SO... COMPUTERS HAVE MASTERED PLAYING CHESS AND DRIVING CARS ACROSS THE DESERT, BUT CAN'T HOLD FIVE MINUTES OF NORMAL CONVERSATION?
PRETTY MUCH.

Source: xkcd

NATURAL LANGUAGE PROCESSING

amitabha mukerjee
iit kanpur

The magic of language

“मोहल्ले का एक लड़का”

“A monkey came in through the window and ate up my lunch.”

The magic of language

- Language is about conveying meaning
- Language is one-dimensional – Meaning is multi-dimensional

- Challenges
  Sounds along one-dimension express multi-dimensional aspects of reality
  - Same sounds map to different meanings [**Polysemy**]
  - Same meanings map to different sounds [**Synonymy**]

Myths about language

- **grammar** is about whether language is correct or incorrect

  It’s me.

  Ganesh is at home?

  There are many small-small holes in this dress.

- Modern view: grammar is about usage
  - descriptive, not prescriptive

Myths about language

- **grammar** is about the correct and incorrectness of language.

  Ganesh is at home? → *Is Ganesh at home?*

  It's me (accusative) → *"It's I"*

  There are many small-small holes in this dress.

- words are separated by spaces.

- how many sounds are there in English? 26

Myths about language

- **grammar** is about the correct and incorrectness of language.

  Ganesh is at home? → *Is Ganesh at home?*

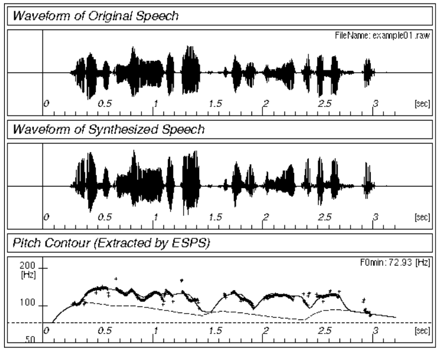
  It's me (accusative) → *"It's I"*

  There are many small-small holes in this dress.

- words are separated by spaces.
  - **words** = meaningful bits of sound
- alphabets are **not** the sounds of language

Levels of Linguistic Analysis

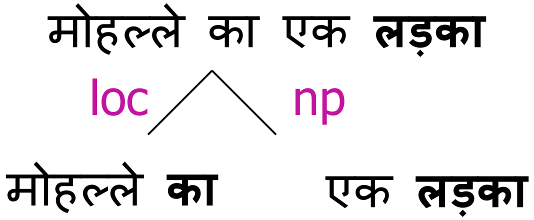
Phonology



⇔ /mohallekaeklaRkA/

Morphology /mohallekaeklaRkA/ ⇔ मोहल्ले का एक लड़का

Syntax mohalle ka ek laRkA


मोहल्ले का एक **लड़का**

loc np

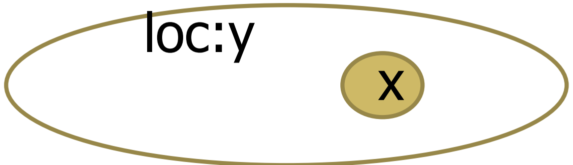
मोहल्ले **का** एक **लड़का**

Semantics

Boolean Logic:

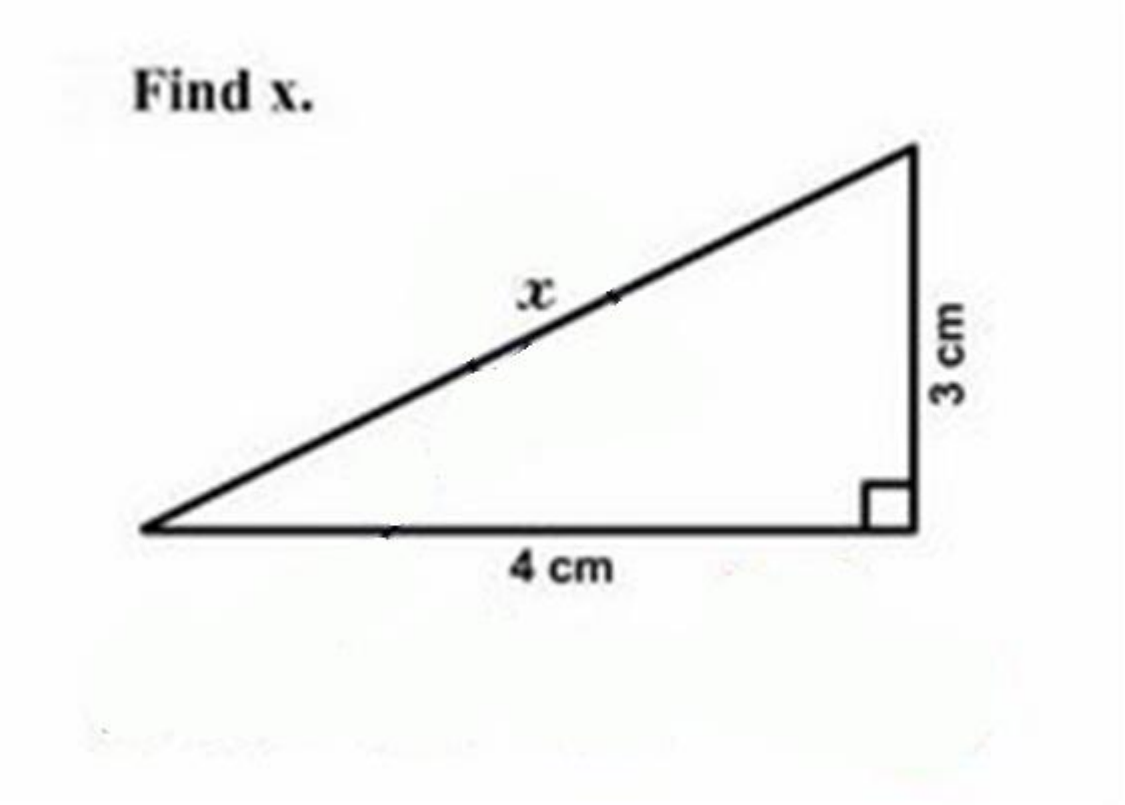
∃x ∃y boy(x) ^ loc(y)^ lives-at(x,y)]

Alternate: Imagistic


loc:y

x

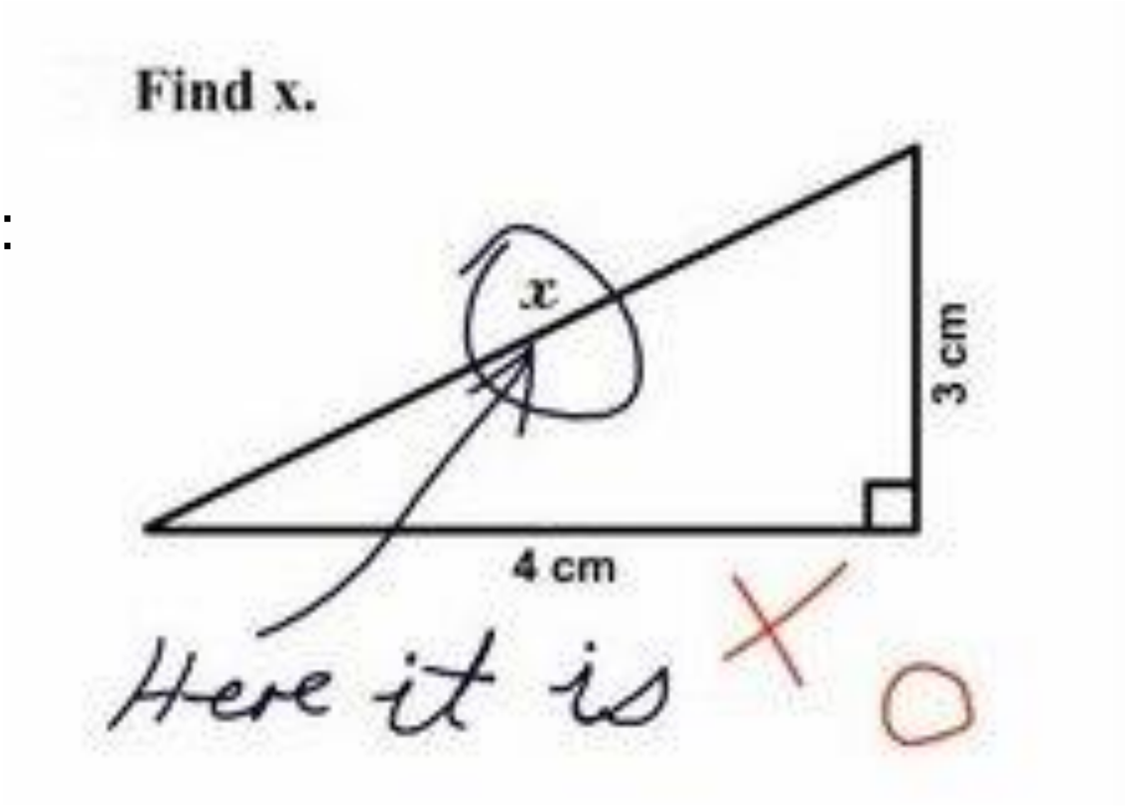
Semantics vs Pragmatics:



Pragmatics: Direct vs Indirect meaning

Traditional thinking:

- Semantics
  - Direct meaning
- Pragmatics
  - Indirect meaning



Pragmatics

- You can't hold two watermelons in one hand

    - Iranian proverb

Pragmatics: Meaning in Context

Traditional levels of analysis:

- **Semantics**: composition from lexical meaning of words [*direct meaning*]
- **Pragmatics**: social / contextual meaning ; [*indirect meaning*]

Psycholinguists:

Retrieval of pragmatic meaning is often faster

LEVELS OF STRUCTURE IN LANGUAGE

Language Structure: Levels

boys like girls

Language Structure: Levels

- **Phonology**
- **Lexicon**
- **Syntax [Morphology]**
- **Discourse**

- **Semantics / Compositionality**
- **Pragmatics / Discourse**

Language Structure: Levels

- **Phonology :** sounds of speech

  phoneme /b/ /oy/ /z/
- **Lexicon :** set of meaning-bearing units, **lexemes**
- **Syntax :** composing lexemes **composition**
  - **Word =** base + affixes / suffixes
  - **Phrase =** [[[boys] like] girls]
- **Discourse :** Boy likes girl. They meet.

NLP: Goals

NL Understanding

Language → NLP → Decision

NL Translation

Language 1 → Machine Translation → Language 2

NL description (Generation)

Situation → NLP → Language

Phonology

Phonemes

- Which sounds change a meaning?

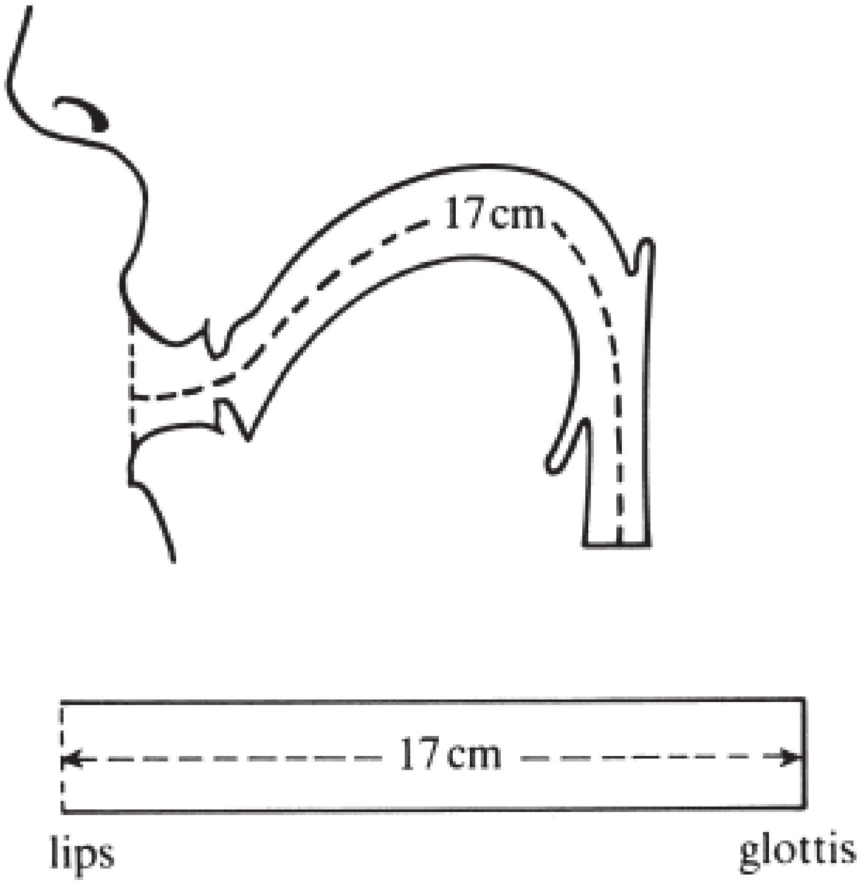
pin, tin, kin, fin, thin, sin, shin

dim, din, ding, did, dig, dish

pin, pen, pan, pun, pain, pine, pawn

- **Phoneme =** minimum distinction in sound that changes meaning
- Phonemes at middle of syllable: **vowel**
  start or end: **consonant**

