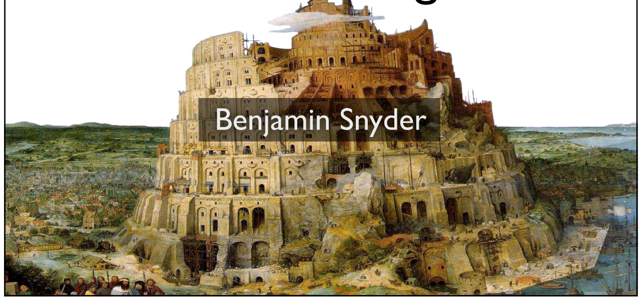
CS 545: Natural Language Processing



Benjamin Snyder

Administrivia

- Prereqs:
  - Comfort doing simple math (probability, stats, a tiny bit of calculus and linear algebra -- we will review as necessary)
  - Programming experience
  - Basic algorithms knowledge (e.g. dynamic programs)
  - Interest in language / linguistics

Administrivia

- Grading:
  - 6-8 Homework assignments (50% of grade)
  - Midterm quiz (15% of grade)
  - Project / Final (20% of grade)
  - Attendance / participation (15% of grade)

Administrivia

- Room change!
- Starting on thursday, we will meet in Room 2255 EH.
- Will send an email reminder.

Administrivia

- Communication via classlist and Piazza
- Benjamin Snyder: bsnyder@cs.wisc.edu, Room 6395 in CS building, Office hours TBD
- TA: Nisha Kiran, nkiran@cs.wisc.edu, Office hours time and location TBD

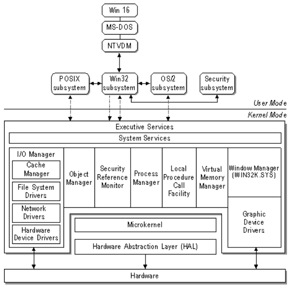
Tentative Syllabus

	1/24/12	Intro
	1/26/12	Math Review
	1/31/12	Perl + Python
HW1 Due	2/2/12	Words + Lexicons
	2/7/12	Language Models
	2/9/12	Smoothing
	2/14/12	Speech Recognition
HW2 Due	2/17/12	-
	2/21/12	Spelling
	2/23/12	Text Classification
HW3 Due	2/28/12	Part-of-speech tagging
	3/1/12	Hidden Markov Models
	3/6/12	Bayesian Probability
HW4 Due	3/8/12	Formal language and Natural Language
	3/13/12	Syntactic Parsing I
	3/15/12	Syntactic Parsing II
HW5 Due	3/20/12	Machine Translation I
	3/22/12	Machine Translation II
	3/27/12	Midterm review
	3/29/12	Midterm
	4/3/12	-
	4/5/12	-
	4/10/12	Projects
	4/12/12	Semantics I
	4/17/12	Semantics II
HW 6 Due	4/19/12	Natural Language Generation
	4/24/12	Information Retrieval + Web Search
	4/26/12	Text encodings
	5/1/12	Deciphering lost languages
	5/3/12	Computational Typology
	5/8/12	?
Project Due	5/10/12	?
	5/15/12	FINAL(?)

Questions?

Survey

- How many people here?
- languages
- Terms: gaussian distribution, Maximum likelihood estimator, entropy, eigenvector, lagrange multiplier, morpheme, dynamic program
- know python, perl



Win 16

MS-DOS

NTVDM

POSIX subsystem

Win32 subsystem

OS/2 subsystem

Security subsystem

User Mode

Kernel Mode

Executive Services

System Services

I/O Manager

Cache Manager

File System Drivers

Network Drivers

Hardware Device Drivers

Object Manager

Security Reference Monitor

Process Manager

Local Procedure Call Facility

Virtual Memory Manager

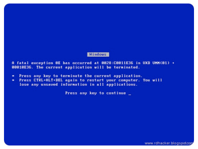
Window Manager (WIN32K.SYS)

Graphic Device Drivers

Microkernel

Hardware Abstraction Layer (HAL)

Hardware

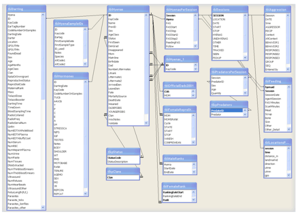


```
                    Windows

A fatal exception 0E has occurred at 0028:C0011E36 in VXD VMM(01) +
00010E36. The current application will be terminated.

*  Press any key to terminate the current application.
*  Press CTRL+ALT+DEL again to restart your computer. You will
   lose any unsaved information in all applications.

              Press any key to continue _
```



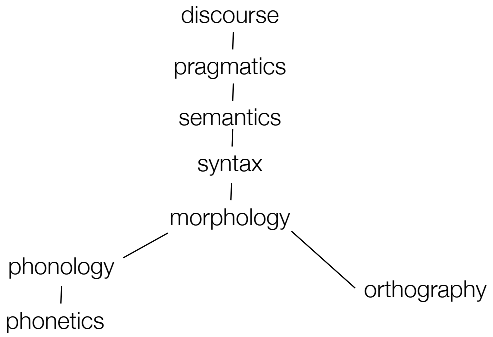
All that stuff is important, but...

