Introduction to Natural Language Processing

Markus Dickinson
Linguistics
@georgetown.edu
With some modifications

Languages and Intelligence

- **Languages:** Natural (mandarin) and Artificial (prolog/lisp), and in between?
- Processing of Computer languages is easy: why?
- Intelligence: Natural? and Artificial (AI).
- What is AI?: behave as humans do! Whatever that means! Turing Test
- Language related activities: acquisition, speech,… manifestation of intelligence

Languages and Intelligence

- Historically: peaks and valleys.
- Tendency to underestimate the problem and overestimate the computing power.
- NLP Goes along with AI: Rise and Decline
- Now is a peak, 20 years ago may have been a valley.
- The internet, small devices may be a driving force.

What is NLP?

- **Natural Language Processing (NLP)**
  - Computers use to process (analyze, understand, generate) natural language
  - A somewhat applied field
- Computational Linguistics (CL)
  - Computational aspects of the human language faculty
  - More theoretical

Why Study NLP?

- Human language interesting & challenging
  - NLP offers insights into language
- Language is the medium of the web
- Interdisciplinary: Linguistics, CS, psychology, math, EE,
- Help in communication
  - With computers (ASR, TTS)-Interfaces-HCI
  - With other humans (MT)
- Ambitious yet practical

Goals of NLP

- Scientific Goal
  – *Identify the computational machinery needed for an agent to exhibit various forms of linguistic behavior*
- Engineering Goal
  – *Design, implement, and test systems that process natural languages for practical applications*
- Ups and downs-Historically: more with the web

Applications

- speech processing: *get flight information or book a hotel over the phone, TTS (for the blind)*
- information extraction: *discover names of people and events they participate in, from a document*
- machine translation: *translate a document from one human language into another*
- question answering: *find answers to natural language questions in a text collection or database*
- summarization: *generate a short biography of Noam Chomsky from one or more news articles*
- OCR: both print and handwritten.
- More inportant with Web and Mobile devices!

General Themes

- Ambiguity of Language
- Language as a formal system
- Rule-based vs. Statistical Methods
- The need for efficiency
- Syntax/Semantics/Data mining
- Multilingual across/languages Support

Ambiguity of language

- Phonetic
  - [raIt] = *write*, *right*, *rite* *(Soundex)* سائد-صائد
- Lexical
  - *can* = noun, verb, modal (ذهب)
- Structural
  - *I saw the man with the telescope* رأيته فقط بالنظارة
- Semantic

→ *dish* = physical plate, menu item(مشروع)

- Referential: The son asked the father to drive him home
  - (طلبت الأم من البنت تصفيف شعرها)

→ All of these make NLP difficult

Language as a formal system

- We can treat parts of language formally
  - Language = a set of acceptable strings in an alphabet
  - Define a model to recognize/generate language (but recall: language before grammar)
- Works for different levels of language (phonology, morphology, etc.)
- Can use finite-state automata, context-free grammars, etc. to represent language

Rule-based & Statistical Methods

- Theoretical linguistics captures abstract properties of language
- NLP can more or less follow theoretical insights
  - Rule-based: model system with linguistic rules
  - Statistical: model system with probabilities of what normally happens: trust what people usually do/write/say
- Hybrid models combine the two

The need for efficiency

- Simply writing down linguistic insights isn't sufficient to have a working system
- Programs need to run in real-time, i.e., be efficient
  - There are thousands of grammar rules which might be applied to a sentence
- Use insights from computer science
  - To find the best parse, use chart parsing, a form of dynamic programming
- Recall: computers are powerful but "never" as powerful as need be.

Preview of Topics

1. ***Finding Syntactic Patterns in Human Languages: Lg. as Formal System***
2. Meaning from Patterns
3. Patterns from Language in the Large
4. Bridging the Rationalist-Empiricist Divide
5. Applications
6. Conclusion

The Problem of Syntactic Analysis

- Assume input sentence S in natural language L
- Assume you have rules (*grammar* G) that describe syntactic regularities (patterns or structures) found in sentences of L
- Given S & G, find syntactic structure of S
- Such a structure is called a parse tree