Vocal organs



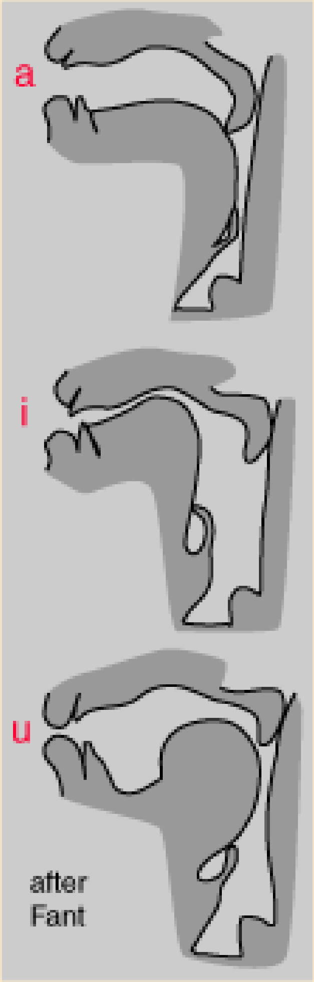
tube model of vocal tract (for most neutral vowel)

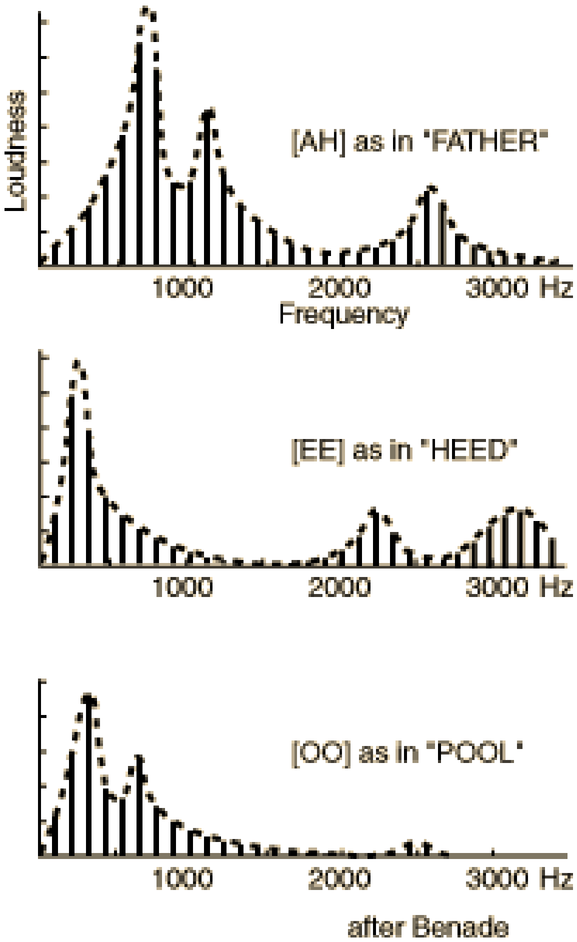
[malmkjaer 02]

Vowels : Formants

formant frequencies:
peaks in the harmonic spectrum of vowel sounds

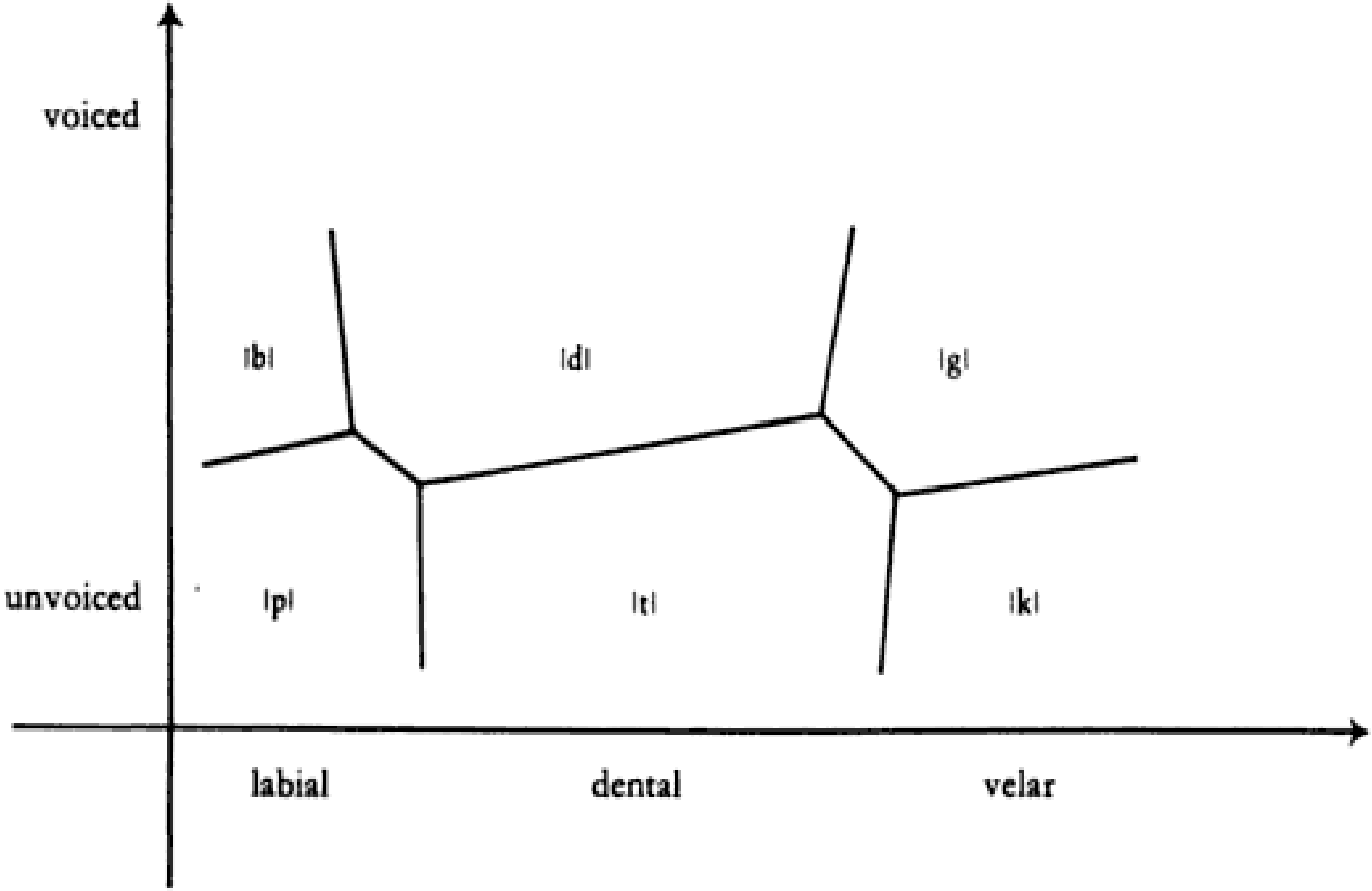
first three:
F1, F2, F3





http://hyperphysics.phy-astr.gsu.edu/hbase/music/vowel.html

Partitioning the speech sound space



[petitot 1989], [gardenfors 00]

Writing : Consonants

stop consonants

voiceless		voiced		nasal	
inaspirate	aspirated	in-	aspirated		
क	ख	ग	घ	ङ	[velar]
च	छ	ज	झ	ञ	[palatal]
ट	ठ	ड	ढ	ण	[retroflex]
त	थ	द	ध	न	[dental]
प	फ	ब	भ	म	[labial]

Consonants

stop consonants

voiceless		voiced		nasal	
inaspirate	aspirated	in-	aspirated		
k	kh	g	gh	N	[velar]
c	chh	j[dz]	jh[dzh]	n~	[palatal]
T	Th	D	Dh	N	[retroflex]
t	th	d	dh	n	[dental]
p	ph	b	bh	m	[labial] (bilabial)

Grammar of Phonology

“cats” → “cat” + /s/

“boys” → “boy” + /z/

WHY ISN'T THE PLURAL OF SMURF SMURVES?
ONE HOUSE, TWO HICE.
For some reason, nobody wants to talk to us!

Source: urbanblah

Morphology and syntax

Syntax (morphosyntax)

- Regularity in how larger structures are assembled from units or smaller structures

- **morphology**

  cook-er / read-er / *-ercook

- **phrase syntax**

  smart woman / *woman smart

- **sentence syntax**

  boys like girls / girls like boys / *like boys girls

Lexicon vs Grammar

- Assumption:
  larger structures are assembled from smaller ones

- *Q. Is this assumption valid?*

- Smallest meaning-bearing structures = unit

- **morpheme :** less likely to appear independently

  -er , -s, -ly, -able

- **lexeme**

  cat, boy, smart, undergraduate student, cook, cooker

Lexicon vs Grammar

- lexicon = mental inventory of units

  = set of all lexemes

- Is “cats” a lexeme?

cook → cooks : grammatical (rule-driven, inflection)

→ cooker : cook + er (not fully a rule; derivation)

Older thinking : lexicon is separate from grammar

at present : lexicon - grammar is a continuum

Syntax vs Morphology

- **Syntax :** how words can be assembled into phrases / sentences:
  - *I found an unopened bottle of wine*
  - ** I found a bottle unopened of wine*
- **Morphology:** internal form of words
  - *unopened* – not **openuned* or any other order
- But this distinction is not crisp (since notion of "morpheme" or "word" is graded) → **Morphosyntax**

Morpheme examples

- निरीक्षक = नि- [रीक्ष] -क
- prefix suffix

- bound / free morphemes:
  -क vs -कर्ता (e.g. अपहरणकर्ता)

- Morphemes often cause changes to the stem
  - bAngla: kin- , buy

Ami kinlAm	uni kenen	kenAkATA
I buy+PAST	he (honorific) buy+PRES	buying (noun)

Stemming (baby lemmatization)

- Assumption : surface form = root . affix

- Reduce a word to the main morpheme

automate *automates* *automatic* *automation*	→	*automat*
run *runs* *running*	→	*run*

- Widely used in Information Retrieval

Porter Stemmer (1980)

- Most common algorithm for stemming English
  - Results suggest it’s at least as good as other stemming options
- Multiple sequential phases of reductions using rules, e.g.
  - sses → ss
  - ies → i
  - ational → ate
  - tional → tion
- http://tartarus.org/~martin/PorterStemmer/

Stemming example

Candidate = candid + ate

This is a poorly constructed example using the Porter stemmer.

This is a poorli construct example us the Porter stemmer.

http://maya.cs.depaul.edu/~classes/ds575/porter.html
Code:
http://snowball.tartarus.org/algorithms/english/stemmer.html

Inflection vs Derivation

Inflections and Derivations

- **Inflection:** e.g. *sing* → *sang* ; *cat* → *cats*

  variation in form due to tense, person, etc.