What can computers do with human language?

A Dream

- Make computers more useful by getting them to ...
  - Answer questions using the Web
  - Translate documents from one language to another
  - Do library research (what papers to read? summarize!)
  - Manage email intelligently (what's urgent? what's spam?)
  - Help us make informed decisions (which phone to buy)
  - Follow directions given by any user (your Grandma)
  - Fix your spelling or grammar
  - Write poems or novels
  - Give advice, psychotherapy
  - Predict world events (elections, financial markets)
- Major obstacle to this fantasy: **language**!

Different Levels of Linguistic Knowledge



discourse

pragmatics

semantics

syntax

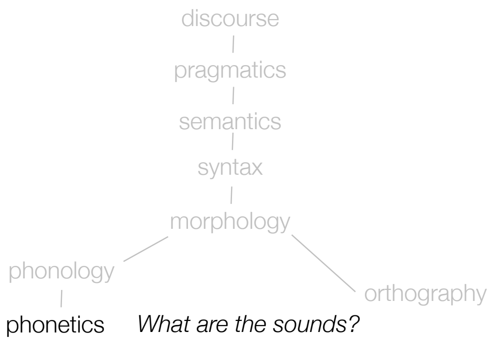
morphology

phonology

phonetics

orthography

Different Levels of Linguistic Knowledge


discourse

pragmatics

semantics

syntax

morphology

phonology

orthography

phonetics *What are the sounds?*

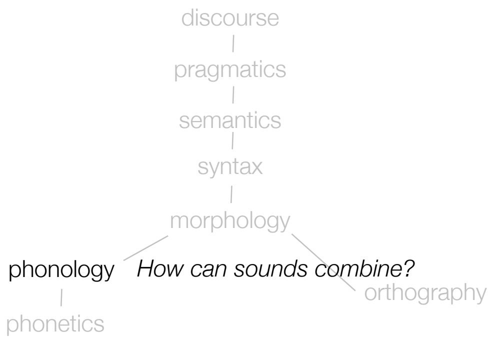
The sound of air moving through my larynx while my tongue is raised in the back of my throat is `iy` (as in teach).

The sound of air gurgling up from my stomach is not speech.

How to distinguish/generate the waveforms for speech sounds?

phonetics *What are the sounds?*

Different Levels of Linguistic Knowledge


discourse

pragmatics

semantics

syntax

morphology

phonology *How can sounds combine?*

orthography

phonetics

"Let's pee in the corner. Let's pee in the spotlight."

www.kissthisguy.com

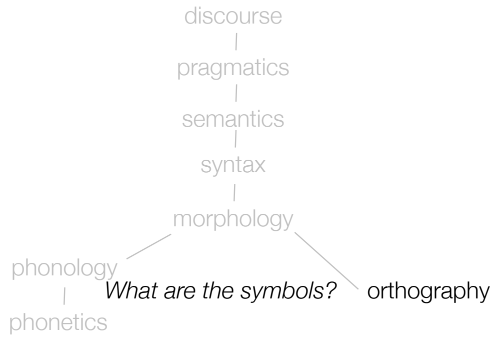
"I left my brains down in Africa"

Sociological factors:
- merry/marry/Mary
- pin/pen/pan/paean
- caught/cot
- goin'/going
- something/suttin'/sumpin'/sumthin'

Words like because and about can be realized many different ways.

phonology *How can sounds combine?*

Different Levels of Linguistic Knowledge



What are the symbols?

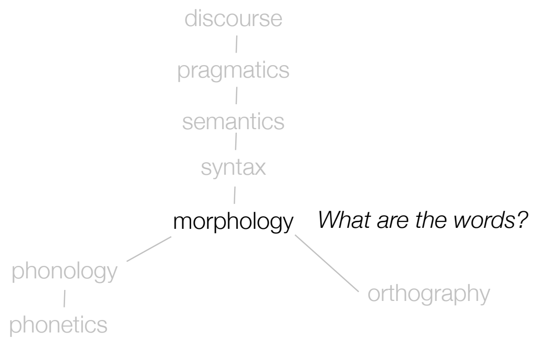
第二阶段的奥运会体育比赛门票与残奥会开闭幕式门票的预订工作已经结束,现在进入门票分配阶段。在此期间,我们不再接受新的门票预订申请。

NATRL LNGUAG PRCSSNG

"Let's go see the praade!"
"They have a mouse in they're house."
"What do you want for desert?"