Example 1

S → NP VP

VP → V NP

VP → V

NP → I

NP → he

V → slept

V → ate

V → drinks

S → NS| VS

NS → M K

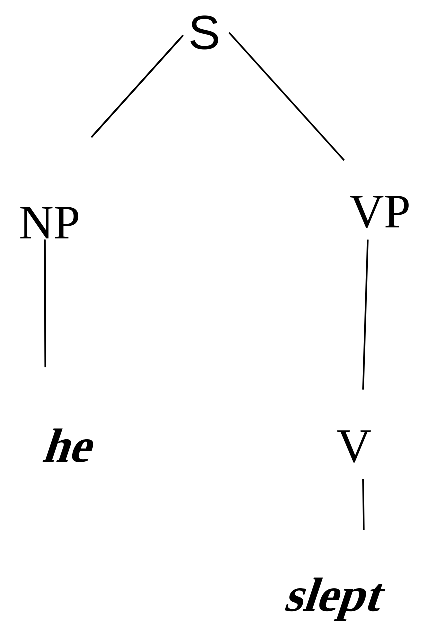
M → N

K → N

VS → V NP NP

NP → N|NS

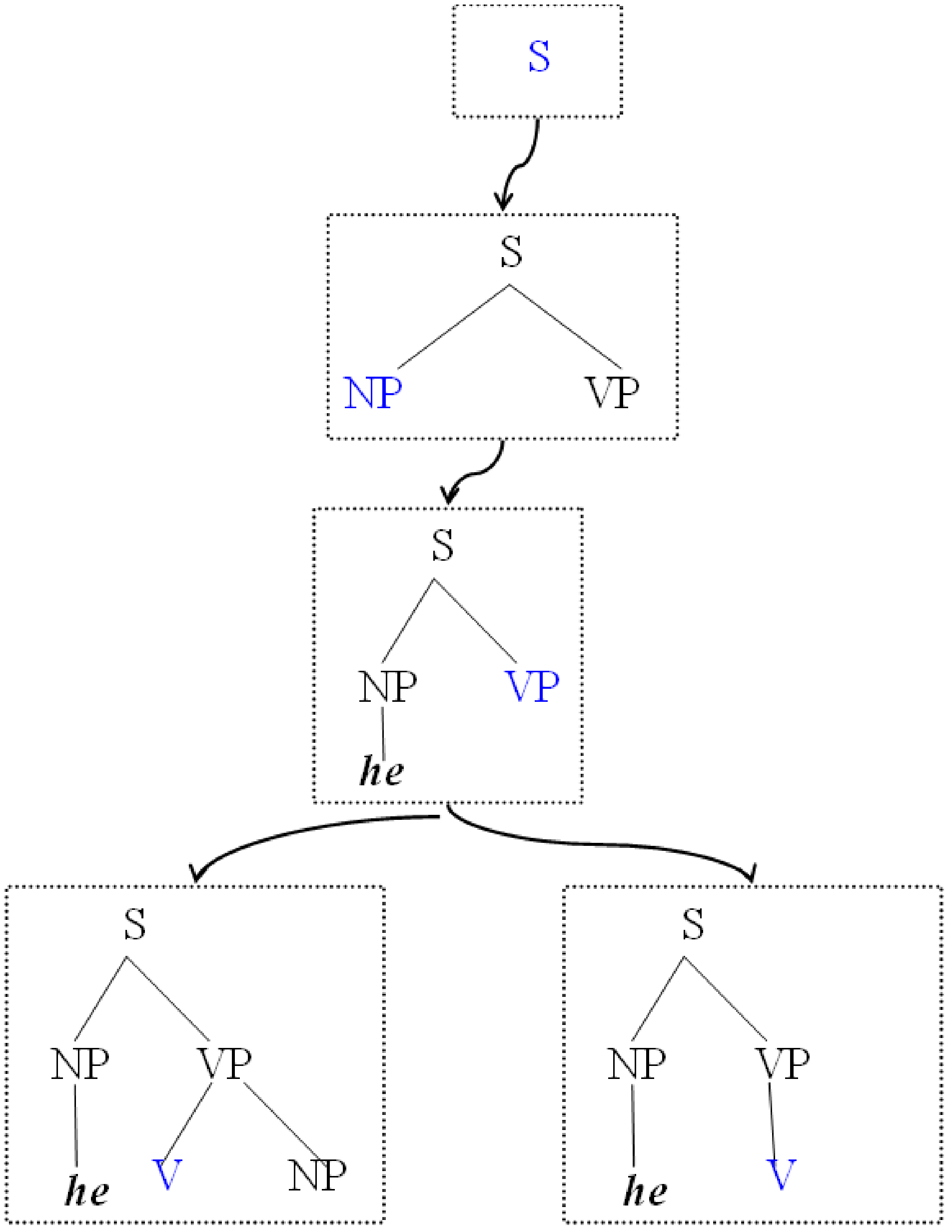
Grammar



Parse Tree

Parsing Example 1

- S → NP VP
- VP → V NP
- VP → V
- NP → I
- *NP → he*
- *V → slept*
- *V → ate*
- *V → drinks*



More Complex Sentences

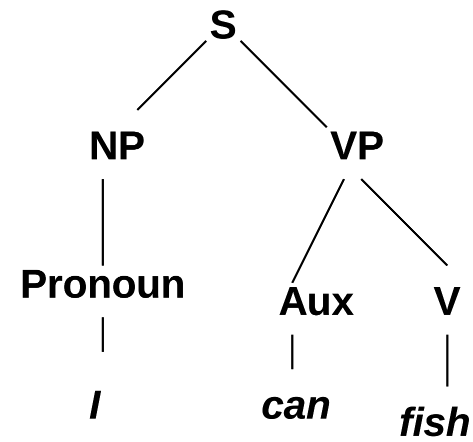
- *I can fish.*
- *I saw the elephant in my pajamas.*

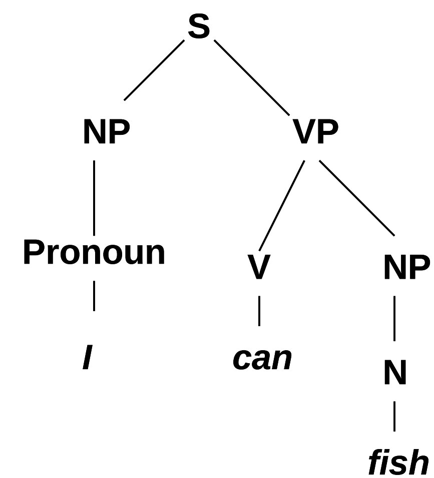
- These sentences exhibit ambiguity
- Computers will have to find the acceptable or most likely meaning(s).
- *I can fish. (يستطيع، يعلب؟)*

Example 2

- S → NP VP
- VP → Aux V
- VP → V NP
- VP → V
- VP → Aux V NP
- NP → D N
- NP → N
- NP → Pronoun

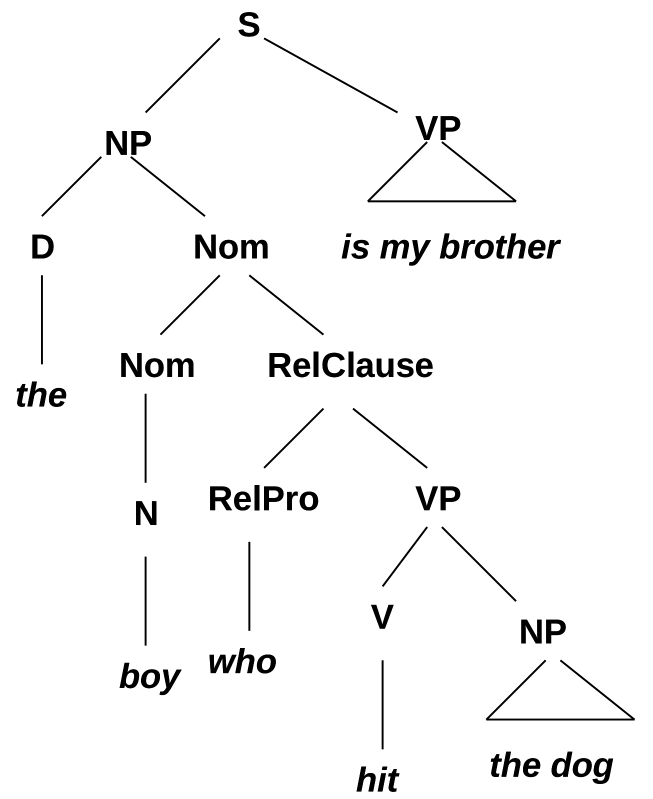
V → can
V → fish
V → dance
Aux → can
D → the
N → fish
N → dance
Pronoun → I





