

Seminar in NLP

החלום



わたしは「おばあさん」
です。 あなたは？

watashi wa
"Obaasan" desu.
Anata wa?

אני אובסן,
ואתה?



CICLing Conference

Devoted to

- computational linguistics (CL),
- intelligent text processing,
- natural language processing (NLP),
- human language technologies (HLT),
- natural-language human-computer interaction (HCI), and
- speech processing and speech recognition (SR).

Topics

- text processing,
- computational morphology,
- tagging,
- stemming,
- syntactic analysis,
- parsing and shallow parsing,
- chunking,
- recognizing textual entailment,
- ambiguity resolution,
- semantic analysis,
- pragmatics,
- lexicon,
- lexical resources,
- dictionaries and machine-readable dictionaries (MRD),

Topics -- Continued

- grammar,
- anaphora resolution,
- word sense disambiguation (WSD),
- machine translation (MT),
- information retrieval (IR),
- information extraction (IE),
- document handling,
- document classification and text classification,
- text summarization,
- text mining (TM), and
- spell checking (spelling).

Definitions

Linguistics

- **phonology**: study sound patterns of language
- **morphology**: study structure and meaning of words
- **syntax**: study sentence structure
- **semantics**: study linguistic meaning

CL vs NLP

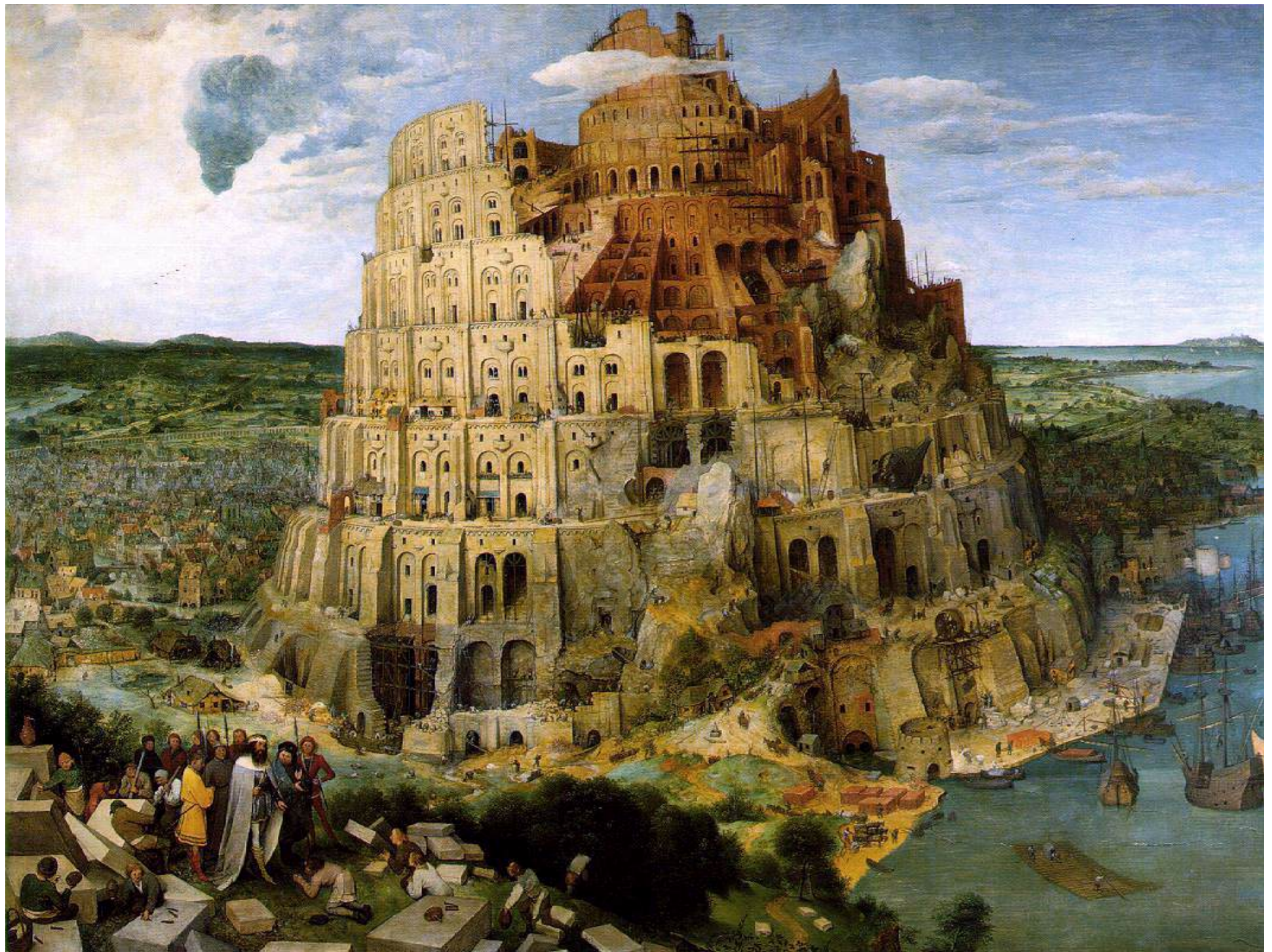
- Computational Linguistics (CL)
 - Modeling people
 - Computer-based tools
- Natural Language Processing (NLP)
 - Computer-based applications
 - Linguistic tools