  - does not change primary meaning,
  - same part-of-speech
  - applies to nearly entire class of lexemes

- **Derivation:** e.g. *sing* → *singer*

  changes meaning, changes part-of-speech

- Like much in grammar, not very crisp distinction

  e.g *cyclic* → *cyclical* = derivation

- treat as new word

Productive Morphemes

- A morpheme is productive if it applies to all words of a given type.
- Inflections – almost fully productive
- Derivations – very limited

I LIKE TO VERB WORDS.
WHAT?
I TAKE NOUNS AND ADJECTIVES AND USE THEM AS VERBS. REMEMBER WHEN "ACCESS" WAS A THING? NOW IT'S SOMETHING YOU DO. IT GOT VERBED.
VERBING WEIRDS LANGUAGE.
MAYBE WE CAN EVENTUALLY MAKE LANGUAGE A COMPLETE IMPEDIMENT TO UNDERSTANDING.
WATTERSON

Semantics of morphemes

- **inflections:**

  e.g. “-ed” : past tense = events in the past

  - *The course started last week.*

  But: often does not refer to past, e.g.:

  - *I thought the course started next week.*
  - *If the course started, everyone would be pleased.*

- past time = **primary** or most common characteristic
- many other interpretations possible (in many languages)

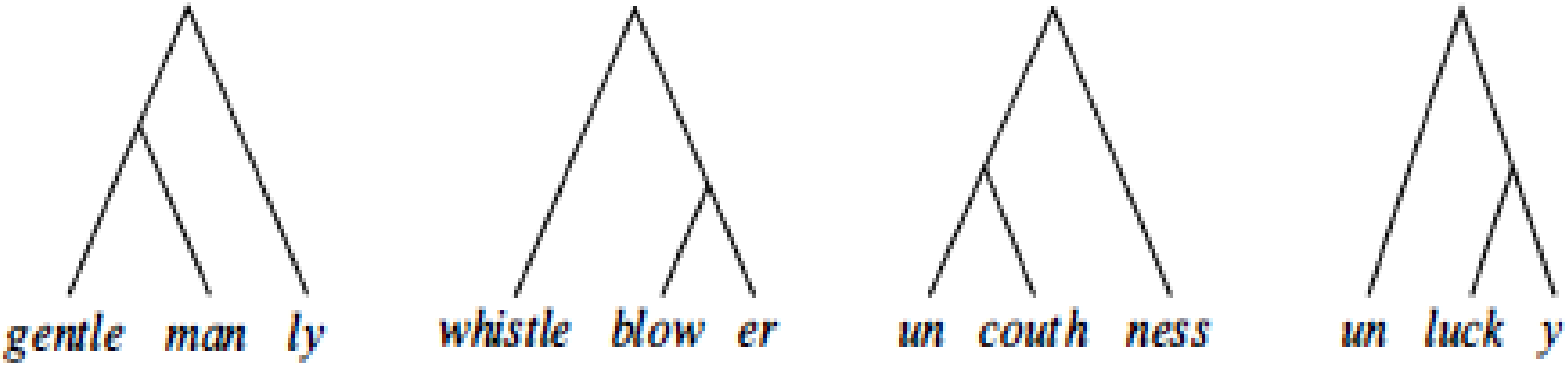
  → past tense = grammatical form, varied semantics

Derivations

- e.g. **ungentlemanly:** un + gentle + man + ly
- not all lexemes of a class will take all these particles, nor will they have the same meaning.
- how to break up (**parse**) the lexeme?
  - [[un+gentle] + man] + ly
  - [un + [gentle + man] + ly

many interpretations are possible

Derivations : Parsing

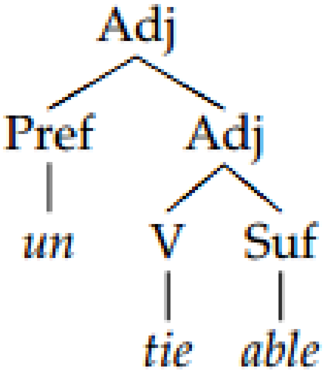


- Differing parses → different semantics :
- e.g. unlockable

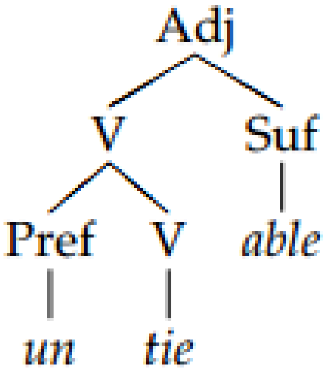
  “can’t be locked” or “can be unlocked”?

Huddleston & Pullum 05

Derivations : Ambiguity



This knot can’t be done
– it’s **untieable**



This rope is too slippery –
it’s **untieable**

- Semantics : not fully systematic –
  e.g. anomalous usage of *un-* :
  loosen same as *unloosen*

Semantics of composition

- **derivations:**

  e.g. “-er” : usually agentive – *builder, writer, teacher*

  But may be instrumental – e.g. *cooker*

  - However, meaning is constrained (not arbitrary)

- **compounds**: composed from multiple lexemes
  - *doghouse, darkroom* (endocentric, tatpuruSha) : ‘house’, ‘room’ is the head
  - *redcoat,* Hindi: *nIlakanTha* (exocentric, bahuvrihi) : refers to neither red nor coat

Models of Syntax

Structure in language

पांच फिरंगी अफसरों ___ फांसी पर ____ दिया

what can go in the blanks?

what can NOT go there?

Syntax

Sentences are built from “words”.

boys like girls
germans drink beer

sentence = noun verb noun

Syntactic Composition

- Constituency : *like girls* = verb phrase VP
  head : *like* V
  constituent: *girls* N-plural

- Grammatical Function (maps to semantics?):
  subject: boys
  predicate: like
  arguments: boys, girls

- Hierarchy and Control

One Version of Constituent Structure

- Lexicon:

  the, a, small, nice, big, very, boy, girl, sees, likes

- Grammatical sentences:
  - (the) boy (likes a girl)
  - (the small) girl (likes the big girl)
  - (a very small nice) boy (sees a very nice boy)
- Ungrammatical sentences:
  - *(the) boy (the girl)
  - *(small) boy (likes the nice girl)

N-gram language models
Word Segmentation

NLP Tasks

Word segmentation:

- Chinese: 浮法像蝴蝶.

  (“float like a butterfly)

- Hindi

  पांचफिरंगीअफसरोंकोफांसीपरलटकादिया

  - Q. Letter-or Syllable- based?
  - Which positions have low “sequence” probability?

NLP tasks : Probabilistic Models

- Other problems
  - Machine Translation:
    - P(**high** winds tonite) > P(**large** winds tonite)
  - Spell Correction
    - The office is about fifteen **minuets** from my house
      - P(about fifteen **minutes** from) > P(about fifteen **minuets** from)
  - Speech Recognition
    - P(I saw a van) >> P(eyes awe of an)
  - Verb argument structure discovery
    - Via factorization of syntactic parses to discover
    - Argument structure (syntax ?)
    - Selection preference (semantics)
  - + Summarization, question-answering, etc.,

Models of Syntax

- Linguistic Rules and Hierarchies:
  - Phonology :
  - Morphology :
  - Lexical :

} categories + rules
= syntax (e.g. CFG)

- Probabilistic models
  - Bayesian models – PCFG
  - N-grams

Probabilistic Language Modeling

- Goal: compute the probability of a sentence or sequence of words:

  $P(W) = P(w_1,w_2,w_3,w_4,w_5\ldots w_n)$

- Related task: probability of an upcoming word:

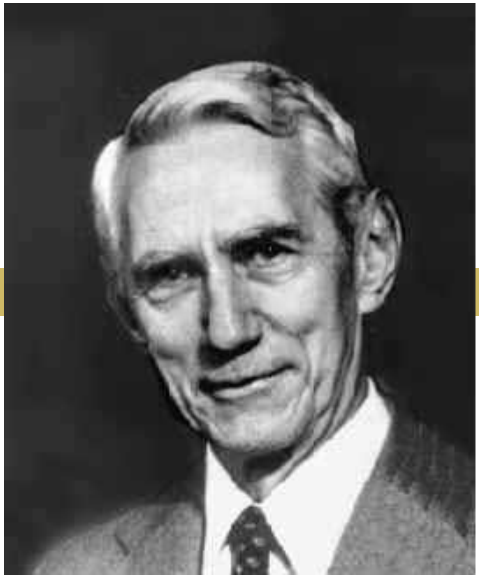
  $P(w_5|w_1,w_2,w_3,w_4)$

- A model that computes either of these:

  $P(W)$ or $P(w_n|w_1,w_2\ldots w_{n-1})$ is called a **language model**.

- Better: **the grammar** But **language model** or **LM** is standard

Shannon Entropy



- Predict the next word/letter,
  given (*n-1*) previous items
  Fn = entropy = $\mathrm{SUM}_i\ (p_i \log p_i)$
- probabilities p_i (of n-grams) from corpus:
  - F_0 (only alphabet) = $\log_2 27$ = 4.76 bits per letter
  - F_1 (1-gram frequencies p_i) = 4.03 bits
  - F_2 (bigram frequencies) = 3.32 bits
  - F_3 (trigrams) = 3.1 bits
  - F_{word} = 2.62 bits

(avg word entropy = 11.8 bits per 4.5 letter word)

Claude E. Shannon. “Prediction and Entropy of Printed English”,
Bell System Technical Journal 30:50-64. 1951.

Shannon Entropy : Human

- Ask human to guess the next letter:

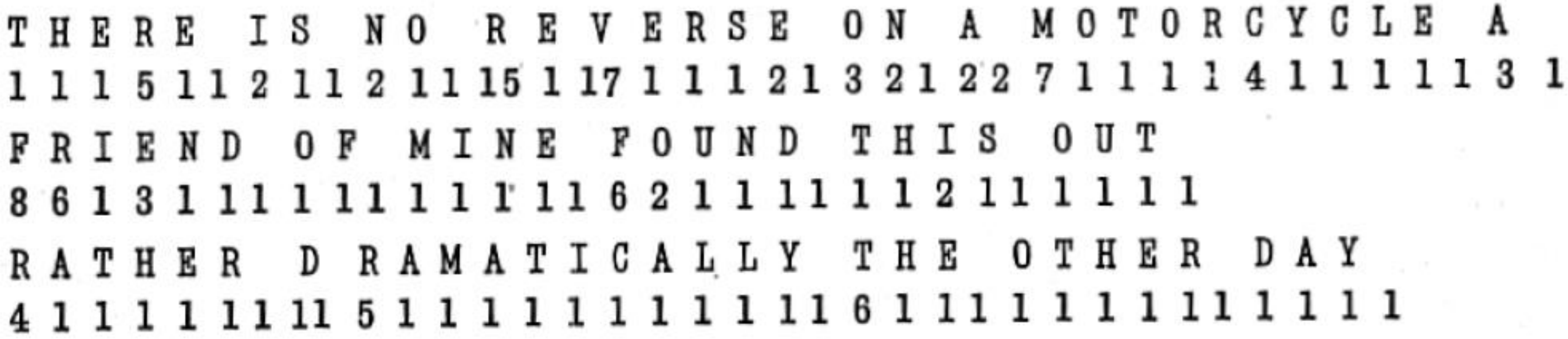
```
THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG
----ROO------NOT-V-----I------SM----OBL---

READING LAMP ON THE DESK SHED GLOW ON
REA----------O------D----SHED-OLD--O-

POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
P-L-S-----O---BU--L-S—O-------SH-----RE—C------
```

- 69% guessed on 1st attempt [“-” = 1st attempt]

Claude E. Shannon. “Prediction and Entropy of Printed English”, *Bell System Technical Journal* 30:50-64. 1951.

Shannon Entropy : Human

- Count number of attempts:



```
T H E R E  I S  N O  R E V E R S E  O N  A  M O T O R C Y C L E  A
1 1 1 5 11 2 11 2 11 15 1 17 1 1 1 21 3 21 22 7 1 1 1 1 4 1 1 1 1 1 3 1
F R I E N D  O F  M I N E  F O U N D  T H I S  O U T
8 6 1 3 1 11 1 11 1 1 1 11 6 2 1 1 11 1 1 2 11 1 1 1 1
R A T H E R  D R A M A T I C A L L Y  T H E  O T H E R  D A Y
4 1 1 1 1 11 11 5 1 1 1 1 1 1 1 1 1 11 6 1 11 1 1 1 1 11 1 1 1 1
```

- Entropy: F_1 =3.2, 4.0 F_{10} =1.0, 2.1 F_{100} = 0.6, **1.3**

Claude E. Shannon. “Prediction and Entropy of Printed English”, *Bell System Technical Journal* 30:50-64. 1951.

LANGUAGE MODELING

NL Corpora

Creating a Corpus

1961 : W. Nelson Francis and Henry Kucera of Brown Univ
500 samples of 2,000 words each from various text genres
→ American English

1970s : Lancaster-Oslo-Bergen corpus: British English
also 500 x 2000 = 1mn words – genres similar to Brown Corpus
Geoffrey Leech of Lancaster U.

