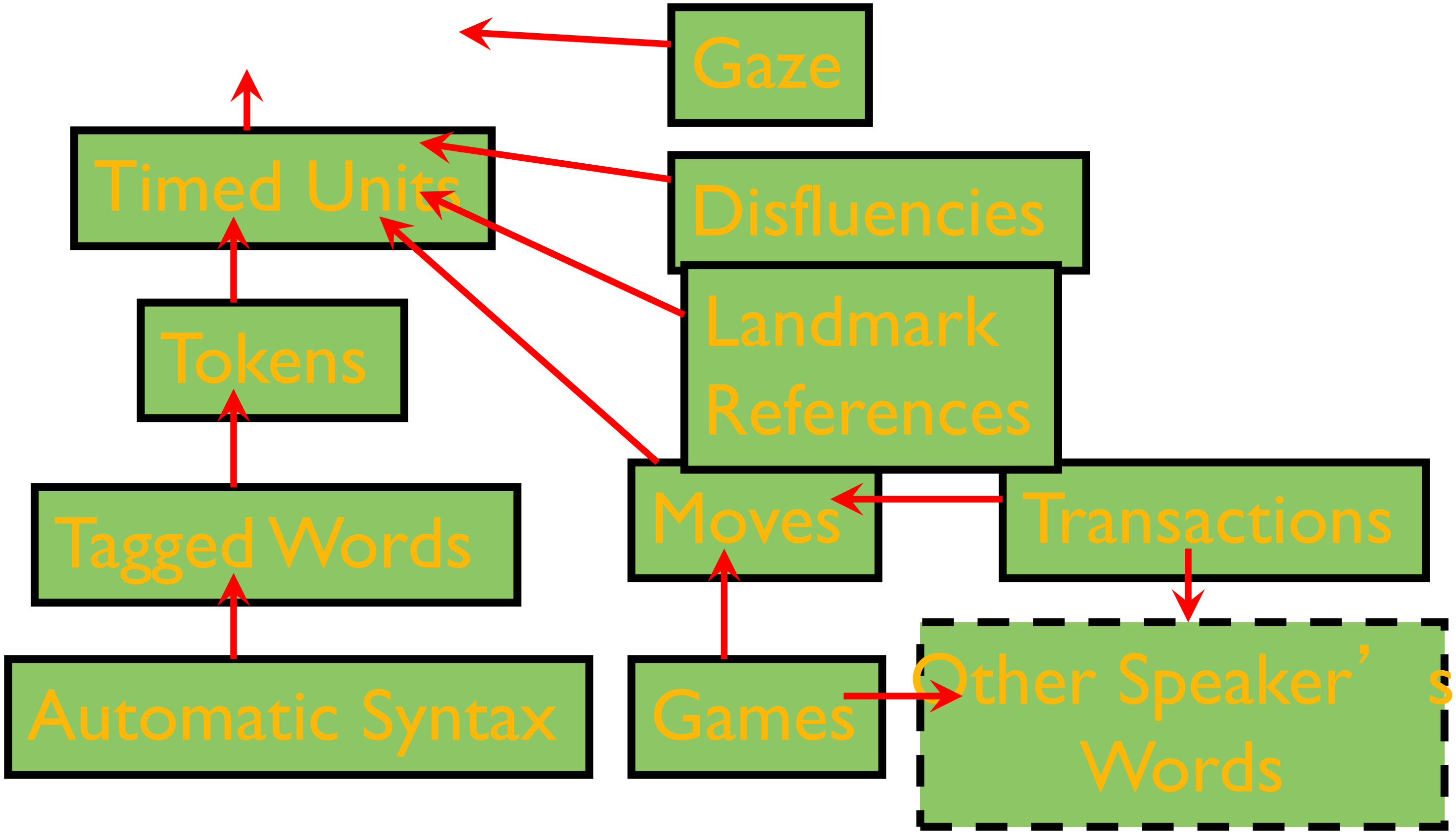
An Italian MapTask: IPAR

F008: okay [straniero] si" l<ll> l<ll> da qui e" il punto di partenza e" il viale della ve+ <esit> della felicita “
G009: <eh> si"
<pb>
F010: quindi poi ?
G011: diciamo<oo> <ehm> allora guardando la mappa tu ce l'hai<ii> a sinistra la partenza , no ?
F012: si “
G013: di viale della felicita" <inspirazione> , okay [straniero] ?
F014: si" <RUMORE>
G015: <inspirazione> allora vai<ii> avant+ <eh> con la penna quindi
F016: <mm>
G017: vai<ii> avanti