CL

- Using computers for the scientific study of human language
 - Historically, associated with *Generative Linguistics* (Chomsky; innate rule patterns)
 - Connections with *Cognitive Science*
 - study of how humans produce, process and understand language
 - Formerly, included study of formal languages and programming languages (now, CS, proper)

NLP

- Computational models of aspects of human language processing:
 - Understanding a textbook
 - Writing a letter
 - Carrying on a conversation
 - Translating a document
 - Searching for information



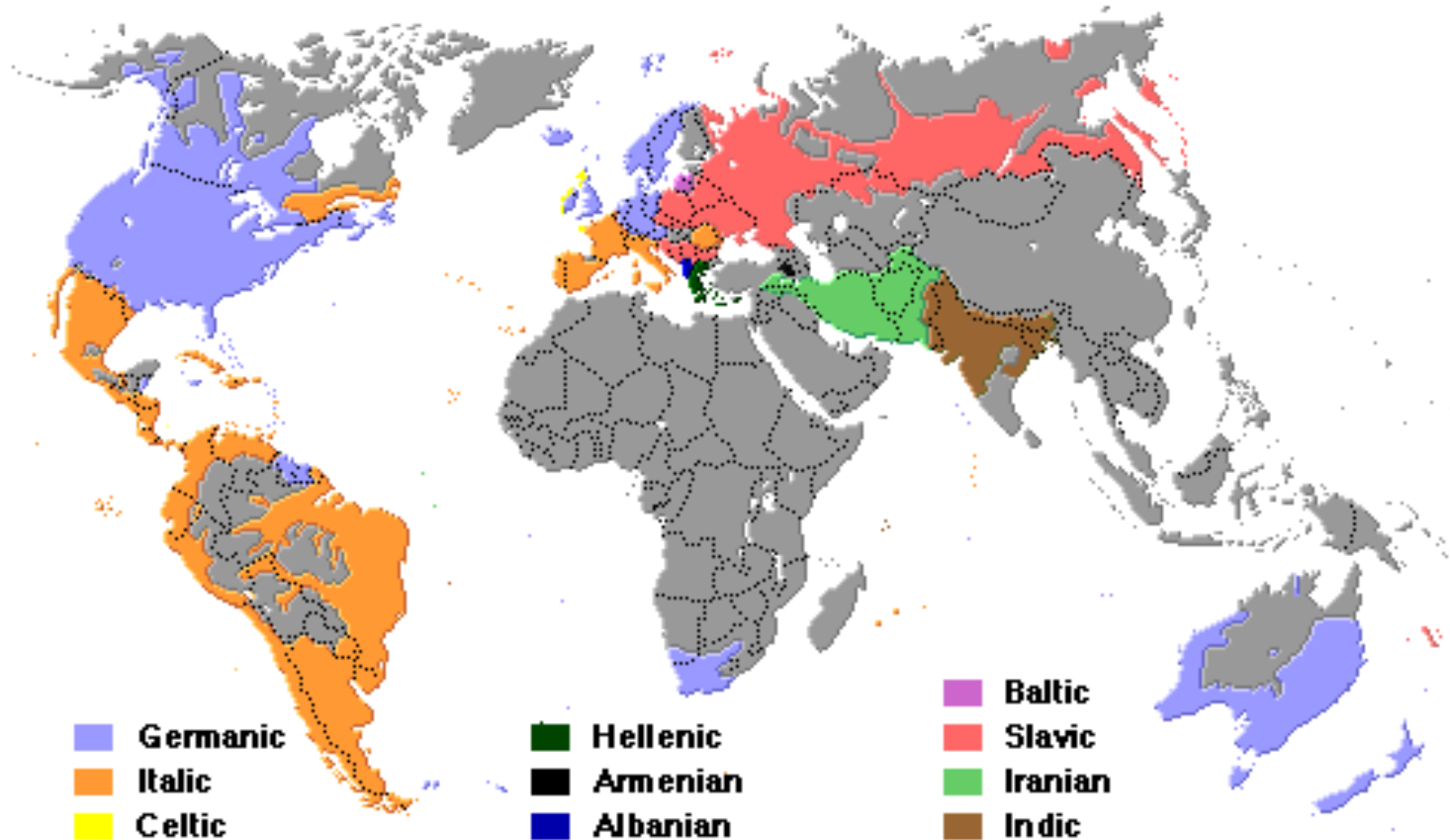
Human Languages

- 6809 known spoken human languages (debated)
- 4% of languages spoken by 97% of world's population
- 96% of languages spoken by 3% of world
- 10% on Internet
- 48 different *first* languages in Israel
- Mandarin Chinese has 845,000,000 speakers
- 543 extinct languages (at least)
- 417 on verge of extinction (2 speak Ter Sami)
- Half on way to extinction (spoken by <10,000)

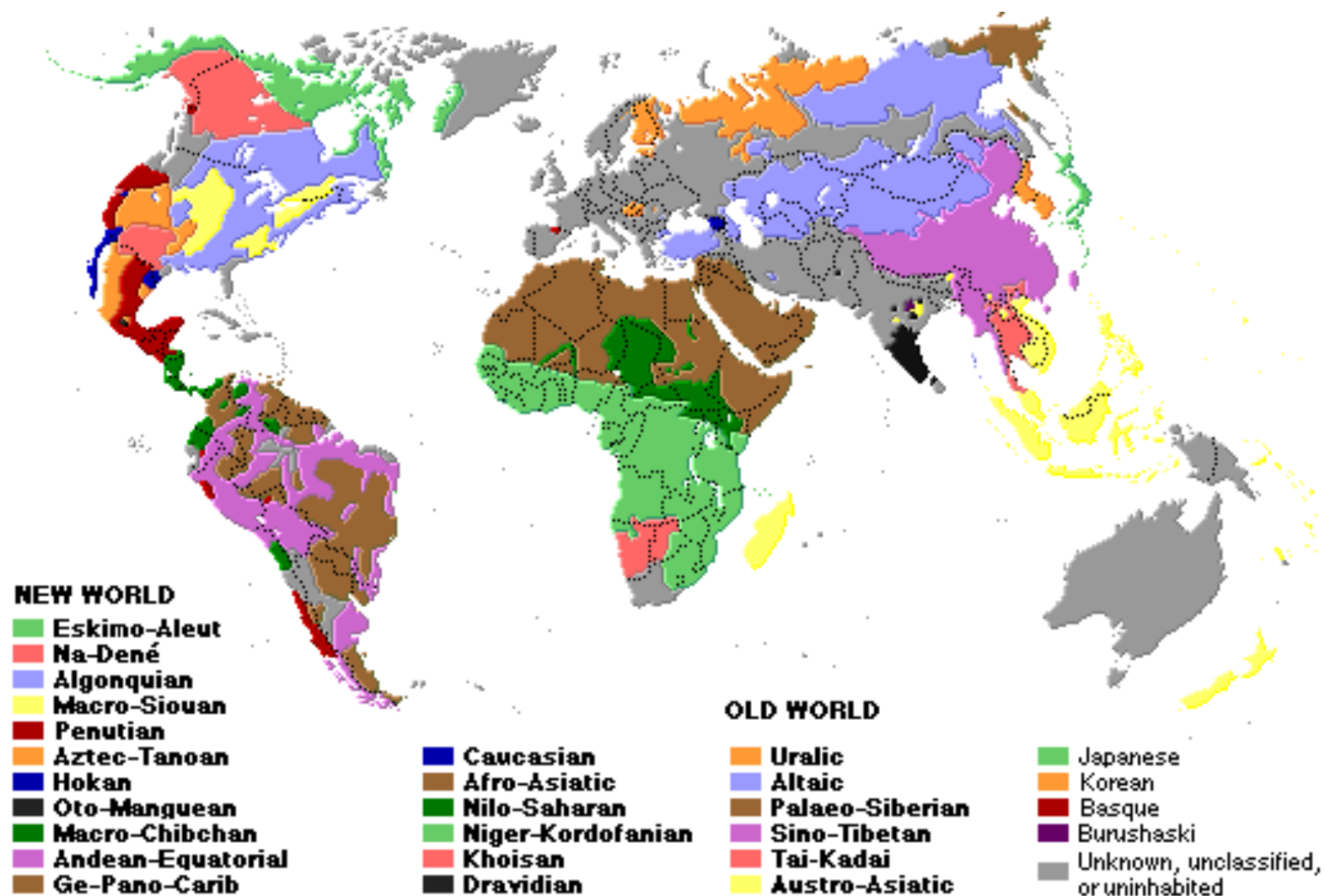
Table 2. Distribution of languages by number of first-language speakers