1994: British National Corpus – 100mn words
Oxford U, Lancaster, Longman / Chambers dictionaries
10% : transcripts of spoken English

2000s: Google corpora: American english 155 bn words; British : 34bn

[Lindquist 2009]: Corpus linguistics and the description of English

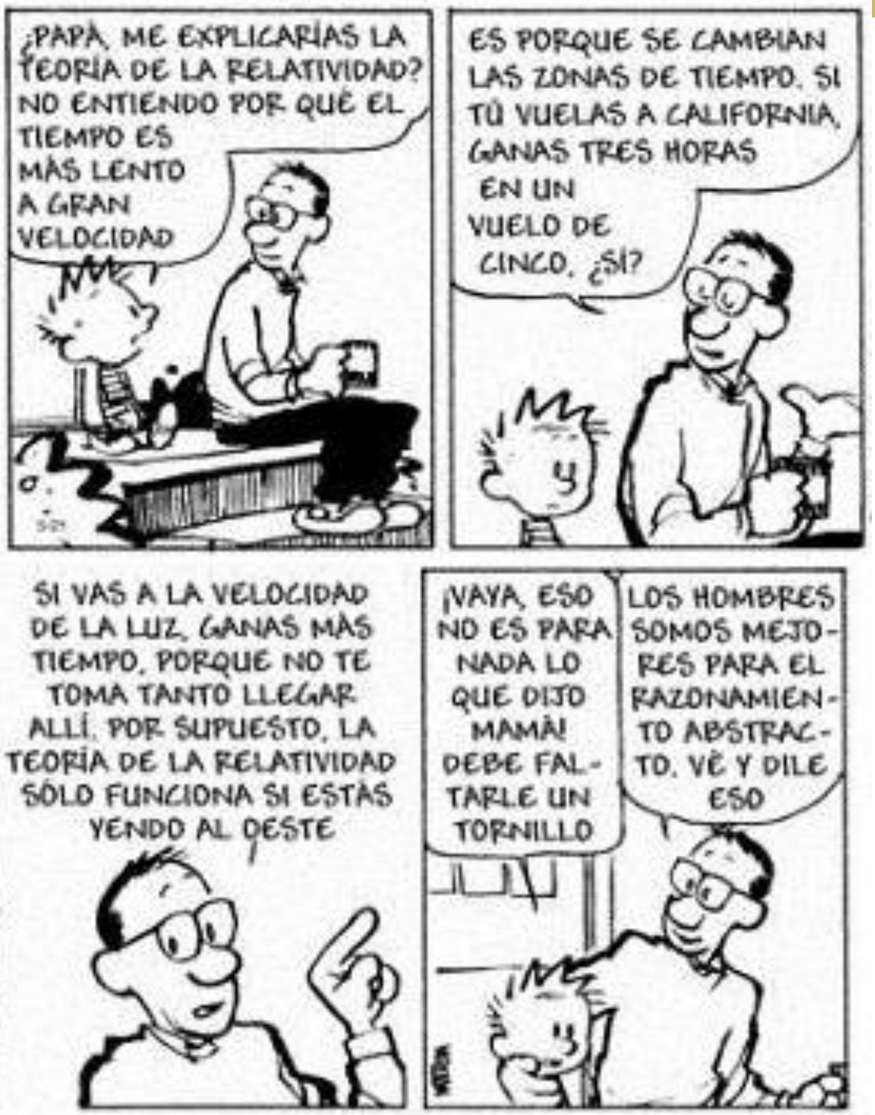
The Brown Corpus

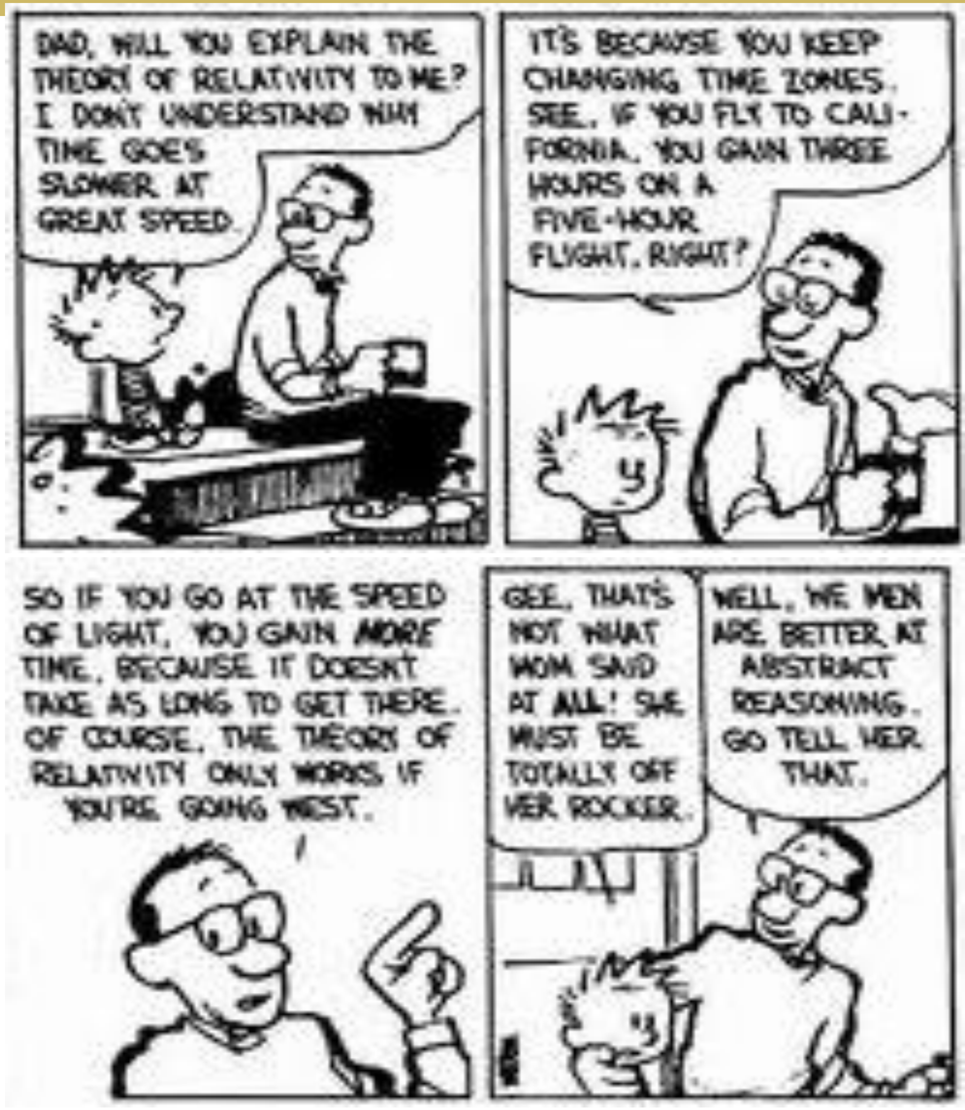
		# texts	%age
A	Press: reportage (newspapers)	44	8.8%
B	Press: editorial (including letters to the editor)	27	5.4%
C	Press: reviews (theatre, books, music, dance)	17	3.4%
D	Religion	17	3.4%
E	Skills and hobbies	36	7.2%
F	Popular lore	48	9.6%
G	Belles letters, biography, memoirs etc.	75	15.0%
H	Miscellaneous (mainly government documents)	30	6.0%
J	Learned (academic texts)	80	16.0%
K	General fiction (novels and short stories)	29	5.8%
L	Mystery and detective fiction	24	4.8%
M	Science fiction	6	1.2%
N	Adventure and Western fiction	29	5.8%
P	Romance and love story	29	5.8%
R	Humour	9	1.8%
	Non-fiction subtotal	374	75%
	Fiction subtotal	126	25%
	Total	500	100%

News: political, sports, society "spot news", financial, cultural)

[Lindquist 2009]: Corpus linguistics and the description of English

Parallel Corpora


¿PAPÁ, ME EXPLICARÍAS LA TEORÍA DE LA RELATIVIDAD? NO ENTIENDO POR QUÉ EL TIEMPO ES MÁS LENTO A GRAN VELOCIDAD
ES PORQUE SE CAMBIAN LAS ZONAS DE TIEMPO. SI TÚ VUELAS A CALIFORNIA, GANAS TRES HORAS EN UN VUELO DE CINCO. ¿SÍ?
SI VAS A LA VELOCIDAD DE LA LUZ, GANAS MÁS TIEMPO, PORQUE NO TE TOMA TANTO LLEGAR ALLÍ. POR SUPUESTO, LA TEORÍA DE LA RELATIVIDAD SÓLO FUNCIONA SI ESTÁS YENDO AL OESTE
¡VAYA, ESO NO ES PARA NADA LO QUE DIJO MAMÁ! DEBE FALTARLE UN TORNILLO
LOS HOMBRES SOMOS MEJORES PARA EL RAZONAMIENTO ABSTRACTO. VÈ Y DILE ESO


DAD, WILL YOU EXPLAIN THE THEORY OF RELATIVITY TO ME? I DON'T UNDERSTAND WHY TIME GOES SLOWER AT GREAT SPEED.
IT'S BECAUSE YOU KEEP CHANGING TIME ZONES. SEE, IF YOU FLY TO CALIFORNIA, YOU GAIN THREE HOURS ON A FIVE-HOUR FLIGHT, RIGHT?
SO IF YOU GO AT THE SPEED OF LIGHT, YOU GAIN MORE TIME, BECAUSE IT DOESN'T TAKE AS LONG TO GET THERE. OF COURSE, THE THEORY OF RELATIVITY ONLY WORKS IF YOU'RE GOING WEST.
GEE, THAT'S NOT WHAT MOM SAID AT ALL! SHE MUST BE TOTALLY OFF HER ROCKER.
WELL, WE MEN ARE BETTER AT ABSTRACT REASONING. GO TELL HER THAT.

Parallel Corpus

Congress MP from Haryana Birender Singh said at a programme that "once someone had told me that Rs 100 crore was required to get a Rajya Sabha berth.

But he said he got it for Rs 80 crore and saved Rs 20 crore. Now will people who are willing to invest Rs 100 crore, ever think of the poor country."

राज्य सभा सांसद बीरेंद्र सिंह ने एक कार्यक्रम में कहा था, "एक बार की बात है कि मुझे एक व्यक्ति ने बताया कि राज्य सभा की सीट 100 करोड़ रुपए में मिलती है. उसने बताया कि उसे खुद यह सीट 80 करोड़ रुपए में मिल गई, 20 करोड़ बच गए. मगर क्या वे लोग, जो 100 करोड़ खर्च करके यह सीट खरीदने के इच्छुक हैं, कभी इस गरीब देश के बारे में भी सोचेंगे?"

একটি অনুষ্ঠানে তিনি বলেন, 'আমাকে একজন বলেছিলেন, ১০০ কোটি রুপি হলেই রাজ্য সভার একটি আসন পাওয়া যায়।

তবে ৮০ কোটি রুপি দিয়ে তিনি একটি আসন সংগ্রহ করে ২০ কোটি রুপি বাঁচিয়েছেন।'

Matching on parallel Corpus

电脑坏了。

The computer is broken.

电脑死机了。

My computer has frozen.

我想玩电脑。

I want to play on the computer.

我家没有电脑。

I don't have a computer at home.

我有一台电脑。

I have a computer.

你有两台电脑吗？

Do you have two computers?

Parallel Corpus

电脑坏了。

The computer is broken.

电脑死机了。

My computer has frozen.

我想玩**电脑**。

I want to play on the computer.

我家没有**电脑**。

I don't have a computer at home.

我有一台**电脑**。

I have a computer.

你有两台**电脑**吗？

Do you have two computers?

电脑 : *diànnǎo*, computer

[电 : *diàn* lightning, electricity 脑 : *nǎo* brain]

Parallel Corpus

电脑坏了。

The computer is broken.

电脑死机了。

My computer has frozen.

我想玩电脑。

I want to play on the computer.

我家没有电脑。

I don't have a computer at home.

我有一台电脑。

I have a computer.

你有两台电脑吗？

Do you have two computers?