What are the symbols? orthography

Different Levels of Linguistic Knowledge


discourse
pragmatics
semantics
syntax
morphology *What are the words?*
phonology
phonetics
orthography

fax, google, w00t, OMG, Man-fucking-hattan, lol, lolz, unfriend, tweet, Obamacare, coo af

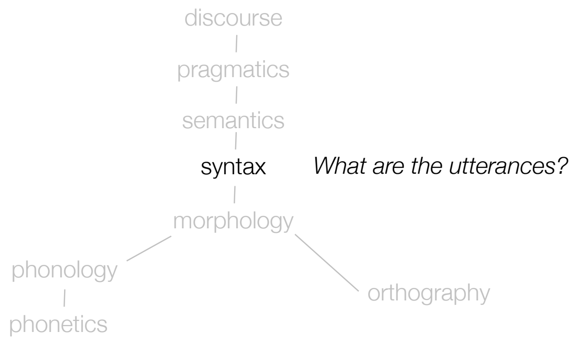
After it sorts each sub-part, it merges them.
After they sort each sub-part, they merge them.
How many merges are needed?
One merge.
Merging is fast.
To split is human, to merge divine.

morphology *What are the words?*

One house among many houses
One mouse among many mouses

uygarlaştıramadıklarımızdanmışsınızcasına
"(behaving) as if you are among those whom we could not civilize"

Different Levels of Linguistic Knowledge

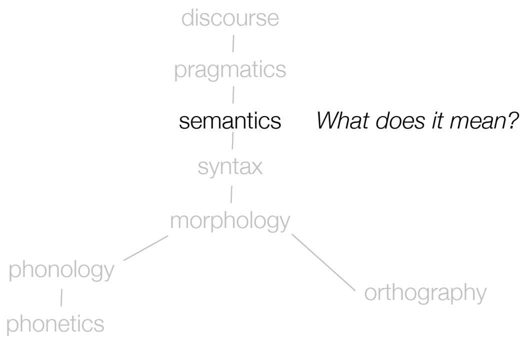

discourse
pragmatics
semantics
syntax *What are the utterances?*
morphology
phonology
phonetics
orthography

Noah gave Kevin the book.
= Noah gave the book to Kevin.
= The book was given to Kevin by Noah.
= The book was given by Noah to Kevin.
*Gave Noah Kevin the book.

syntax *What are the utterances?*

I want a flight to Tokyo.
I want to fly to Tokyo.
I found a flight to Tokyo.
*I found to fly to Tokyo.

Different Levels of Linguistic Knowledge



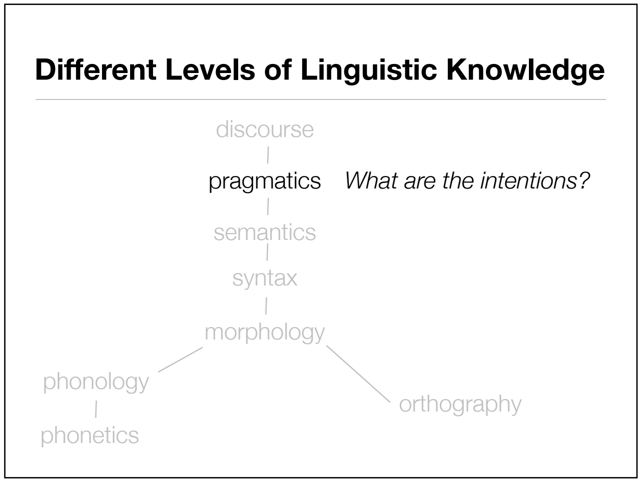
discourse
pragmatics
semantics *What does it mean?*
syntax
morphology
phonology
phonetics
orthography

“Jerusalem - there is no such city!”

In this country a woman gives birth every fifteen minutes.
Our job is to find that woman and stop her.

semantics *What does it mean?*

Colorless green ideas sleep furiously.