Population range	Living languages			Number of speakers		
	Count	Percent	Cumulative	Count	Percent	Cumulative
100,000,000 to 999,999,999	8	0.1	0.1%	2,308,548,848	38.73721	38.73721%
10,000,000 to 99,999,999	77	1.1	1.2%	2,348,900,757	39.38078	78.11797%
1,000,000 to 9,999,999	304	4.4	5.6%	951,916,458	15.97306	94.09103%
100,000 to 999,999	896	13.0	18.6%	283,116,716	4.75067	98.84170%
10,000 to 99,999	1,824	26.4	45.0%	60,780,797	1.01990	99.86160%
1,000 to 9,999	2,014	29.2	74.1%	7,773,810	0.13044	99.99204%
100 to 999	1,038	15.0	89.2%	461,250	0.00774	99.99978%
10 to 99	339	4.9	94.1%	12,560	0.00021	99.99999%
1 to 9	133	1.9	96.0%	521	0.00001	100.00000%
Unknown	277	4.0	100.0%			
Totals	5,909	100.0		5,959,511,717	100.00000	

Indo-European



Rest



Syntax

SVO common: *English, French, Danish, Chinese, Swahili (Tanzania)*

SOV common: *German, Turkish, Japanese, Persian, Korean*

VSO rare: *Gaelic (Ireland), Arabic, Welsh (UK)*

VOS rare: *Mopán Maya (Belize), Bushi (Madagascar), Fijian (Fiji)*

OSV extremely rare: *Xavante (Brazil)*

OVS extremely rare: *Panare (Venezuela), Macushi (Guyana), [Klingon]*

Morphology

- ***Isolating / Analytic***: Most words can't be changed; almost no inflection. Often have rich particle systems instead, i.e. a lot of small separate words for marking case, tense, topic, etc. *Examples: Chinese, Vietnamese*
- ***Inflected / Fusional / Synthetic***: Words with some affixes. An affix can have more than one grammatical function or meaning. *Examples: Latin, Greek, Arabic*
- ***Agglutinative***: Rich, but strict, inflection system. Every affix has a fixed grammatical function or meaning. *Examples: Finnish, Turkish, Japanese*
- ***Amalgamating / Polysynthetic***: Vast number of morphemes to combine to very complex words. One word might express what in other languages would be expressed by a sentence. *Examples: Inuktit, Mohawk*

Hebrew

- 5000 roots
- 50,000 dictionary entries
- 2,000,000 prefix-free words
- 10,000,000 words
- $\aleph_0$ sentence
- 10,000,000 speakers