有 : “in possession of”

[**又** (“hand”) + 月 (肉) (“meat”) = a hand holding meat]

LANGUAGE MODELING

Generalization and zeros

The perils of overfitting

- N-grams only work well for word prediction if the test corpus looks like the training corpus
  - In real life, it often doesn’t
  - We need to train robust models that generalize!
  - One kind of generalization: Zeros!
    - Things that don’t ever occur in the training set
      - But occur in the test set

Zeros

- □ Training set:

  … denied the allegations

  … denied the reports

  … denied the claims

  … denied the request

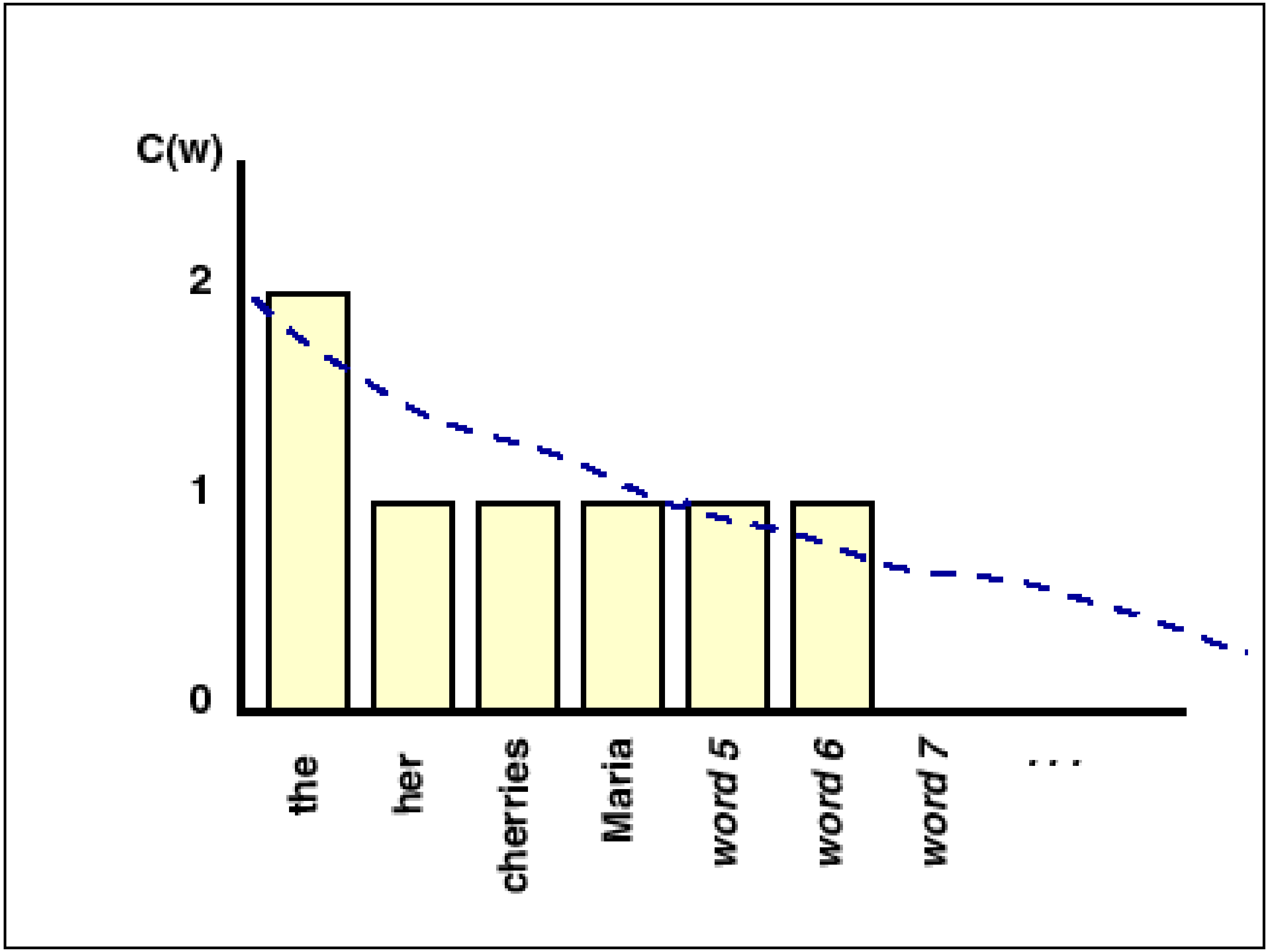
  P(“offer” | denied the) = 0

- Test set

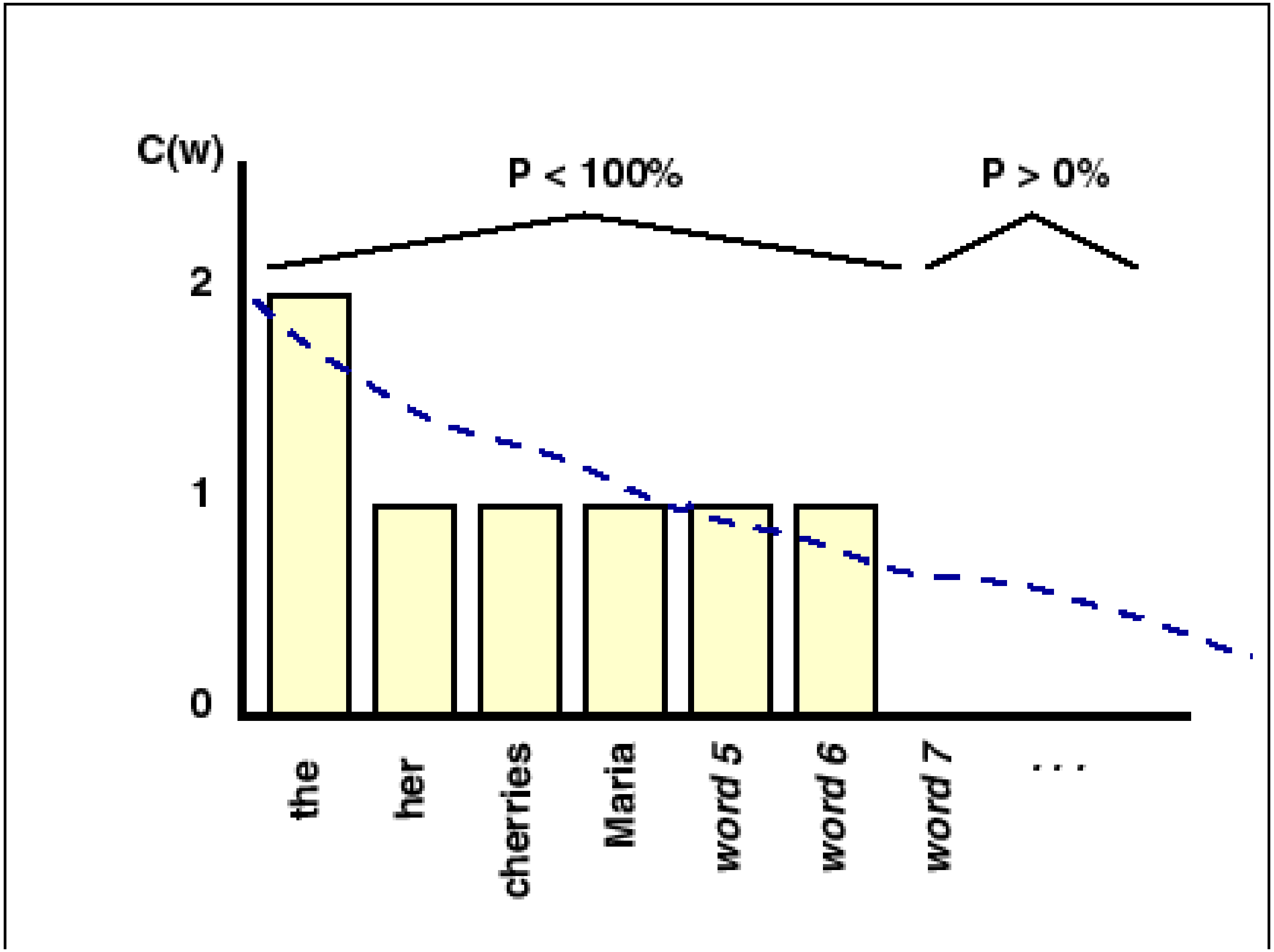
  … denied the offer

  … denied the loan

Actual Probability Distribution:



Actual Probability Distribution:

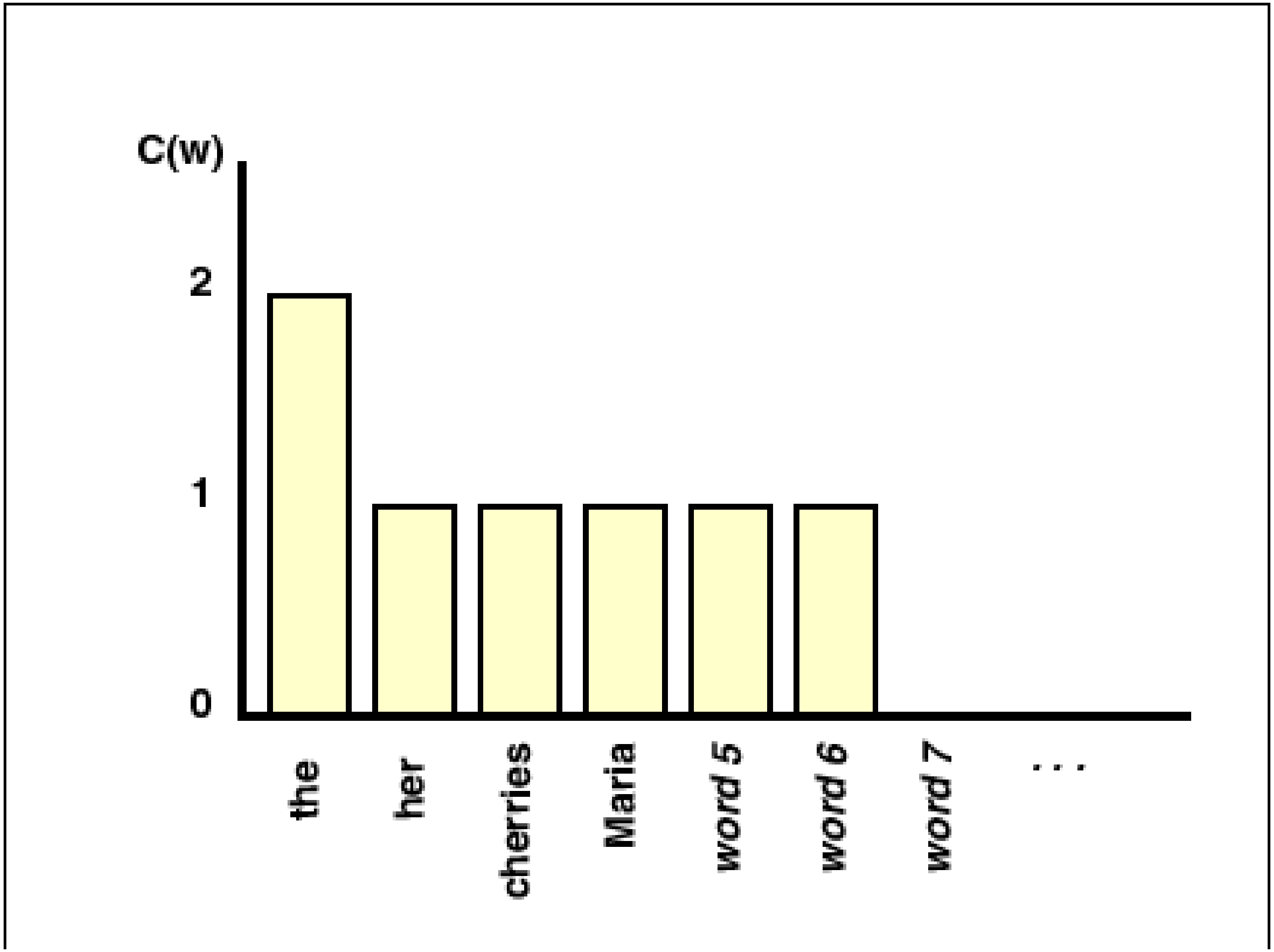


“Smoothing”

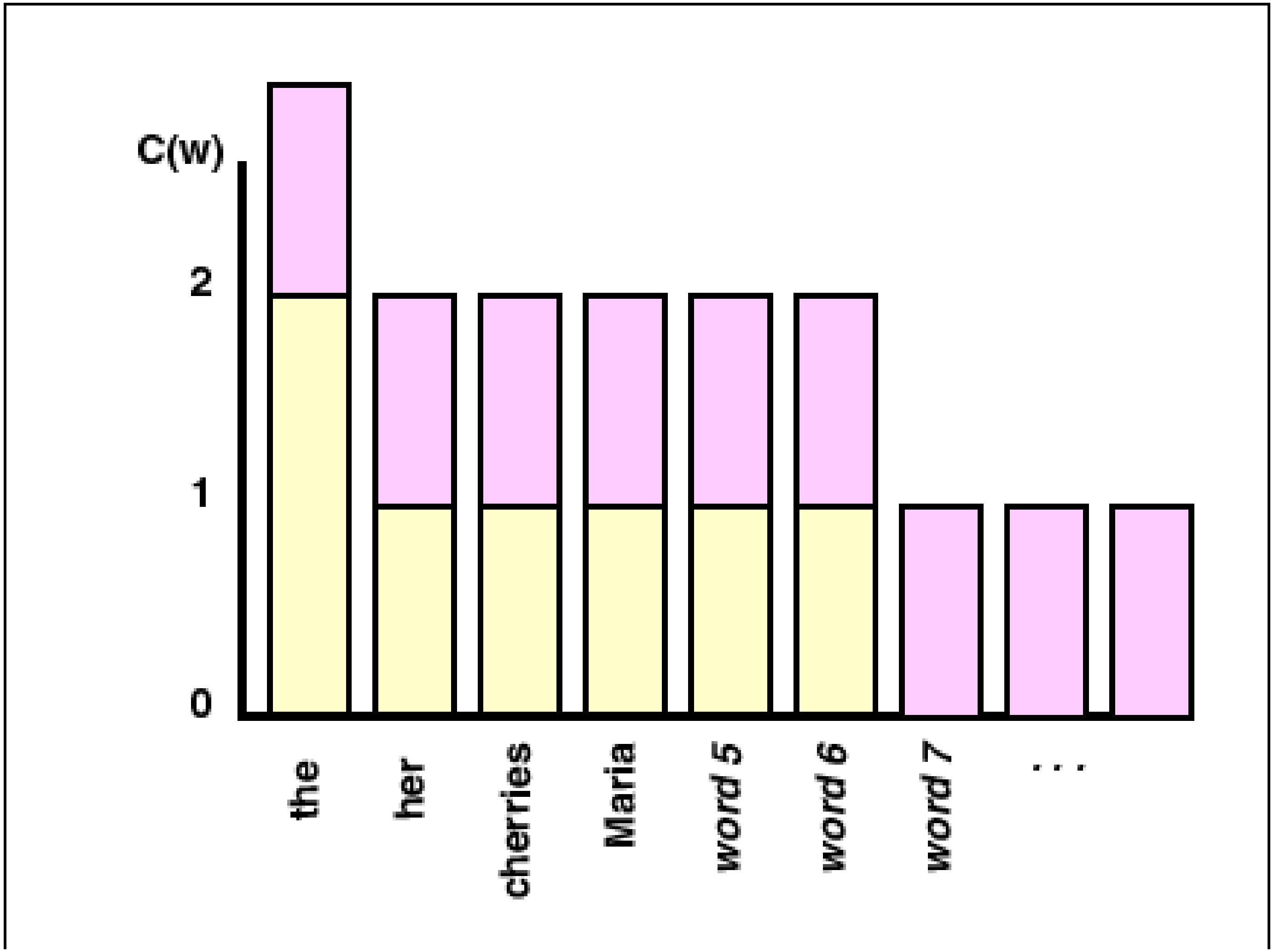
- Develop a model which decreases probability of seen events and allows the occurrence of previously unseen n-grams
- a.k.a. “Discounting methods”
- “Validation” – Smoothing methods which utilize a second batch of test data.