Different Levels of Linguistic Knowledge

discourse

pragmatics *What are the intentions?*

semantics

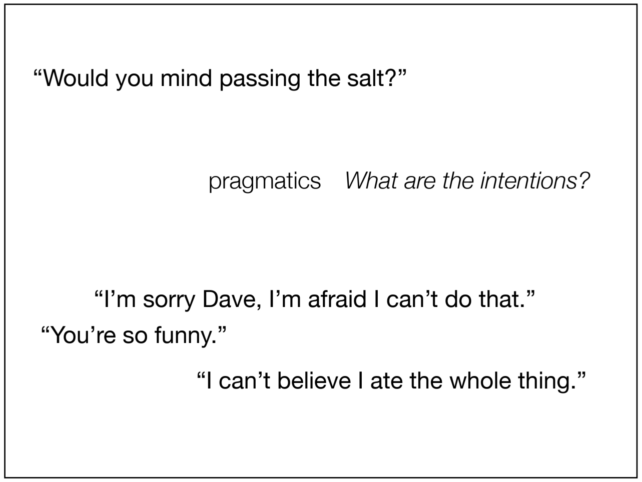
syntax

morphology

phonology

phonetics

orthography



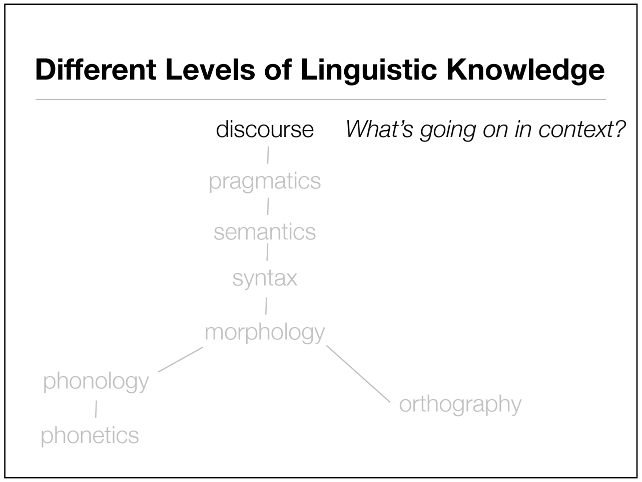
"Would you mind passing the salt?"

pragmatics *What are the intentions?*

"I'm sorry Dave, I'm afraid I can't do that."

"You're so funny."

"I can't believe I ate the whole thing."



Different Levels of Linguistic Knowledge

discourse *What's going on in context?*

pragmatics

semantics

syntax

morphology

phonology

phonetics

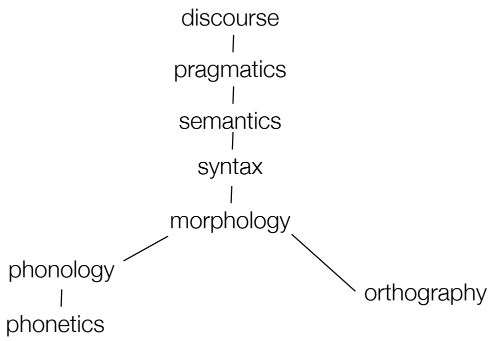
orthography

discourse *What's going on in context?*

The Tin Woodman went to the Emerald City to see the Wizard of Oz and ask for a heart. After he asked for it, the Woodman waited for the Wizard's response.

Any time you got nothing to do - and lots of time to do it - come on up.

Different Levels of Linguistic Knowledge



discourse
pragmatics
semantics
syntax
morphology
phonology
phonetics
orthography

Ambiguity

Students hate annoying professors

From Groucho:

- Last night I shot an elephant in my pajamas. What he was doing in my pajamas I'll never know.

Headlines:

- Kids Make Nutritious Snacks
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Iraqi Head Seeks Arms

From Facebook:

- I'd rather have Kissed a Girl stuck in my head than the Girl from Ipanema.

4,000 living languages

- Research has focused on English
- Most languages beyond reach of NLP:
  - Lack of data
  - Variations in linguistic structure

Linguistic Typology:

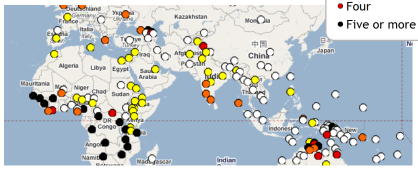
The study of language difference

Subject Verb Object Positioning

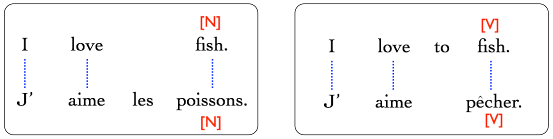
Number of Genders

Definite Article

○ None	(145 languages)
🟡 Two	(50 languages)
🟠 Three	(26 languages)
🔴 Four	(12 languages)
● Five or more	(24 languages)



Variations in Ambiguity



[N]
I love fish.
J' aime les poissons.
[N]

[V]
I love to fish.
J' aime pêcher.
[V]

English: fish (noun) / fish (verb)
French: poissons (noun) / pêcher (verb)

Variations in Ambiguity

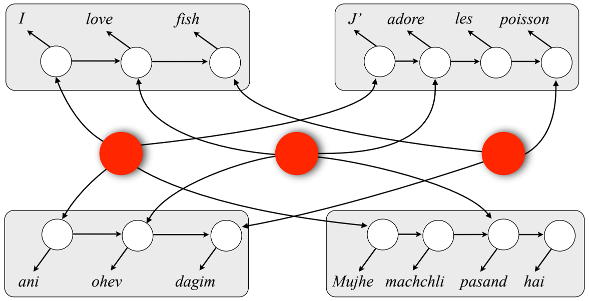
- Differences in morphology

English:	in my country	*separate words*
Hebrew:	בארצי	

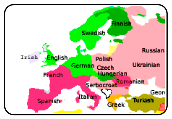
- Differences in syntax

Japanese:	チーズ<u>の</u>スパゲティを食べた	*genitive marker*
English:	I ate pasta with cheese.	