Example 3

- NP → D Nom
- Nom → Nom RelClause
- Nom → N
- RelClause → RelPro VP
- VP → V NP
- *D → the*
- *D → my*
- *V → is*
- *V → hit*
- *N → dog*
- *N → boy*
- *N → brother*
- *RelPro → who*



Topics

1. Finding Syntactic Patterns in Human Languages

2. Meaning from Patterns

3. Patterns from Language in the Large

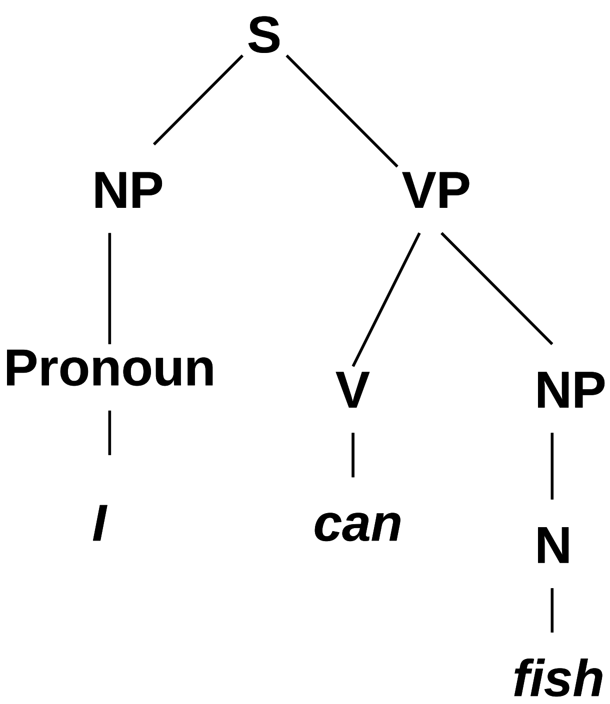
4. Bridging the Rationalist-Empiricist Divide

5. Applications

6. Conclusion

Meaning from a Parse Tree

- *I can fish.*
- We want to understand
  - Who does what?
    - the *canner* is me, the *action* is canning, and the *thing canned* is fish.
    - e.g. Canning(ME, FishStuff)
    - This is a logic representation of meaning

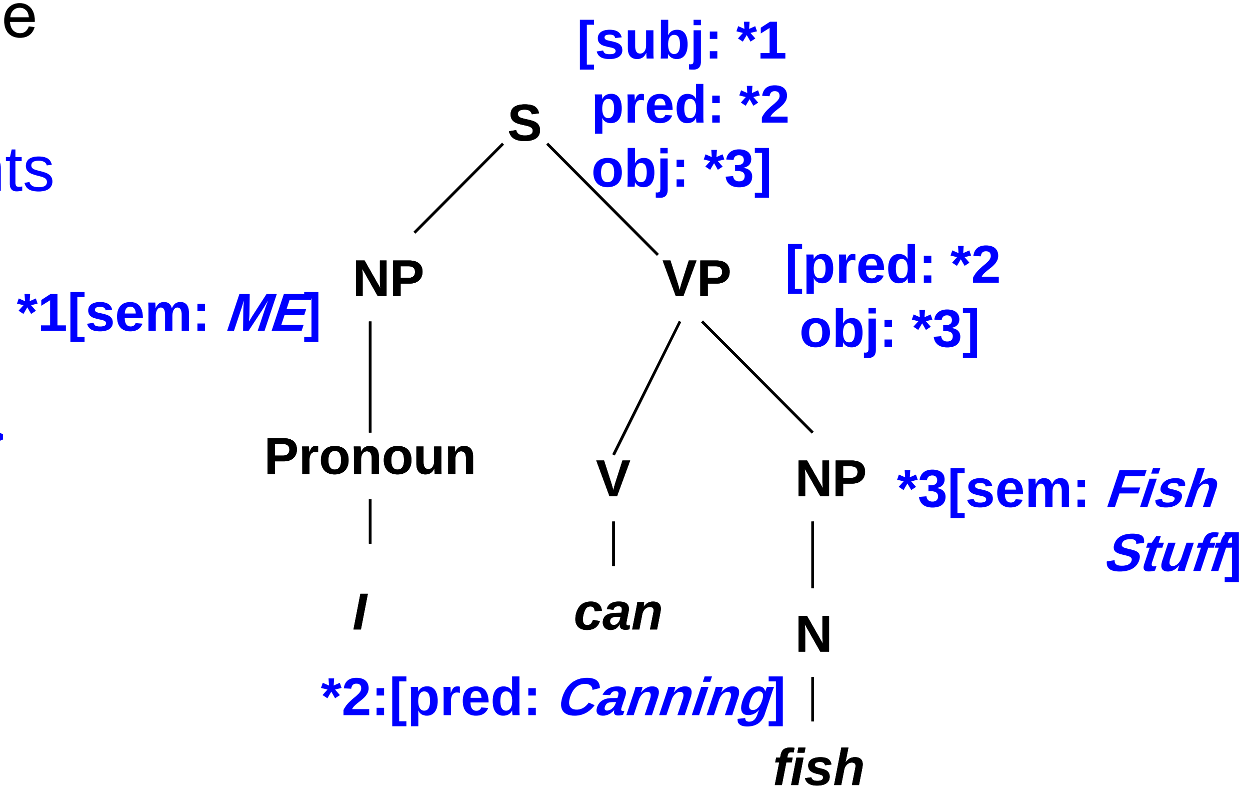


We can do this by

- associating meanings with lexical items in the tree
- then using rules to figure out what the S as a whole means