based on Manning and Schütze

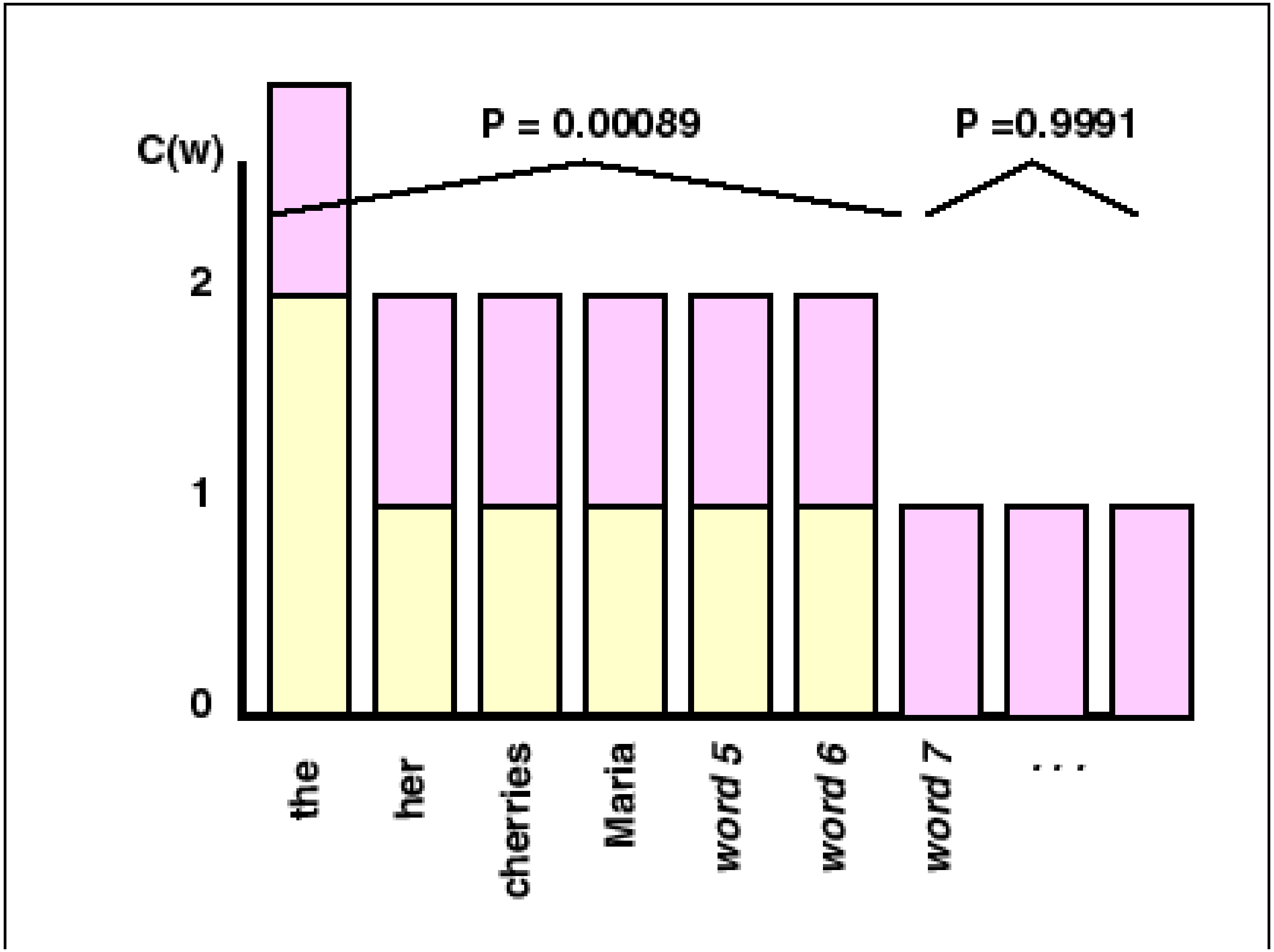
Smoothing



Smoothing: +1



Smoothing: +1


C(w)
P = 0.00089
P =0.9991
2
1
0
the
her
cherries
Maria
word 5
word 6
word 7
. . .

Spelling correction w bigram language model

- "a stellar and versatile **acress** whose combination of sass and glamour…"
- Counts from the Corpus of Contemporary American English with add-1 smoothing
- P(actress|versatile)=.000021 P(whose|actress) = .0010
- P(across|versatile) =.000021 P(whose|across) = .000006

- **P("versatile actress whose") = .000021*.0010 = 210 $\mathbf{x10^{-10}}$**
- P("versatile across whose") = .000021*.000006 = 1 $x10^{-10}$

LANGUAGE MODELING

Estimating N-gram Probabilities

Probabilistic Language Modeling

- Goal: determine if a sentence or phrase has a high acceptability in the language

  → compute the probability of the sequence of words

  E.g. “its water is so transparent that”

  - P(its, water, is, so, transparent, that)

Probabilistic Language Modeling

$$P(W) = P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$$

- Related task: probability of an upcoming word:

$$P(w_5|w_1,w_2,w_3,w_4)$$

Reliability vs. Discrimination

- larger n: more information about the context of the specific instance (greater discrimination)

- smaller n: more instances in training data, better statistical estimates (more reliability)

The Chain Rule

- Chain Rule in General

  $P(x_1,x_2,x_3,\ldots,x_n) =$

  $P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$

- Proof:
  - Holds for n=2 (Product rule)
  - Assume is true for $X = x_1 \ldots x_{n-1.}$

    $P(X\ ,\ x_n) = P(X)\ P\ (x_n|X)$ → General chain rule

Markov Assumption

Andrei Markov
1856-1922, Russia

- Simplifying assumption:

  Depends only on *k*-nearby text

- *First-order* Markov Process (k= 1):

$$P(\text{the} \mid \text{its water is so transparent that}) » P(\text{the} \mid \text{that})$$

- or *Second-order* (k=2):

$$P(\text{the} \mid \text{its water is so transparent that}) » P(\text{the} \mid \text{transparent that})$$

Estimating bigram probabilities

- The Maximum Likelihood Estimate

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

N-gram Text Generation

Sentence Generation

Unigram Model: No dependencies on previous words

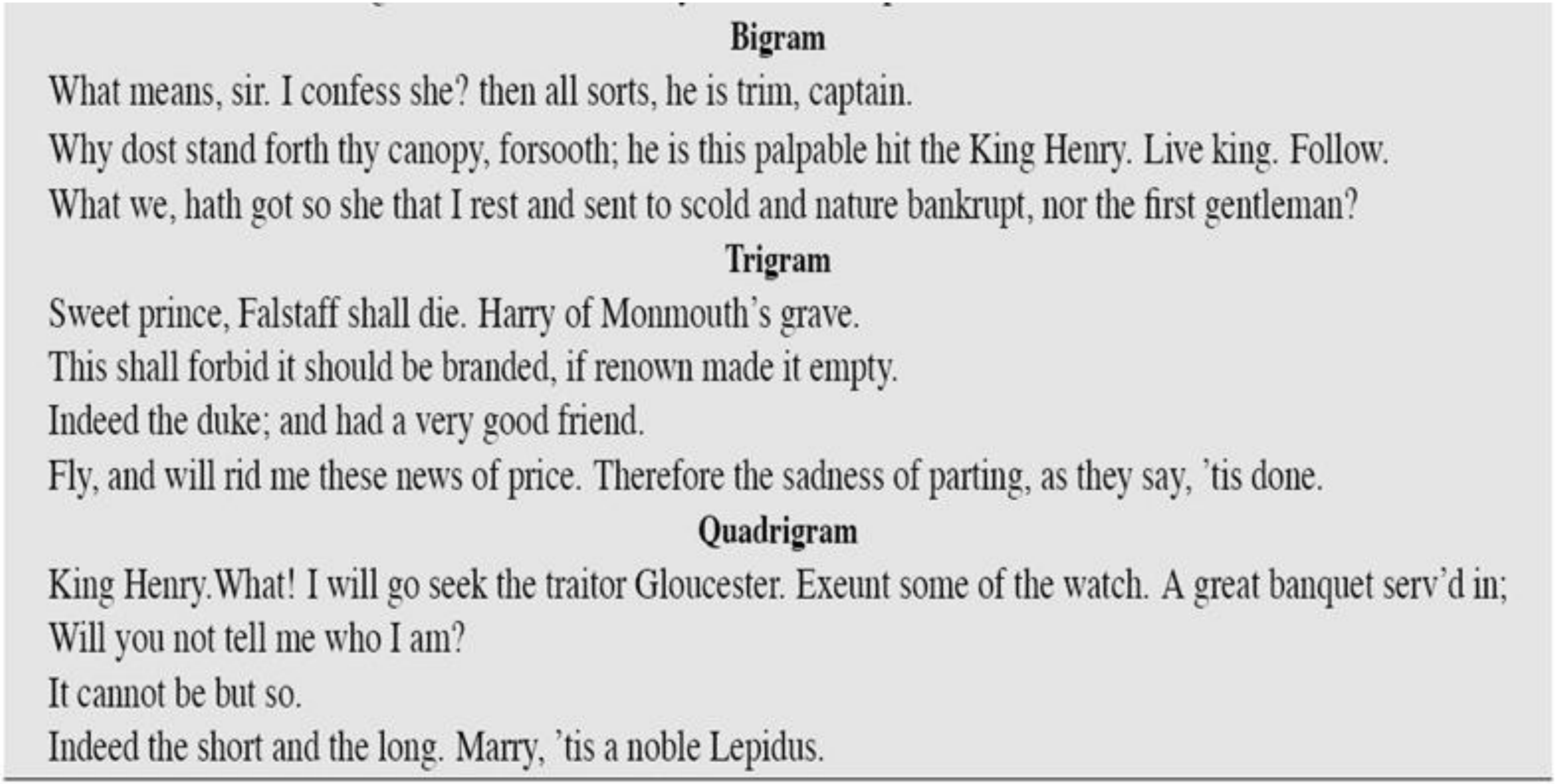
$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i)$$

Bigram Model : Depends on 1 previous word

$$P(w_i \mid w_1 w_2 \dots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

The Corpus matters

- What corpus was used to generate these:



Bigram

What means, sir. I confess she? then all sorts, he is trim, captain.

Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.

What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?

Trigram

Sweet prince, Falstaff shall die. Harry of Monmouth's grave.

This shall forbid it should be branded, if renown made it empty.

Indeed the duke; and had a very good friend.

Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.