A Multilingual Probabilistic Model



Corpus

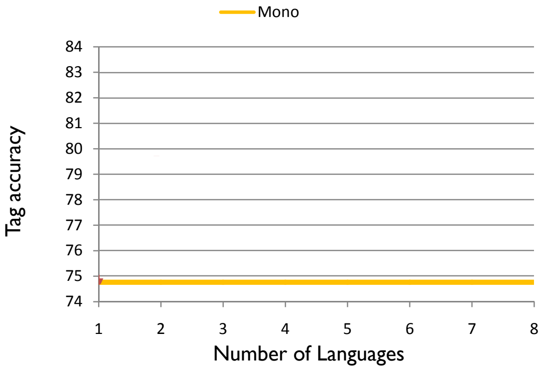


Orwell's *Nineteen Eighty Four* (~100k words)

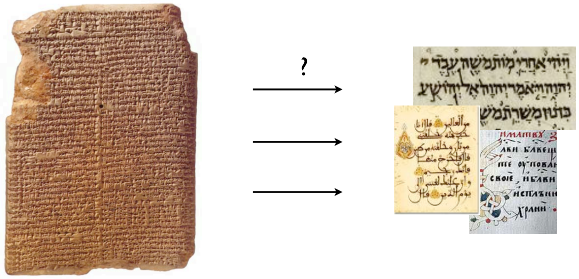
Slavic:	Bulgarian, Czech, Serbian, Slovene
Uralic:	Hungarian, Estonian
Romance:	Romanian
Germanic:	English

Task: Part-of-speech Induction

As we add Languages…



Archeological Decipherment



lost language

known languages

Linguistic Assumptions

- Related languages have *cognates*

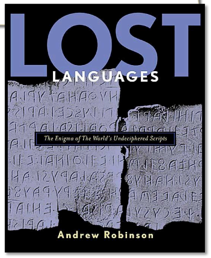
Arabic:	dheker	ذَكَر
Syriac:	dukrā	ܕܘܟܪܐ
Hebrew:	zakhar	זָכָר

fr:	masculin
it:	maschile
sp:	macho

- Systematic mapping between alphabets

ذ ↔ ז ↔ ܕ
(dh) *(z)* *(d)*

successful archaeological decipherment has turned out to require a synthesis of logic and intuition based, as already remarked, on wide linguistic, archaeological, historical and cultural knowledge that computers do not (and presumably cannot) possess.



The Ugaritic Language

Family	: Northwest Semitic
Tablets from	: 14th – 12th century BCE
Discovered	: 1928
Deciphered	: 1932 (by WW1 code breakers)

Large portion of vocabulary covered by cognates with Semitic languages

Arabic:	malik	مَلِك
Syriac:	malkā	ܡܲܠܟܵܐ
Hebrew:	melek	מֶלֶךְ
Ugaritic:	malku	𐎎𐎍𐎋

Corpus: 34,105 tokens, 7,386 unique types

The Epic Of Ba'al

(English translation)

**As soon as El sees Her,
He cracks a smile and laughs.
His feet He sets on the footstool,
And twiddles His fingers.
He lifts His voice
And shouts:
"Why has Lady Asherah of the Sea come?
Why came the Creatress of Gods?
Art Thou hungry?
Then have a morsel!
Or art Thou thirsty?
Then have a drink!
Eat!
Or drink!
Eat bread from the tables!
Drink wine from the goblets!
From a cup of gold, the blood of vines!
If the love of El moves Thee,
Yea the affection of The Bull arouses Thee!"**

**And Lady Asherah of the Sea replied:
"Thou art great, O El,
Thou art verily wise!
The gray of Thy beard hath verily instructed Thee!
Here are pectorals of gold for Thy breast.**

**Lo, also it is the time of His rain.
Baal sets the season,
And gives forth His voice from the clouds.
He flashes lightning to the earth.
As a house of cedars let Him complete it,
Or a house of bricks let Him erect it!
Let it be told to Aliyan Baal:
'The mountains will bring Thee much silver.
The hills, the choicest of gold;
The mines will bring Thee precious stones,
And build a house of silver and gold.
A house of lapis gems!'"**

The Decipherment Task

- Given:
  - Corpus of undeciphered language
  - Lexicon of related language (non-parallel)
- Learn:
  - Alphabetic mapping