Meaning from a Parse Tree (Details)

- Let’s augment the grammar with feature constraints
- S → NP VP
  - <S subj> =<NP>
  - <S>=<VP>
- VP→ V NP
  - <VP> = <V>
  - <VP obj> =<NP>



Grammar Induction

- Start with a tree bank = collection of parsed sentences
- Extract grammar rules corresponding to parse trees, estimating the probability of the grammar rule based on its frequency

  $P(A \rightarrow \beta \mid A) = \text{Count}(A \rightarrow \beta) / \text{Count}(A)$

- You then have a probabilistic grammar, derived from a corpus of parse trees

- *How does this grammar compare to grammars created by human intuition?*
- *How do you get the corpus?*

Finite-State Analysis

We can also "cheat" a bit in our linguistic analysis

A finite-state machine for recognizing NPs:

- initial=0; final ={2}
- 0->N->2
- 0->D->1
- 1->N->2
- 2->N->2

An equivalent regular expression for NP's

/D? N$^+$/

A regular expression for recognizing simple sentences

/(Prep D? A* N$^+$)* (D? N) (Prep D? A* N$^+$)* (V_tns|Aux V_ing) (Prep D? A* N$^+$)*/

Topics

1. Finding Syntactic Patterns in Human Languages
2. Meaning from Patterns
3. ***Patterns from Language in the Large***
4. Bridging the Rationalist-Empiricist Divide
5. Applications
6. Conclusion

Empirical Approaches to NLP

- *Empiricism*: knowledge is derived from experience
- *Rationalism:* knowledge is derived from reason
- NLP is, by necessity, focused on ‘performance’, in that naturally-occurring linguistic data has to be processed
  – Have to process data characterized by false starts, hesitations, elliptical sentences, long and complex sentences, input in a complex format, etc.
- The methodology used is corpus-based
  – linguistic analysis (phonological, morphological, syntactic, semantic, etc.) carried out on a fairly large scale
  – rules are derived by humans or machines from looking at phenomena in situ (with statistics playing an important role)

Which Words are the Most Frequent?

Common Words in *Tom Sawyer (71,730 words)*, from Manning & Schutze p.21

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

- Will these counts hold in a different corpus (and genre, cf. Tom)?
- What happens if you have 8-9M words?

Data Sparseness

Word Frequency	**Number of words of that frequency**
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
>100	102

- Many low-frequency words
- Fewer high-frequency words.
- Only a few words will have lots of examples.
- About 50% of word types occur only once
- Over 90% occur 10 times or less.

Frequency of word types in *Tom Sawyer*, from M&S 22.

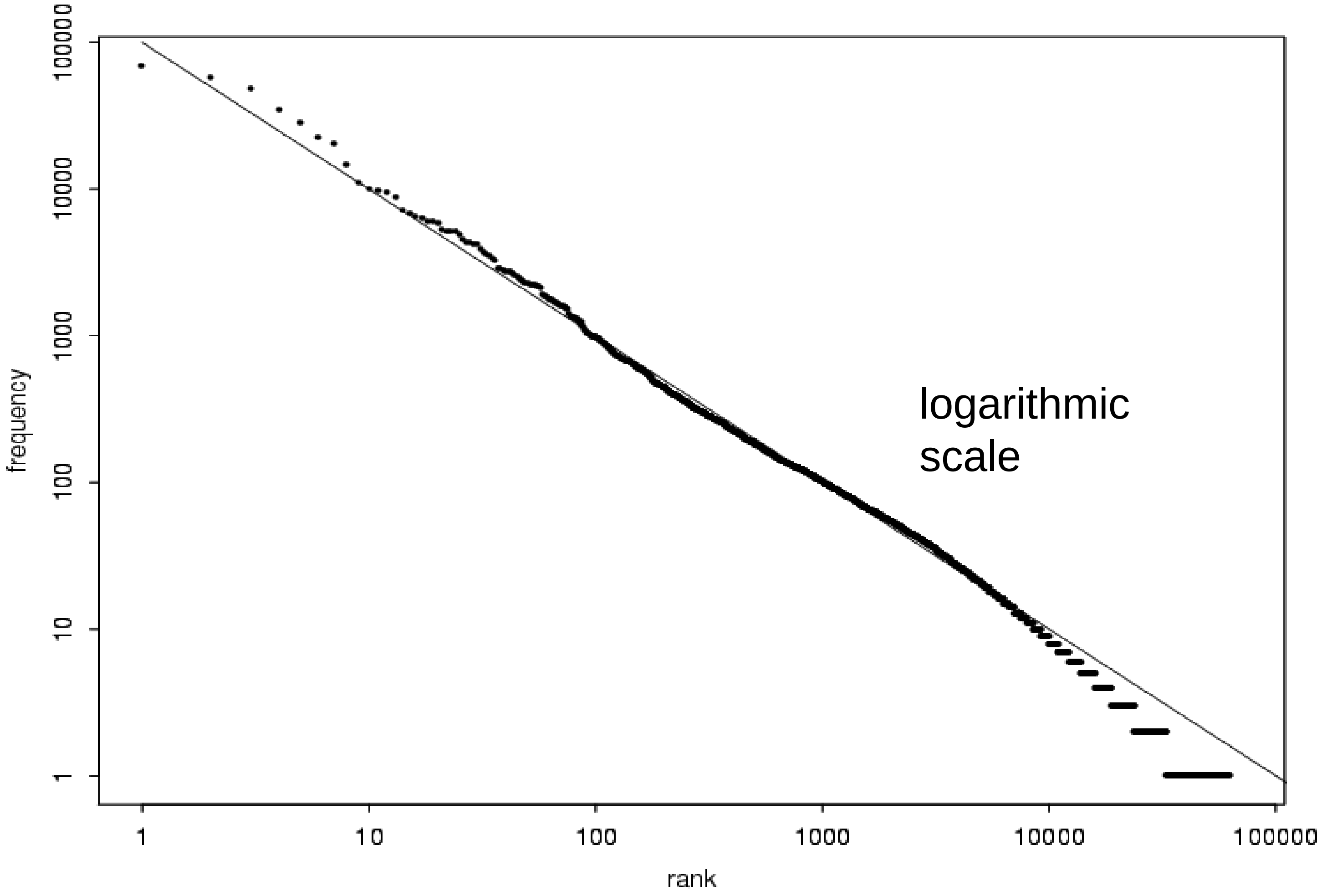
Zipf's Law: *Frequency is inversely proportional to rank*