Quadrigram

King Henry.What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in;

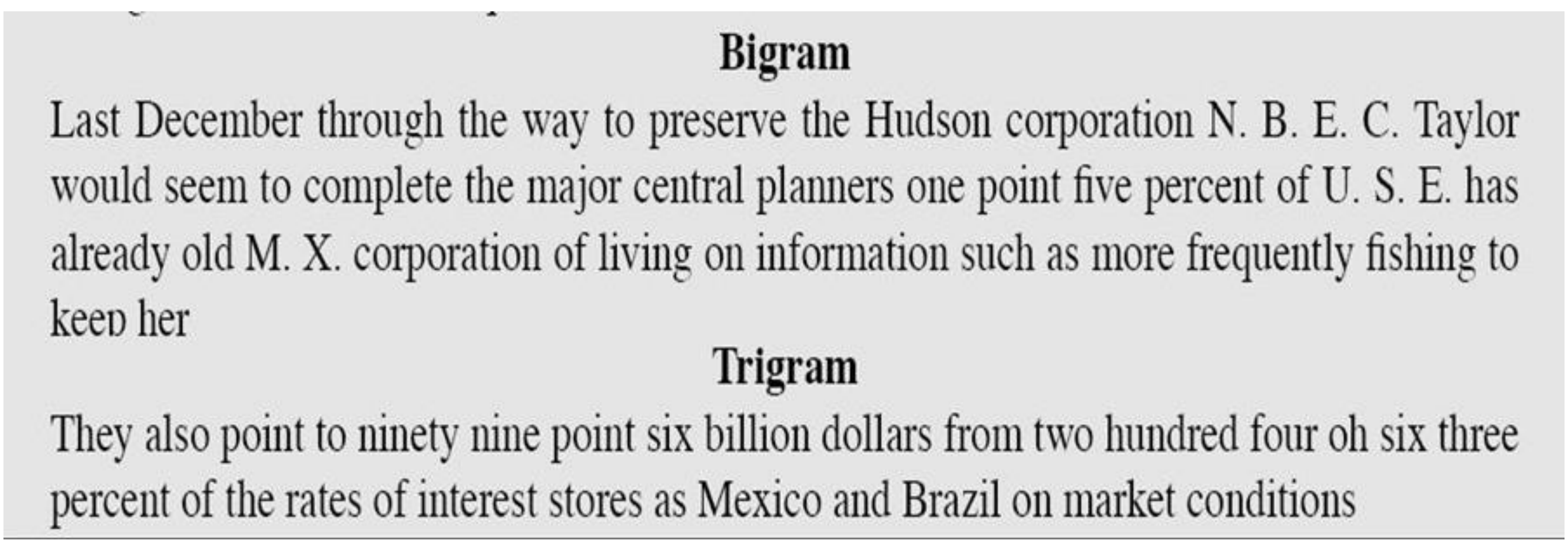
Will you not tell me who I am?

It cannot be but so.

Indeed the short and the long. Marry, 'tis a noble Lepidus.

The Corpus matters

- What corpus was used to generate these:



Bigram

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

Trigram

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

N-gram frequency falls rapidly w N

- Shakespeare Corpus: N=884,647 tokens, V=29,066
- Shakespeare produced 300,000 bigram types out of V^2= 844 million possible bigrams.
  - So 99.96% of the possible bigrams were never seen (have zero entries in the table)
- Quadrigrams worse: Shakespeare had very specific patterns of usage

Limitations of N-gram models

- Advantages:
  - Does not require expensive annotated corpora
  - Annotations are often disputed
  - Efficacy of intermediate representations are doubtful
- We can extend to trigrams, 4-grams, 5-grams
  - Corpus size must grow exponentially larger
- Main Disadvatage: **Long-distance dependencies**:

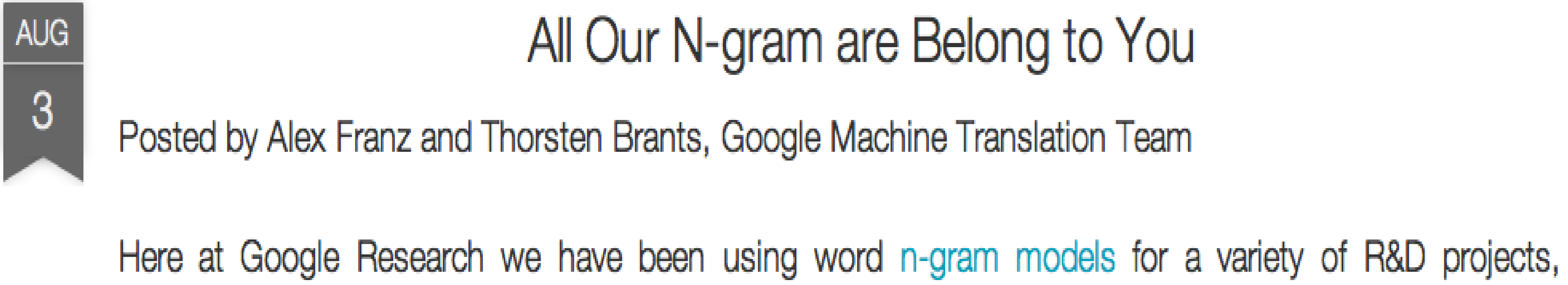
“The computer which I had just put into the machine room on the fifth floor crashed.”

Practical Issues

- We do everything in log space
  - Avoid underflow
  - (also adding is faster than multiplying)

$$\log(p_1 \times p_2 \times p_3 \times p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

Google N-Gram Release, August 2006



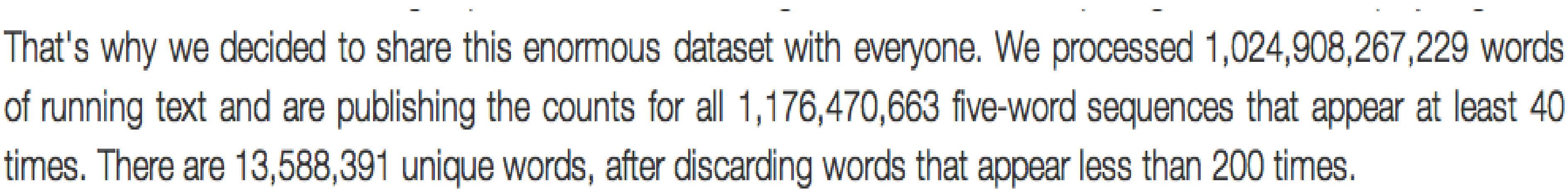
AUG 3

All Our N-gram are Belong to You

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects,

…



That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

Google N-Gram Release

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensible 40
- serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Google N-Gram Release

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensible 40
- serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Computational Morphology

Computational Analysis

- [Goldsmith 01]

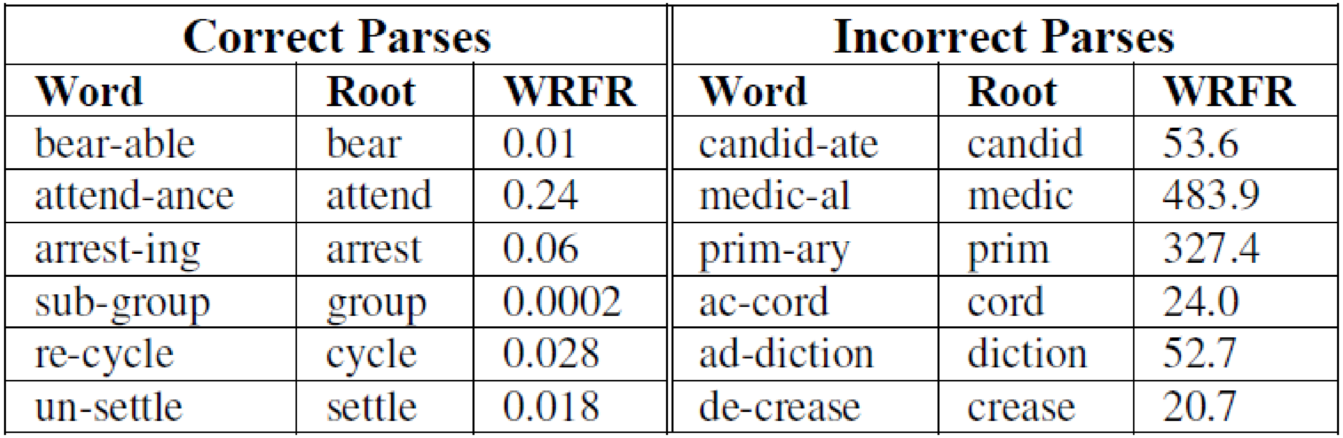
  Information-Theoretic ideas - Minimum Description Length

  Which “signature” (pattern) will results in the most compact description of the corpus?

```
----------------------------------------  Counts ----------
Signature  Example                   Stem # (type)  Token
-----------------------------------------------------------
NULL.ed.ing    betray betrayed betraying 69      864
NULL.ed.ing.s  remain remained           14      516
                   remaining remains
NULL.s.        cow cows                  253     3414
e.ed.es.ing    notice noticed notices    4 62
                      noticing
-----------------------------------------------------------
```

Computational Analysis

- [Dasgupta & V.Ng 07]
  - Simple concatenation not enough for more agglutinated languages.
  - Attempt to discover root word form. (*denial* →*deny*)
  - Assumption: if compound word is common,then root word will also : Word-Root Frequency Ratios (WRFR)



Correct Parses			Incorrect Parses		
Word	**Root**	**WRFR**	**Word**	**Root**	**WRFR**
bear-able	bear	0.01	candid-ate	candid	53.6
attend-ance	attend	0.24	medic-al	medic	483.9
arrest-ing	arrest	0.06	prim-ary	prim	327.4
sub-group	group	0.0002	ac-cord	cord	24.0
re-cycle	cycle	0.028	ad-diction	diction	52.7
un-settle	settle	0.018	de-crease	crease	20.7

STATISTICAL NATURAL LANGUAGE PARSING

POS-Tagging

POS Tagging Approaches

- **Rule-Based**: Human crafted rules based on lexical and other linguistic knowledge (e.g. ENGTWOL 95)
- **Stochastic**: Trained on human annotated corpora like the Penn Treebank
  - **Statistical models**: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF), log-linear models, support vector machines
  - **Rule learning**: Transformation Based Learning (TBL)

- Many English POS-taggers are publicly available
- Hindi / Bangla POS tagger:
  - http://nltr.org/snltr-software/

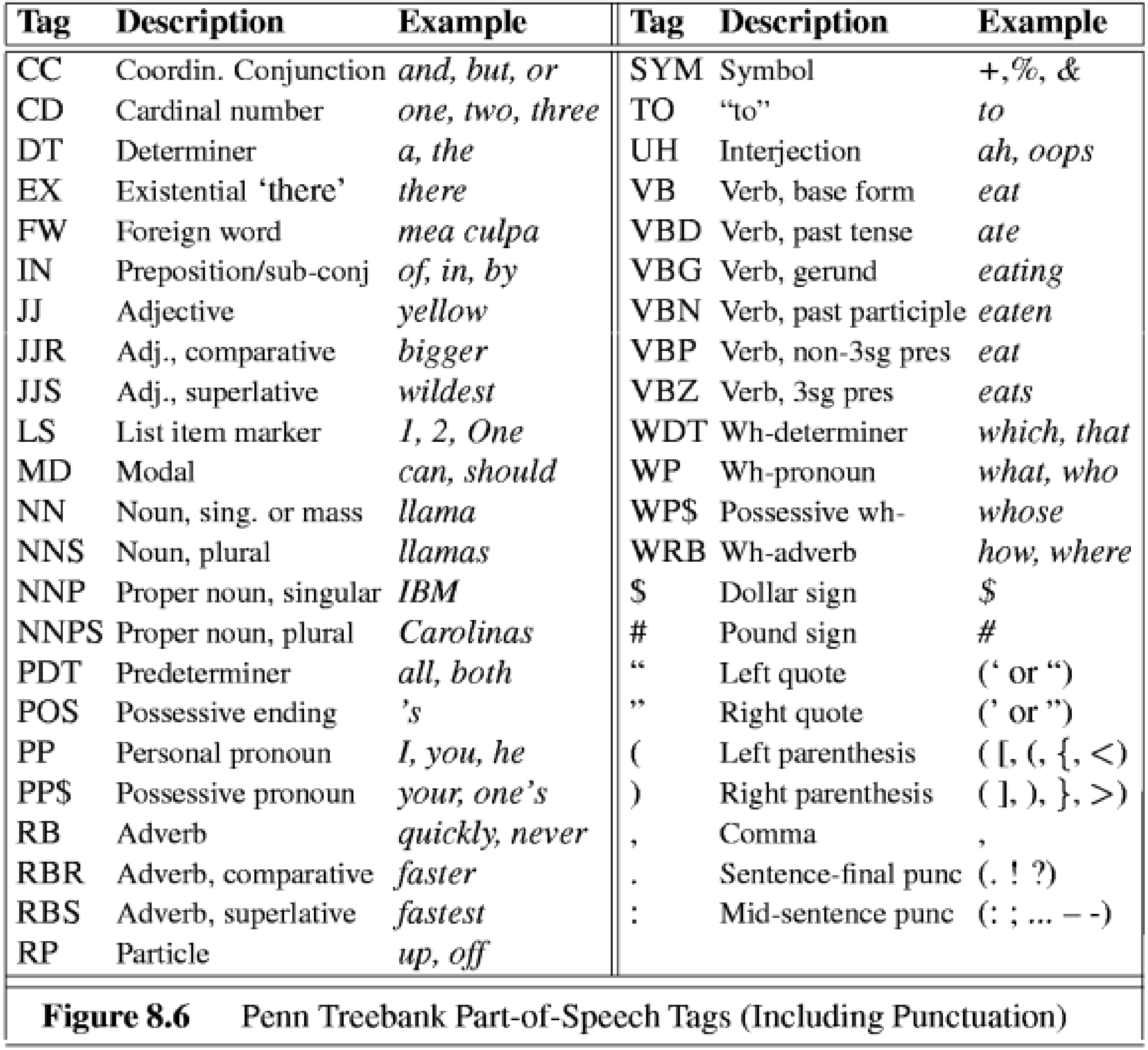
Deciding on a POS tagset

NOUN	The DOG barked.	WE saw YOU.
VERB	The dog BARKED.	It IS impossible.
ADJECTIVE	He's very OLD.	I've got a NEW car.
DETERMINATIVE	THE dog barked.	I need SOME nails.
ADVERB	She spoke CLEARLY.	He's VERY old.
PREPOSITION	It's IN the car.	I gave it TO Sam.
COORDINATOR	I got up AND left.	It's cheap BUT strong.
SUBORDINATOR	It's odd THAT they were late.	I wonder WHETHER it's still there.
INTERJECTOR	OH, HELLO, WOW, OUCH	

from [huddleston-pullum 05] *Student's intro to English Grammar*

Coordinator / subordinator: markers for coordinate / subordinate clauses
POS distinctions based on analysis of syntax and semantics