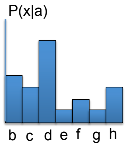
    𐎂 ↔ ג

    𐎄 ↔ ד
  - Word mappings

    𐎎𐎍𐎋 ↔ מלך

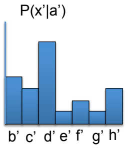
Decipherment Intuition I

- True alphabetic mapping
  ⇒ similar character-level distributions



Ugaritic letters

≈

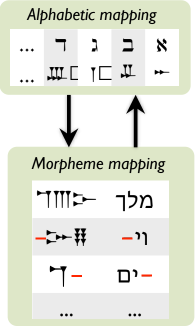


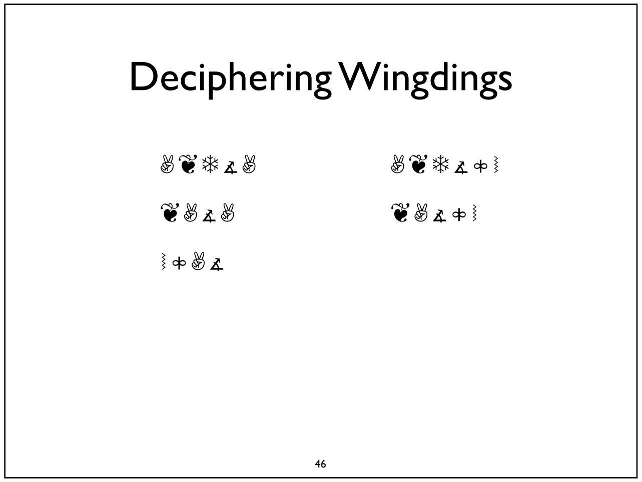
Hebrew letters

Decipherment Intuition II

<u>Interplay between learning:</u>

- Alphabetic mapping
- Higher level morpheme & word correspondences



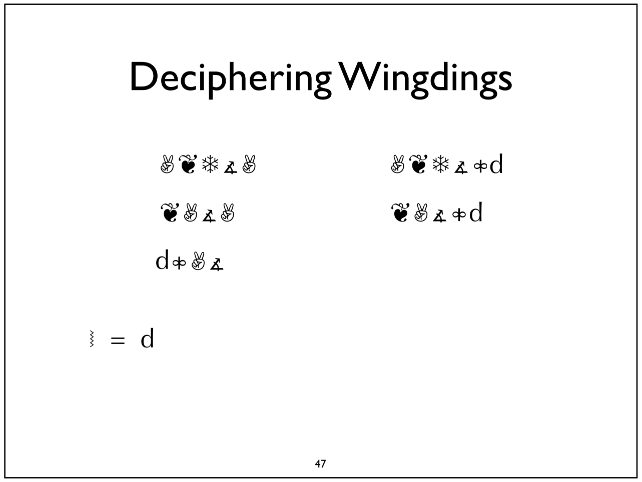


Deciphering Wingdings

✌❦❄⦨✌ ✌❦❄⦨⧞⦚

❦✌⦨✌ ❦✌⦨⧞⦚

⦚⧞✌⦨



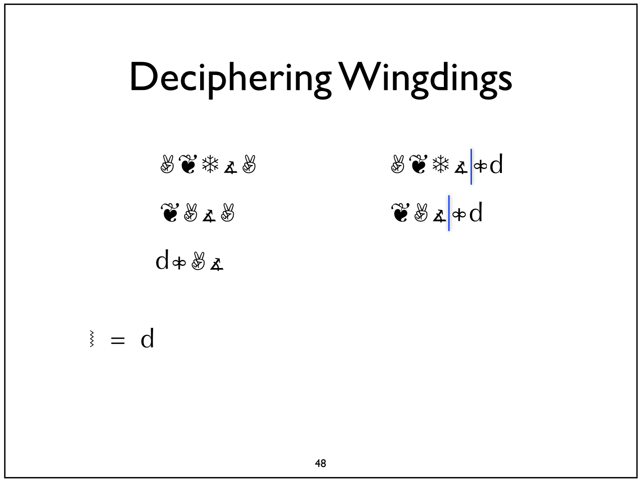
Deciphering Wingdings

✌❦❄⦨✌ ✌❦❄⦨⧞d

❦✌⦨✌ ❦✌⦨⧞d

d⧞✌⦨

⦚ = d



Deciphering Wingdings

✌❦❄⦨✌ ✌❦❄⦨⧞d

❦✌⦨✌ ❦✌⦨⧞d

d⧞✌⦨

⦚ = d

Deciphering Wingdings

✌❦❄⦨✌ ✌❦❄⦨|ed

❦✌⦨✌ ❦✌⦨|ed

de✌⦨

⦚ = d
⧞ = e

Deciphering Wingdings

✌❦❄⦨✌ ✌❦❄⦨|ed

❦✌⦨✌ ❦✌⦨|ed

de✌⦨

⦚ = d
⧞ = e

Deciphering Wingdings

✌❦❄⦨|✌ ✌❦❄⦨|ed

❦✌⦨|✌ ❦✌⦨|ed

de✌⦨

⦚ = d
⧞ = e

Deciphering Wingdings

s ❦ ❄ ∡|s　　s ❦ ❄ ∡|ed