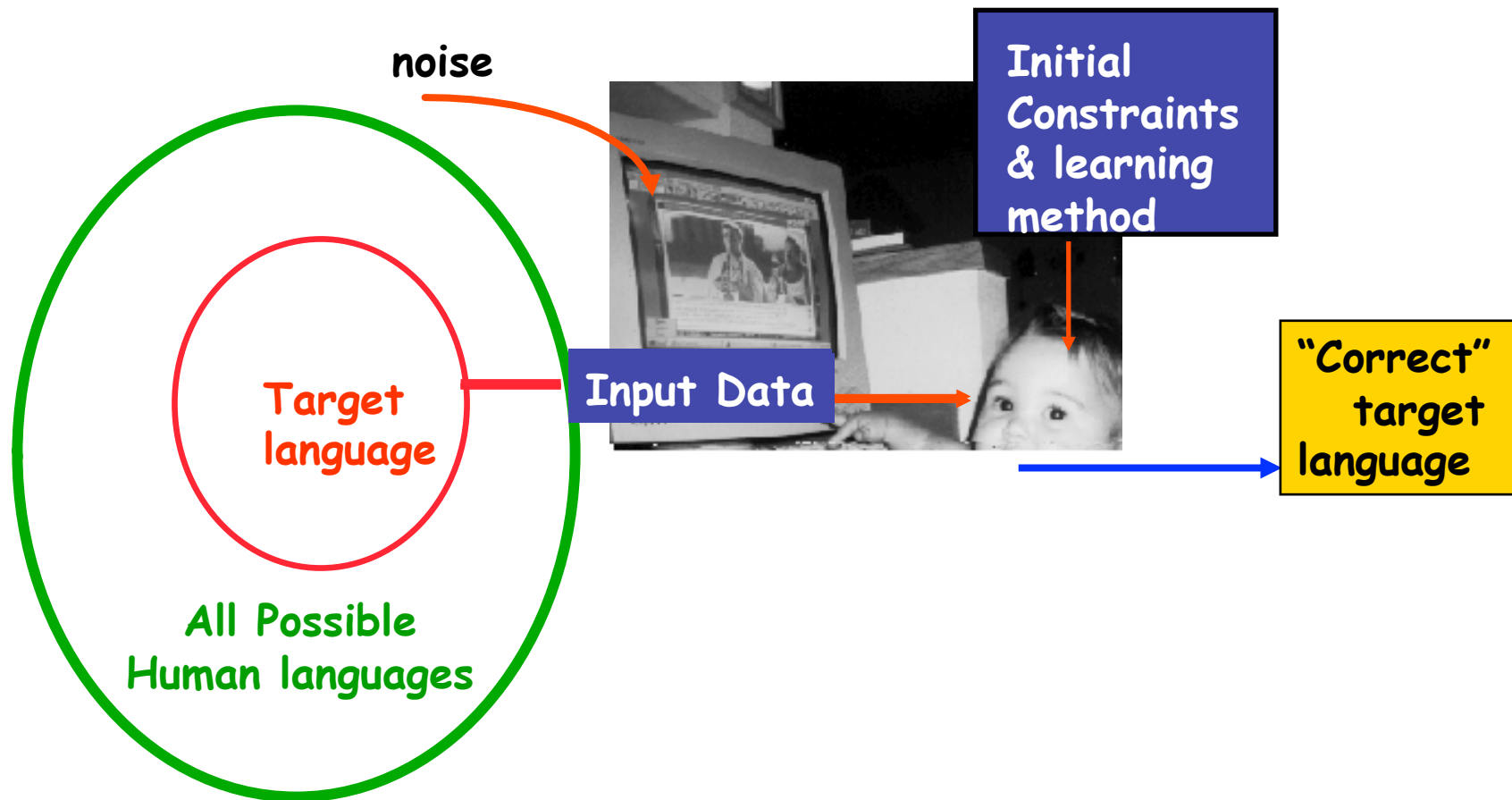
Human Languages

- You know ~50,000 words of primary language, each with several meanings
- Six year old knows ~13000 words
- First 16 years we learn 1 word every 90 minutes of waking time
- Mental grammar generates sentences
 - virtually every sentence is novel
- 3 year olds already have 90% of grammar

The Logical Problem of Language Acquisition



"problem of discovering the character of the mental capacities that make it possible for human beings to attain knowledge of their language on the basis of fragmentary and haphazard early linguistic experience"



Human Spoken Language

- Most complicated mechanical motion of the human body
 - Movements must be accurate to within mm
 - synchronized within 100ths of a second
- We can understand up to 50 phonemes/sec (normal speech 10-15ph/sec)
 - but if sound is repeated 20 times /sec we hear continuous buzz!
- All aspects of language processing are involved and manage to keep apace



This model shows what a man's body would look like if each part grew in proportion to the area of the cortex of the brain concerned with its movement.