Penn Tagset



Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	“to”	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential ‘there’	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	“	Left quote	(‘ or “)
POS	Possessive ending	*’s*	”	Right quote	(’ or ”)
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one’s*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Figure 8.6 Penn Treebank Part-of-Speech Tags (Including Punctuation)

Figure: jurafsky-martin ch.8 (2000)

Penn Treebank

[Marcus etal

"*I miss the good old days when all we had to worry about was nouns and verbs.*"

[palmer: grammar (1984)]

Stochastic POS-tagging

- Markovian assumption : tag depends on limited set of previous tags
- HMM:

  maximize P(word|tag) * P(tag| previous n tags)
- Maximize the probability for whole sentence, not single word

$$S = \arg\max_{t1\ldots tn} \prod_{i=1,n} P(w_i \mid t_i) P(t_i \mid t_{i-1})$$

Stochastic POS-tagging

- Secretariat/NNP is/VBZ expected/VBN to/TO race/VB tomorrow/NN
- People/NNS continue/VBP to/TO inquire/VB the/DT reason/NN for/IN the/DT race/NN for/IN outer/JJ space/NN
- *to race* vs. *the race*

Stochastic POS-tagging

- □ *to/TO race* *the/DT race*

- □ P(VB|TO) P(*race*|VB)
- □ P(NN|TO) P(*race*|NN)

- □ P(NN|TO) = .021 P(race|NN) = .00041
- □ P(VB|TO) = .34 P(race|VB) = .00003

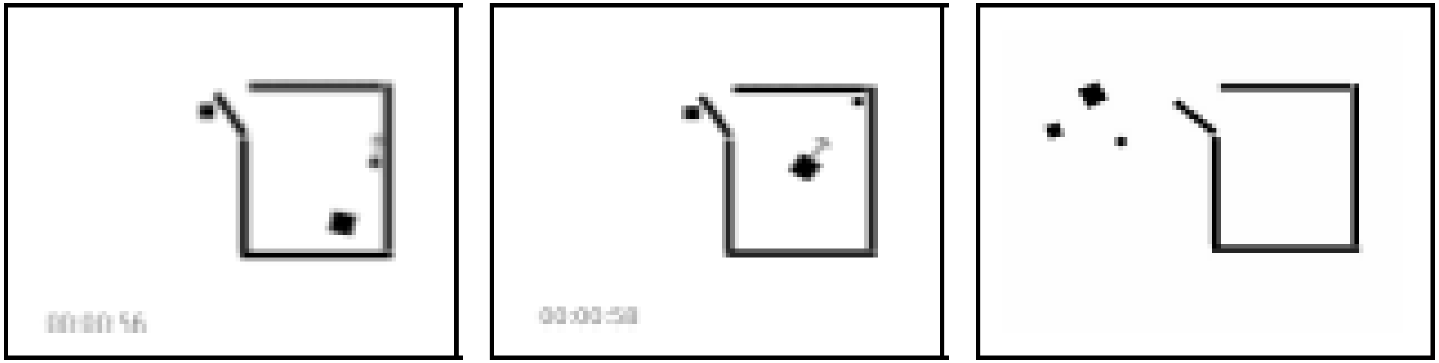
- □ P(VB|TO)P(race|VB) = .00001
- □ P(NN|TO)P(race|NN) = .000007

GROUNDED LANGUAGE MODELS

Unsupervised POS and Syntax: Grounded Models

Language Acquisition : Domains

- Perceptual input



[heider/simmel 1944] [hard/tversky 2003]

- Discovery Targets:
  - semantics: objects, 2-agent actions, relations
  - lexicon : nominal, transitive verbs, preposition
  - lexical categories: N VT P Adj
  - constructions: PP VP S
  - sense extension (metaphor)   [nayak/mukerjee (AAAI-12)]

Linguistic input

- input = description commentaries transcribed into text
  - 48 descriptions in English / 10 : Hindi

- Unconstrained description by different subjects:
  - the little square hit the big square
  - they're hitting each other
  - the big square hit the little square
  - circle and square in [unitelligible stammer]
  - the two squares stopped fighting

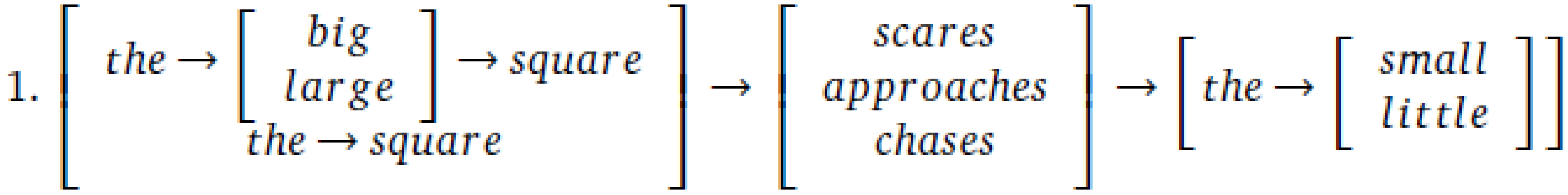
•छोटा बक्सा	बडा बक्सा	मे	कुछ	बातचीत	होती है
little box	big box	between	some	talk	happens

POS categories - Unsupervised

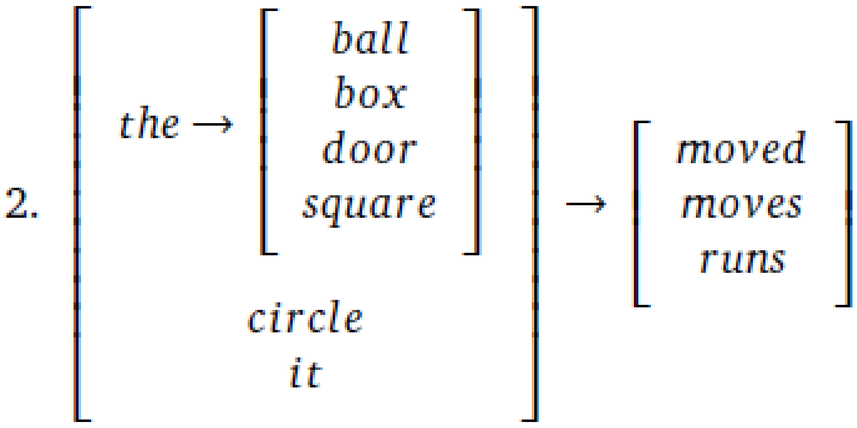
$$\begin{bmatrix} ball \\ block \\ box \\ circle \\ square \end{bmatrix} \begin{bmatrix} in \\ inside \\ into \end{bmatrix} \begin{bmatrix} chases \\ pushes \\ corners \\ the \end{bmatrix} \quad \begin{bmatrix} big \\ large \\ little \\ the \end{bmatrix}$$

[mukerjee nayak 12] based on ADIOS
[solan rupin edelman 05]

Language Structures : Verbs



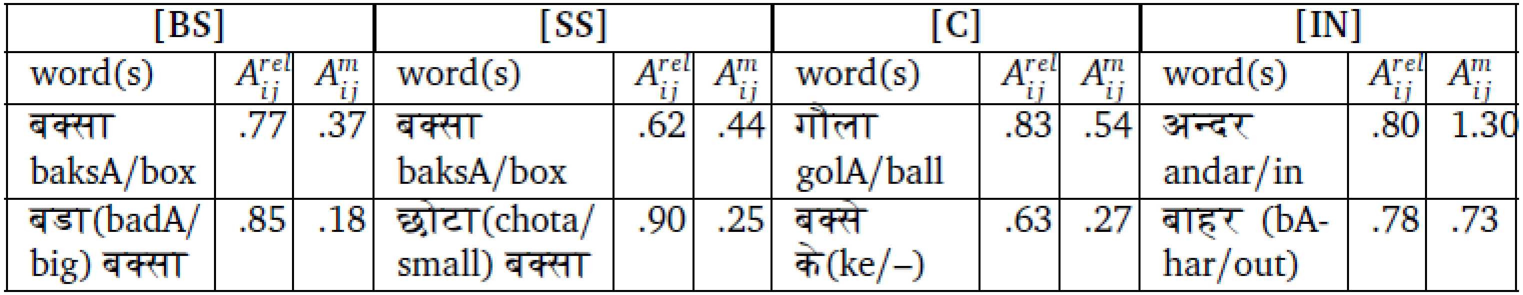
$$1.\ \left[\begin{array}{c} \mathit{the} \to \begin{bmatrix} \mathit{big} \\ \mathit{large} \end{bmatrix} \to \mathit{square} \\ \mathit{the} \to \mathit{square} \end{array}\right] \to \begin{bmatrix} \mathit{scares} \\ \mathit{approaches} \\ \mathit{chases} \end{bmatrix} \to \left[\mathit{the} \to \begin{bmatrix} \mathit{small} \\ \mathit{little} \end{bmatrix} \right]$$



$$2.\ \left[\begin{array}{c} \mathit{the} \to \begin{bmatrix} \mathit{ball} \\ \mathit{box} \\ \mathit{door} \\ \mathit{square} \end{bmatrix} \\ \mathit{circle} \\ \mathit{it} \end{array}\right] \to \begin{bmatrix} \mathit{moved} \\ \mathit{moves} \\ \mathit{runs} \end{bmatrix}$$

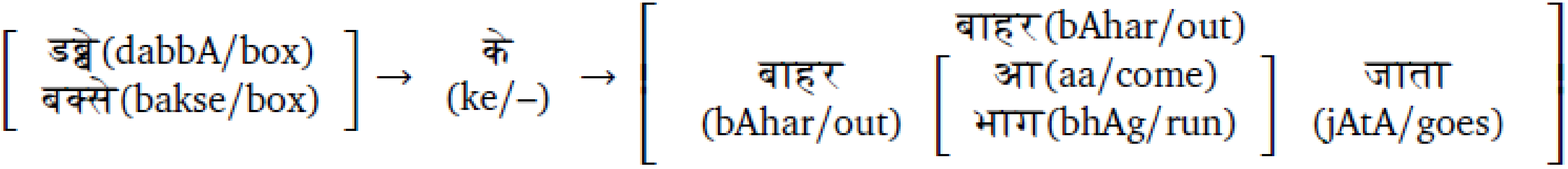
[mukerjee nayak 12]

Hindi Acquisition: Word learning



[BS]			[SS]			[C]			[IN]		
word(s)	A^{rel}_{ij}	A^{m}_{ij}	word(s)	A^{rel}_{ij}	A^{m}_{ij}	word(s)	A^{rel}_{ij}	A^{m}_{ij}	word(s)	A^{rel}_{ij}	A^{m}_{ij}
बक्सा baksA/box	.77	.37	बक्सा baksA/box	.62	.44	गौला golA/ball	.83	.54	अन्दर andar/in	.80	1.30
बडा(badA/ big) बक्सा	.85	.18	छोटा(chota/ small) बक्सा	.90	.25	बक्से के(ke/–)	.63	.27	बाहर (bAhar/out)	.78	.73

Incipient Syntax



[डब्बे(dabbA/box)
बक्से(bakse/box)] → के
(ke/–) → [बाहर
(bAhar/out) [बाहर(bAhar/out)
आ(aa/come)
भाग(bhAg/run)] जाता
(jAtA/goes)]