Word	Freq f	Rank r	f.r
the	3332	1	3332
and	2972	2	5944
a	1775	3	5325
he	877	10	8770
but	410	20	8200
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
oh	116	90	10440
two	104	100	10400

turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000
could	2	4000	8000
applausive	1	8000	8000

Empirical evaluation of Zipf's Law on *Tom Sawyer*, from M&S 23.

Illustration of Zipf's Law



(Brown Corpus, from M&S p. 30)

Empiricism: Part-of-Speech Tagging

- Word statistics are only so useful
- We want to be able to deduce linguistic properties of the text
- Part-of-speech (POS) Tagging = assigning a POS (lexical category) to every word in a text
  - Words can be ambiguous
  - What is the best way to disambiguate?

Part-of-Speech Disambiguation

Secretariat/NNP is/VBZ expected/VBN to/TO race/VB tomorrow/NN

The/DT reason/NN for/IN the/DT race/NN for/IN outer/JJ space/NN is

…

- Given a sentence W1…Wn and a tagset of lexical categories, find the most likely tag C1..Cn for each word in the sentence
- Tagset – e.g., Penn Treebank (45 tags)
- Note that many of the words may have unambiguous tags
- The tagger also has to deal with unknown words

Penn Tree Bank Tagset

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

A Statistical Method for POS Tagging

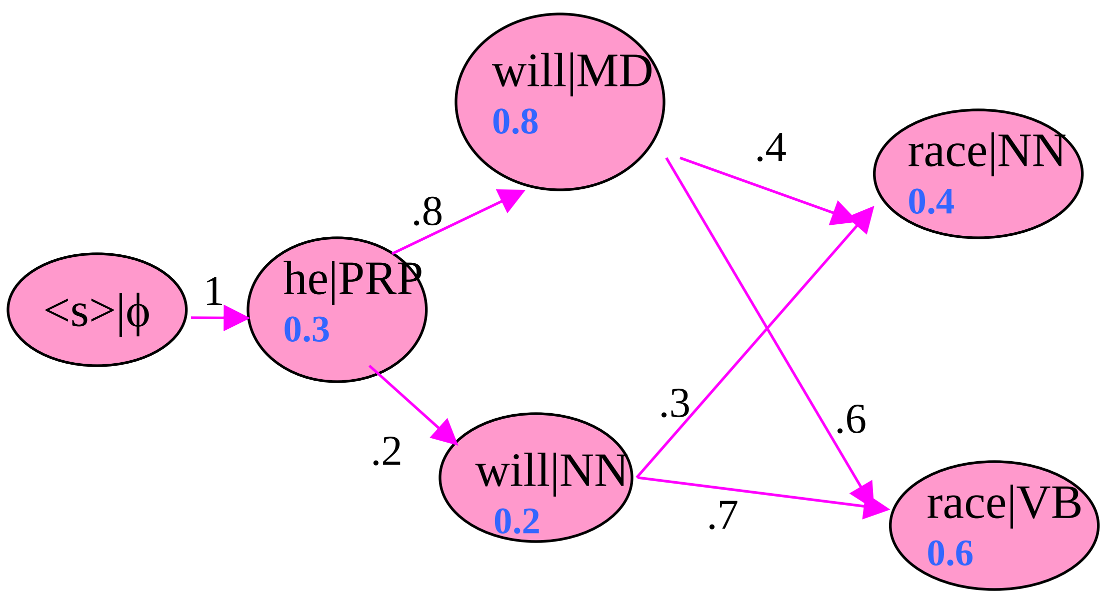
Find the value of C1..Cn which maximizes:

$$\Pi_{i=1,\,n} \quad P(Wi|\ Ci) \quad * \quad P(Ci|\ Ci\text{-}1)$$

lexical generation probabilities — *POS bigram probabilities*

	MD	NN	VB	PRP
he	0	0	0	.3
will	.8	.2	0	0
race	0	.4	.6	0

lexical generation probs



C|R	MD	NN	VB	PRP
MD		.4	.6	
NN		.3	.7	
PRP	.8	.2		
ϕ				1

POS bigram probs

Chomsky's Critique of Corpus-Based Methods

1. Corpora model performance, while linguistics is aimed at the explanation of competence

If you define linguistics that way, linguistic theories will never be able to deal with actual, messy data

2. Natural language is in principle infinite, whereas corpora are finite, so many examples will be missed

Excellent point, which needs to be understood by anyone working with a corpus.

But does that mean corpora are useless?

- Introspection is unreliable (prone to performance factors, cf. only short sentences), and pretty useless with child data.
- Insights from a corpus might lead to generalization/induction beyond the corpus– if the corpus is a good sample of the "text population"

3. Ungrammatical examples won't be available in a corpus

Depends on the corpus, e.g., spontaneous speech, language learners, etc.