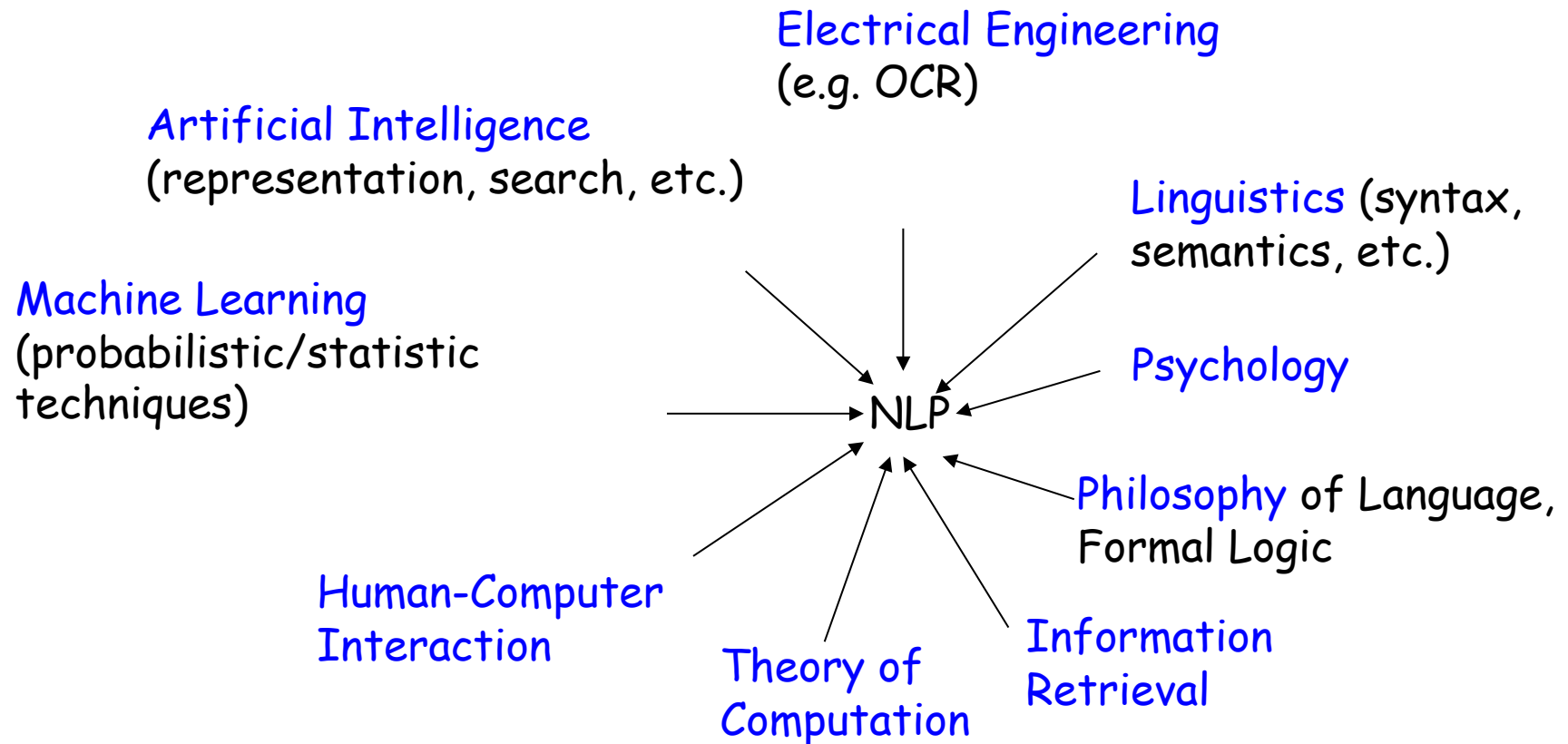
Natural Language Processing

- speech recognition
- natural language understanding
- computational linguistics
- psycholinguistics
- information extraction
- information retrieval
- inference
- natural language generation
- speech synthesis
- language evolution

Applied NLP

- Machine translation
- Spelling/grammar correction
- Information retrieval
- Data mining
- Document classification
- Question answering, conversational agents

Related Disciplines



Major Language Disciplines

Discipline	Typical Problem
Linguists	How do words form phrases and sentences?
Psycholinguists	How do people identify sentence structure?
Philosophers	How do words and sentences acquire meaning?
Computational Linguists	How is the structure of sentences identified?

Other Disciplines

- ★ **Linguistics**: formal grammars, abstract characterization of what is to be learned
- ★ **Computer Science**: algorithms for efficient learning or online deployment of such systems
- ★ **Engineering**: stochastic techniques for characterizing regular patterns for learning and ambiguity resolution
- ★ **Psychology**: insights into what linguistic constructions are easy or difficult for people to learn or use

What's involved in an "intelligent" Answer?