❦ s ∡|s　　❦ s ∡|ed

de s ∡

⦚ = d
ф = e
✌ = s

Deciphering Wingdings

s ❦ ❄ k|s　　s ❦ ❄ k|ed

❦ s k|s　　❦ s k|ed

de s k

⦚ = d
ф = e
✌ = s
∡ = k

Deciphering Wingdings

s a ❄ k|s　　s a ❄ k|ed

a s k|s　　a s k|ed

de s k

⦚ = d　　❦ = a
ф = e
✌ = s
∡ = k

Deciphering Wingdings

s a c k|s　　　s a c k|ed

a s k|s　　　a s k|ed

desk

⌇ = d
⚲ = e
✌ = s
∡ = k

❦ = a
❄ = c

Deciphering Wingdings

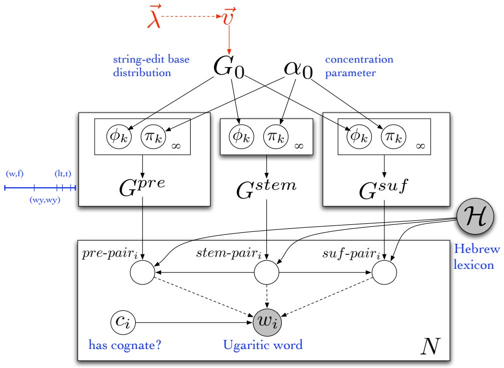
s a c k|s　　　s a c k|ed

a s k|s　　　a s k|ed

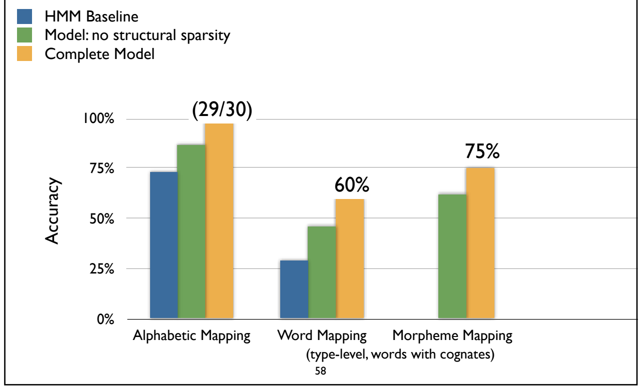
desk

- Used knowledge of English lexicon & morphology (*ask*, *-ed*)
- Discovery of morpheme correspondences ⇔ Discovery of character correspondences

A Probabilistic Decipherment Model

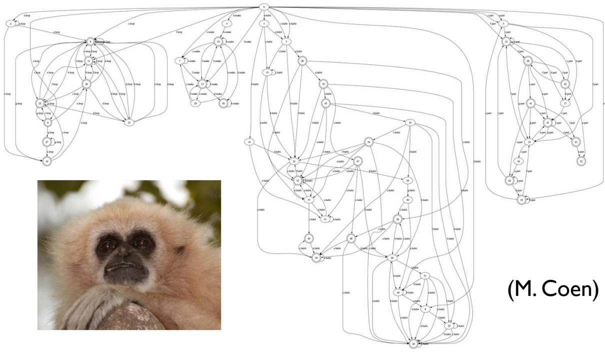


Results



Play Jeopardy (IBM)

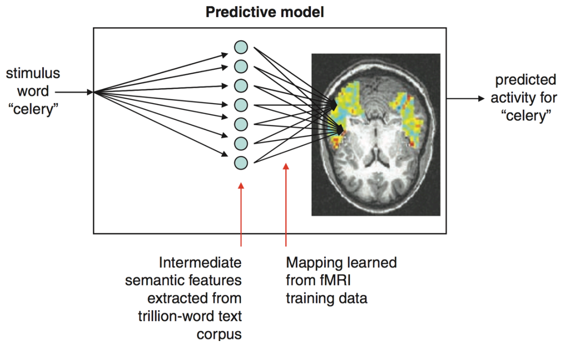
Infer Gibbon Grammar



(M. Coen)

Read your Mind

(T. Mitchell)



Predictive model

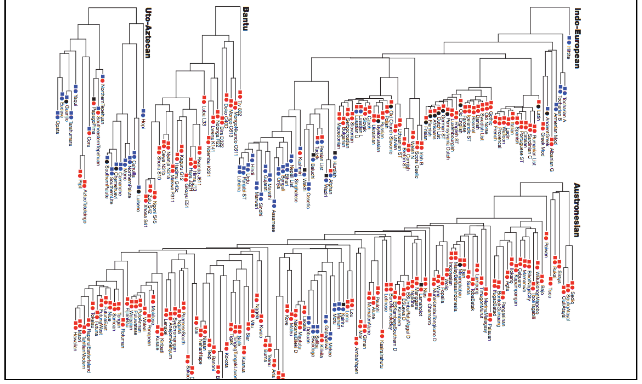
stimulus word "celery"

predicted activity for "celery"