The notion of grammaticality is not that clear

- *Who did you see [pictures/?a picture/??his picture/*John's picture] of?*

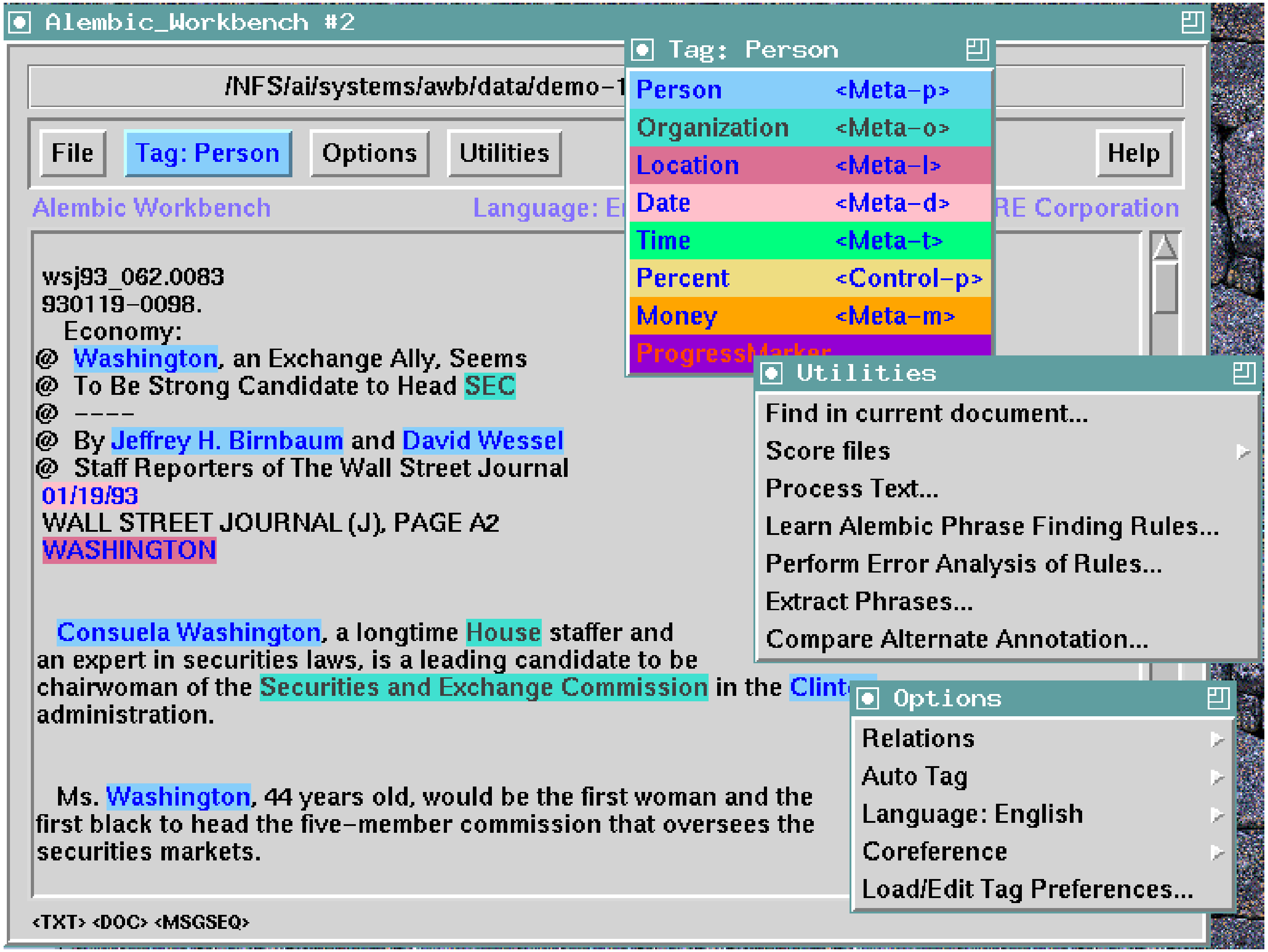
Topics

1. Finding Syntactic Patterns in Human Languages
2. Meaning from Patterns
3. Patterns from Language in the Large
4. ***Bridging the Rationalist-Empiricist Divide***
5. Applications
6. Conclusion