Multiple levels of annotation in the MAPTASK corpus



Standoff annotation in the MAPTASK corpus



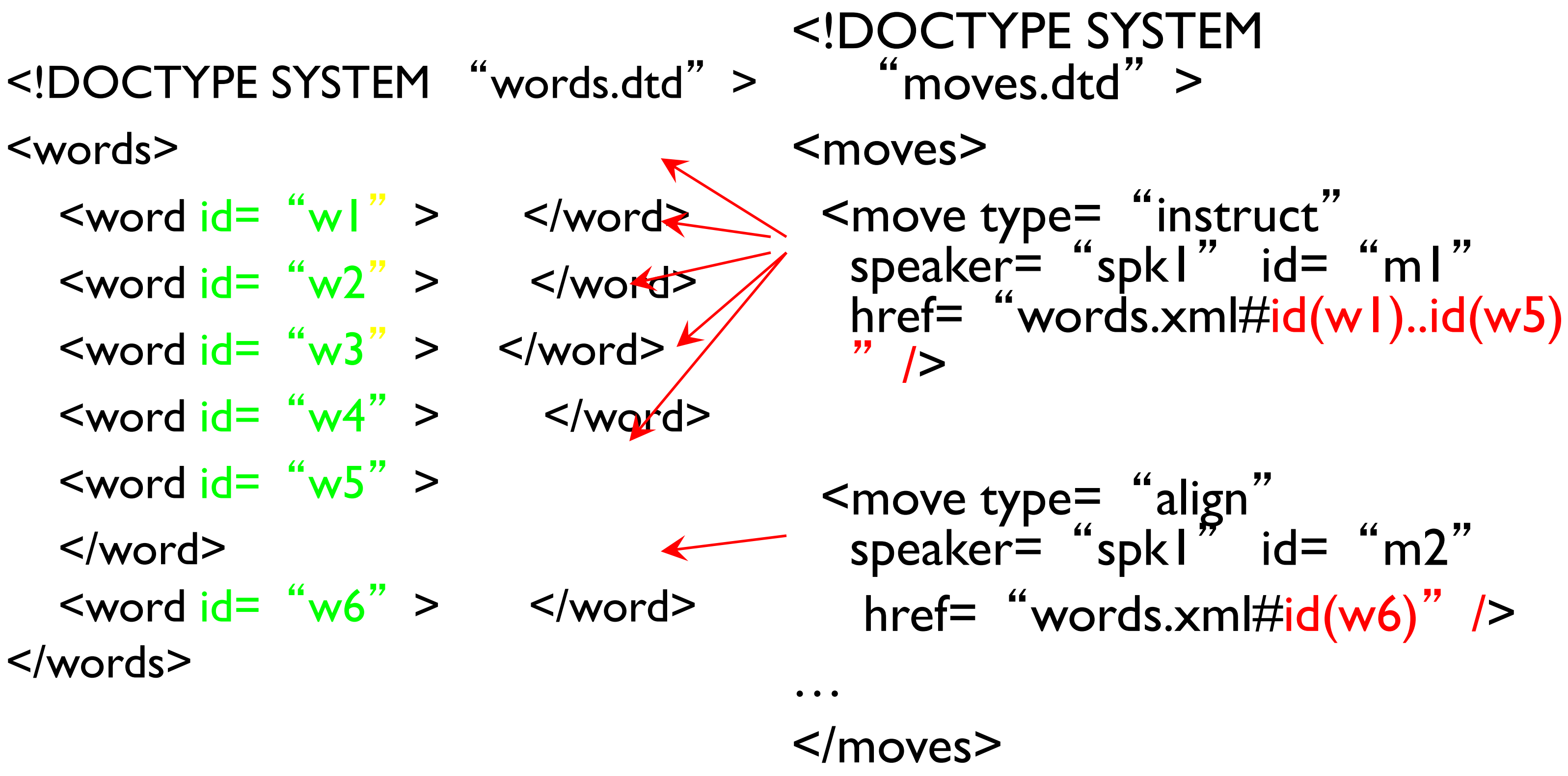
Standoff Example (1): Words XML

```
<!DOCTYPE SYSTEM “words.dtd” >
<words>
    <word id= “w1” >turn</word>
    <word id= “w2” >right</word>
    <word id= “w3” >for</word>
    <word id= “w4” >three</word>
    <word id= “w5” >centimetres</word>
    <word id= “w6” >okay</word>
</words>
```

Standoff Example (2): Moves XML

```
<!DOCTYPE SYSTEM “moves.dtd” >
<moves>
  <move type= “instruct” speaker= “spk1” id= “m1”
     href= “words.xml#id(w1)..id(w5)” />
  <move type= “align” speaker= “spk1” id= “m2”
     href= “words.xml#id(w6)” />
…
</moves>
```

Standoff Example (3): Moves and Words XML



```
<!DOCTYPE SYSTEM  “words.dtd” >
<words>
   <word id= “w1” >      </word>
   <word id= “w2” >      </word>
   <word id= “w3” >    </word>
   <word id= “w4” >       </word>
   <word id= “w5” >
   </word>
   <word id= “w6” >      </word>
</words>
```

```
<!DOCTYPE SYSTEM
     “moves.dtd” >
<moves>
  <move type= “instruct”
    speaker= “spk1”  id= “m1”
    href= “words.xml#id(w1)..id(w5)
    ”  />

  <move type= “align”
    speaker= “spk1”  id= “m2”
     href= “words.xml#id(w6)”  />
…
</moves>
```

Other corpora annotated for anaphoric information (in English)

- The UCREL/IBM corpus (not freely available)
- The Wolverhampton corpus (from the Wolverhampton CL group website)
  - only pronominal anaphora
- The Ge/Charniak corpus (ask Ge or Charniak @ Brown)
  - only pronominal anaphora