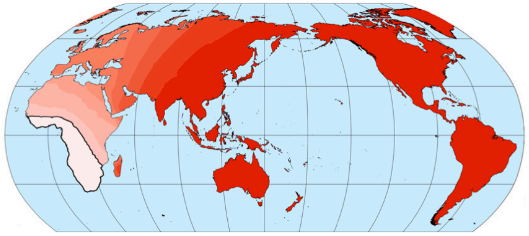
Intermediate semantic features extracted from trillion-word text corpus

Mapping learned from fMRI training data

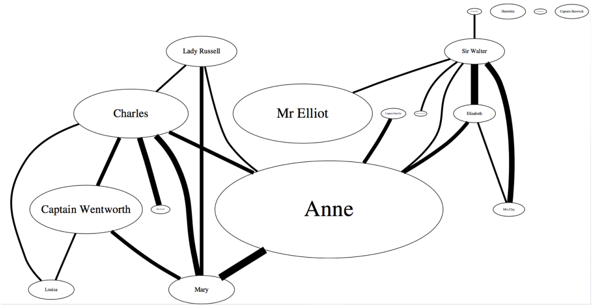
Trace Language History



Locate Origin of Human Language

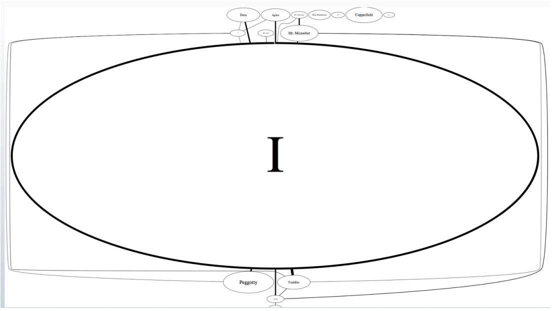


Analyze Fiction



Persuasion by Jane Austen

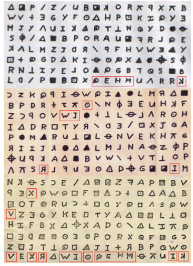
Analyze Fiction



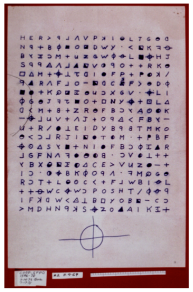
David Copperfield by Charles Dickens

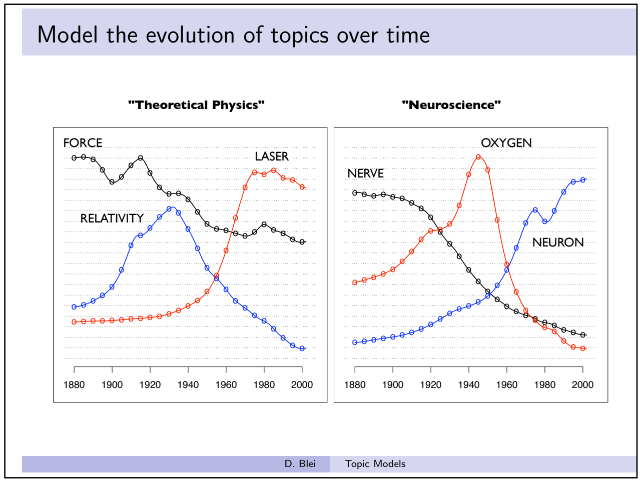
Decipher Messages from Serial Killers

Zodiac 408 (solved, 1969, by couple at breakfast)



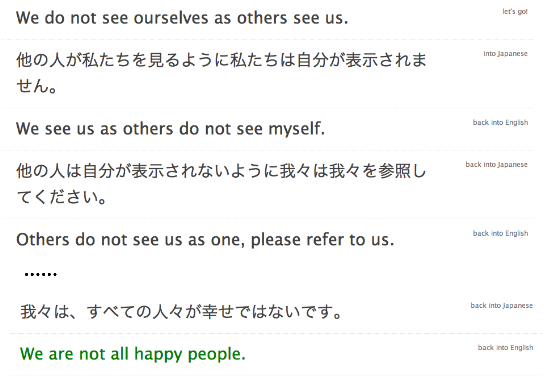
Zodiac 340 (still unsolved)





Model the evolution of topics over time

Translate (?)



We do not see ourselves as others see us. let's go!

他の人が私たちを見るように私たちは自分が表示されません。 into Japanese

We see us as others do not see myself. back into English

他の人は自分が表示されないように我々は我々を参照してください。 back into Japanese

Others do not see us as one, please refer to us. back into English

......

我々は、すべての人々が幸せではないです。 back into Japanese

We are not all happy people. back into English