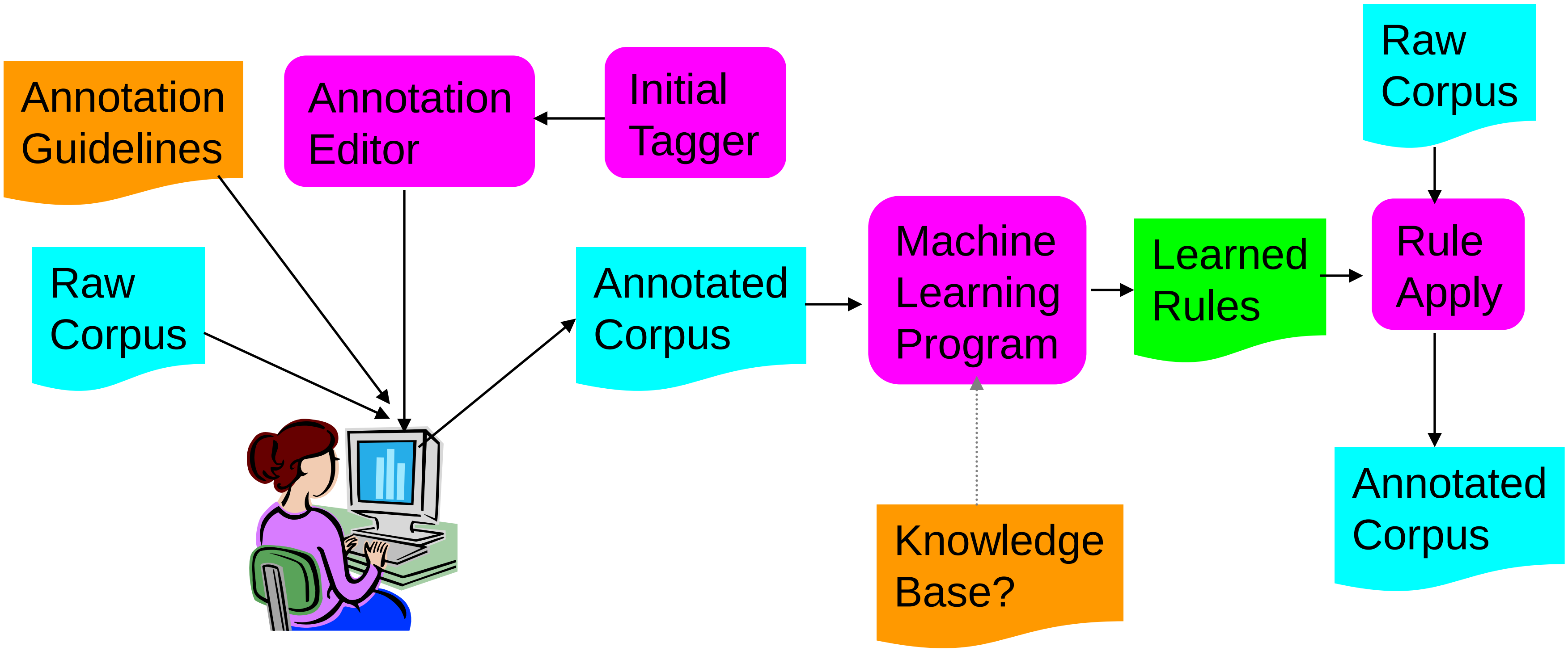
The Annotation of Data

- If we want to learn linguistic properties from data, we need to annotate the data
  - Train on annotated data
  - Test methods on other annotated data
- Through the annotation of corpora, we encode linguistic information in a computer-usable way.

An Annotation Tool



Knowledge Discovery Methodology



Topics

1. Finding Syntactic Patterns in Human Languages
2. Meaning from Patterns
3. Patterns from Language in the Large
4. Bridging the Rationalist-Empiricist Divide

5. Applications

6. Conclusion

Application #1: Machine Translation

- Using different techniques for linguistic analysis, we can:
  - Parse the contents of one language
  - Generate another language consisting of the same content
  - If we have an intermediate language then we may be able to translate between pairs
  - Success modest: check Google translate!

Machine Translation on the Web

http://complingone.georgetown.edu/~linguist/GU-CLI/GU-CLI-home.html

- متظاهران برضوض وجروح في قرية المعصرة الى الجنوب من بيت لحم جراء الاعتداء عليهما من اصيب
تظاهران برضوض وجروح في قرية المعصرة الى الجنوب من بيت لحم جراء الاعتداء عليهما من قبل قوات
لقت المسيرة من امام .الاحتلال الاسرائيلي خلال المسيرة الاسبوعية التي تنظم ضد جدار الفصل العنصري
مدرسة القرية الثانوية لتجوب شوارعها بمشاركة العشرات من المتظاهرين العرب والاجانب وخلال ذلك رفع
المتظاهرون الاعلام الفلسطينية واللافتات المنددة بالاجراءات والممارسات الاسرائيلية المتواصلة بحق الشعب
ي ، كما رفع المتظاهرون صور عدد من الاسرى في اشارة الى محورية قضية الاسرى لدى الشعب الفلسطيني
ى وصول المسيرة الى مشارف اقامة جدار الفصل العنصري واجهها العشرات من جنود الاحتلال الاسرائيلي

- Injured demonstrators sustained bruises and cuts to the mill in the villag
south of Bethlehem in the attack on them by the Israeli occupation forc
during the weekly march organized against the Apartheid Wall.

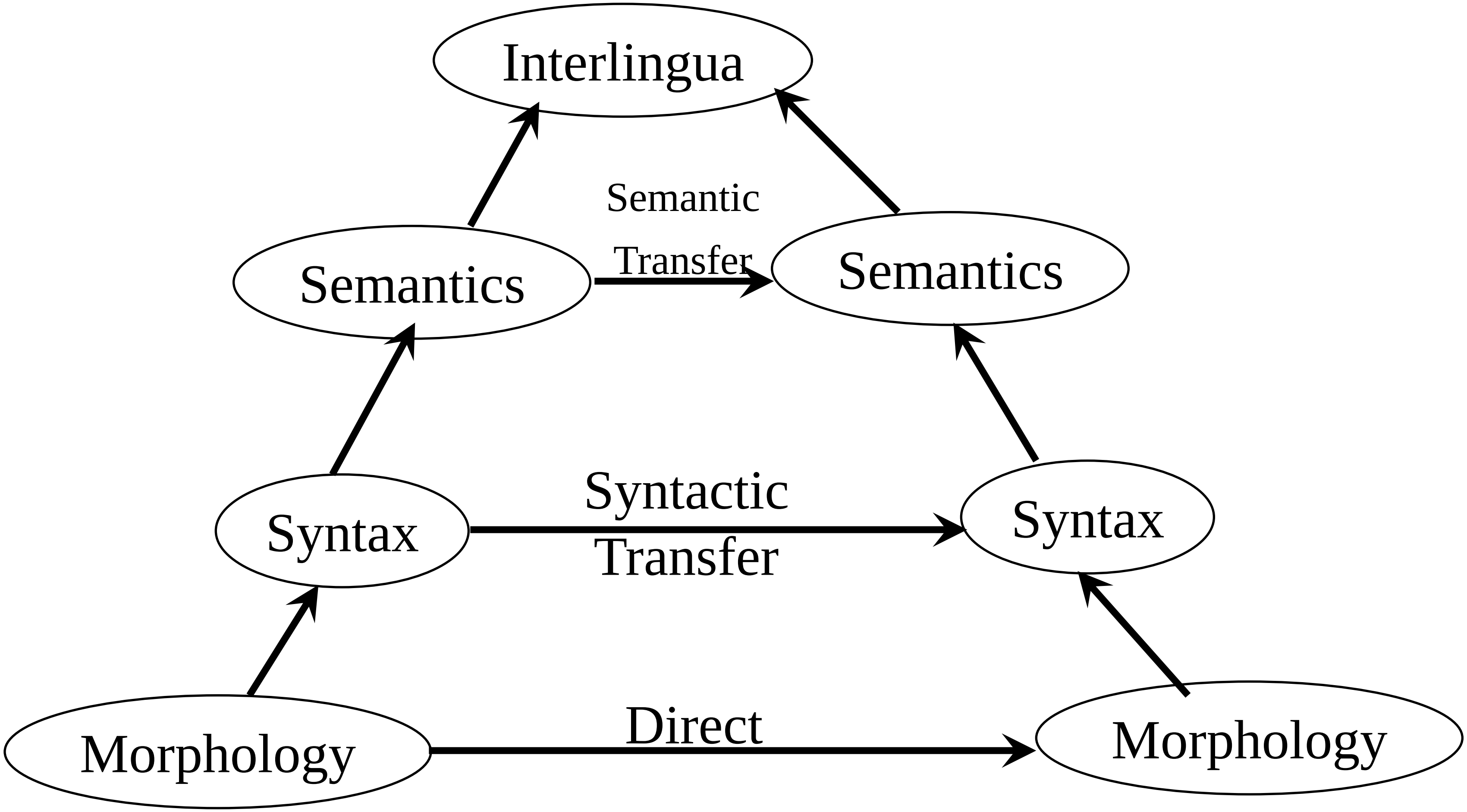
- The march began in front of the village school secondary to roam the str
with dozens of demonstrators Arabs and foreigners, and during that
demonstrators raised Palestinian flags and banners condemning the actio
and practices of Israel's continued right of the Palestinian people, as the
protesters portrayed a number of prisoners in reference to the central i
of the prisoners to the Palestinian people and the arrival of the march to
outskirts of the establishment of the apartheid wall and faced dozens of
Israeli soldiers

If languages were all very similar….

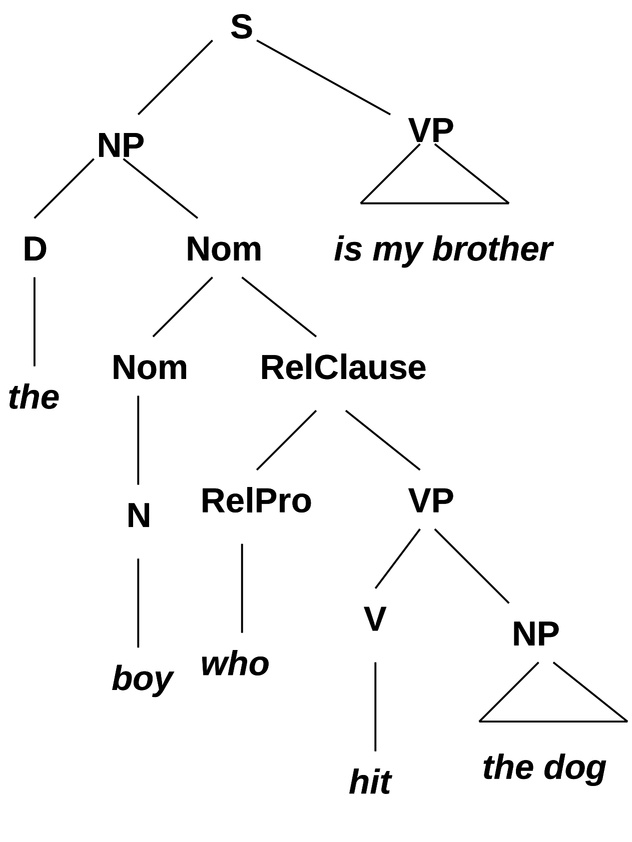
… then MT would be easier

- Dialects
  - http://rinkworks.com/dialect/
- Spanish to Portuguese….
- Spanish to French
- English to Japanese
- ………….

MT Approaches



MT Using Parallel Treebanks



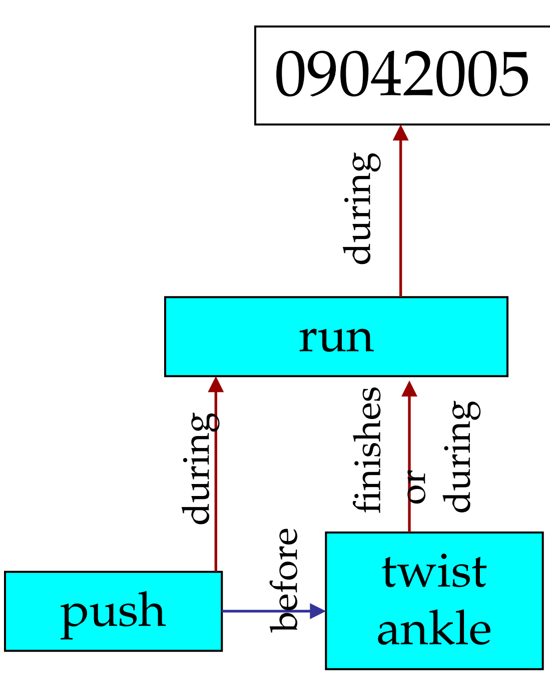
Application #2: Understanding a Simple Narrative (Question Answering)

Yesterday Holly was running a marathon when she twisted her ankle. David had pushed her.

1. *When did the running occur?*
2. *When did the twisting occur?*
3. *Did the pushing occur before the twisting?*
4. *Did Holly keep running after twisting her ankle?*

Question Answering by Computer (Temporal Questions)

Yesterday Holly was running a marathon when she twisted her ankle. David had pushed her.



1. When did the running occur?
Yesterday.
2. When did the twisting occur?
Yesterday, during the running.
3. Did the pushing occur before the twisting?
Yes.
4. Did Holly keep running after twisting her ankle?
Maybe not????

Application #3: Information Extraction

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Cp., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

⇨

$Company_{NG}$ $Set\text{-}UP_{VG}$ $Joint\text{-}Venture_{NG}$ with $Company_{NG}$

$Produce_{VG}$ $Product_{NG}$

KEY:

Trigger word tagging

Named Entity tagging

Chunk parsing: NGs, VGs, preps, conjunctions

Information Extraction: Filling Templates

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

Activity:
- Type: PRODUCTION
- Company:
- Product: *golf clubs*
- Start-date:

The joint venture, Bridgestone Sports Taiwan Cp., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

Activity:
- Type: PRODUCTION
- Company: *Bridgestone Sports Taiwan Co*
- Product: *iron and "metal wood" clubs*
- Start-date: DURING *1990*

Conclusion

- NLP programs can carry out a number of very interesting tasks
  - Part-of-speech disambiguation
  - Parsing
  - Information extraction
  - Machine Translation
  - Question Answering
- These programs have impacts on the way we communicate
- These capabilities also have important implications for cognitive science