Analysis:

Decomposition of the signal (spoken or written) eventually into meaningful units.

This involves ...

Speech/Character Recognition

- Decomposition into words, segmentation of words into appropriate phonemes or letters
- Requires knowledge of phonological patterns:
 - I'm enormously proud.
 - I mean to make you proud.

Morphological Analysis

- Inflectional
 - duck + s = [N duck] + [plural s]
 - duck + s = [V duck] + [3rd person s]
- Derivational
 - kind, kindness
- Spelling changes
 - drop, dropping
 - hide, hiding

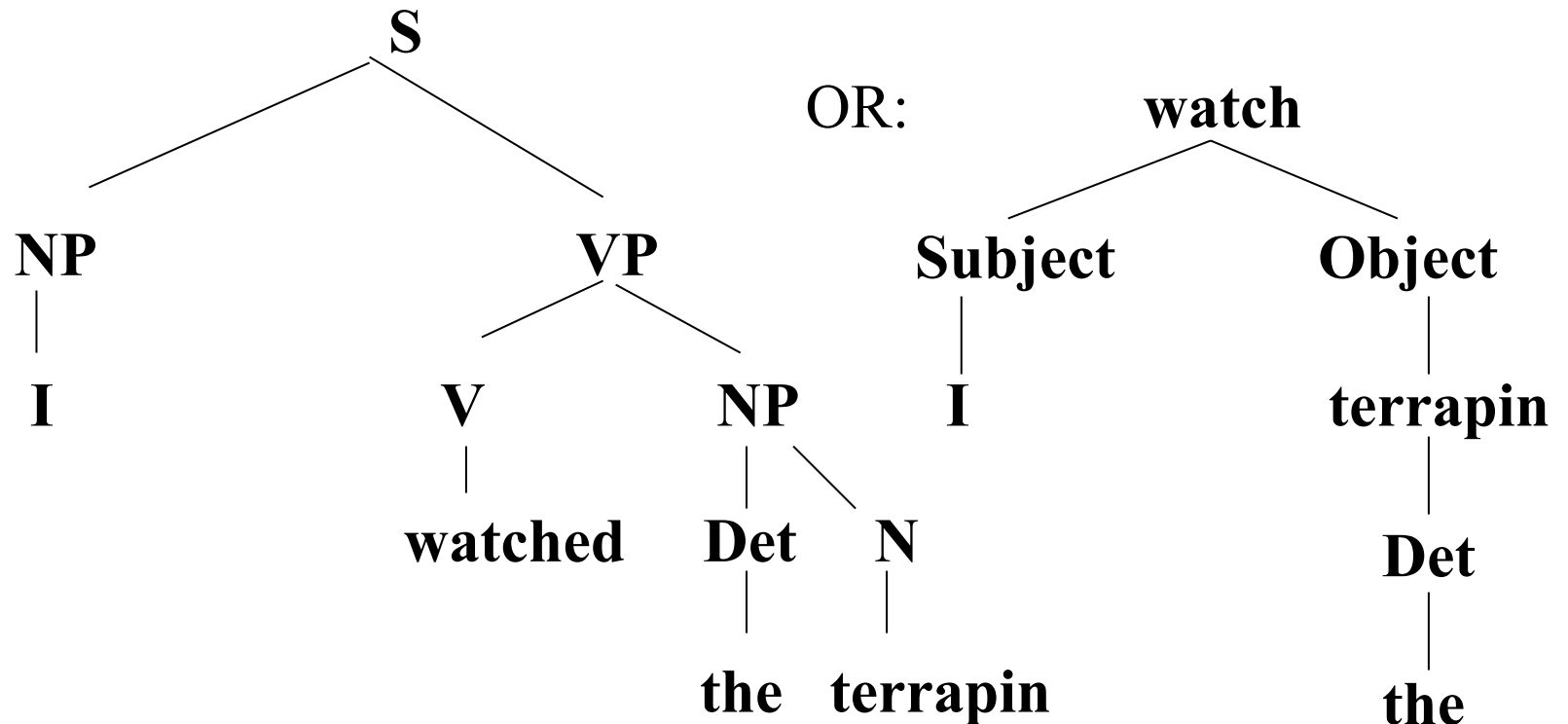
I watched the terrapin.



Stu Weiss

Syntactic Analysis

- Associate constituent structure with string
- Prepare for semantic interpretation



Semantic Interpretation

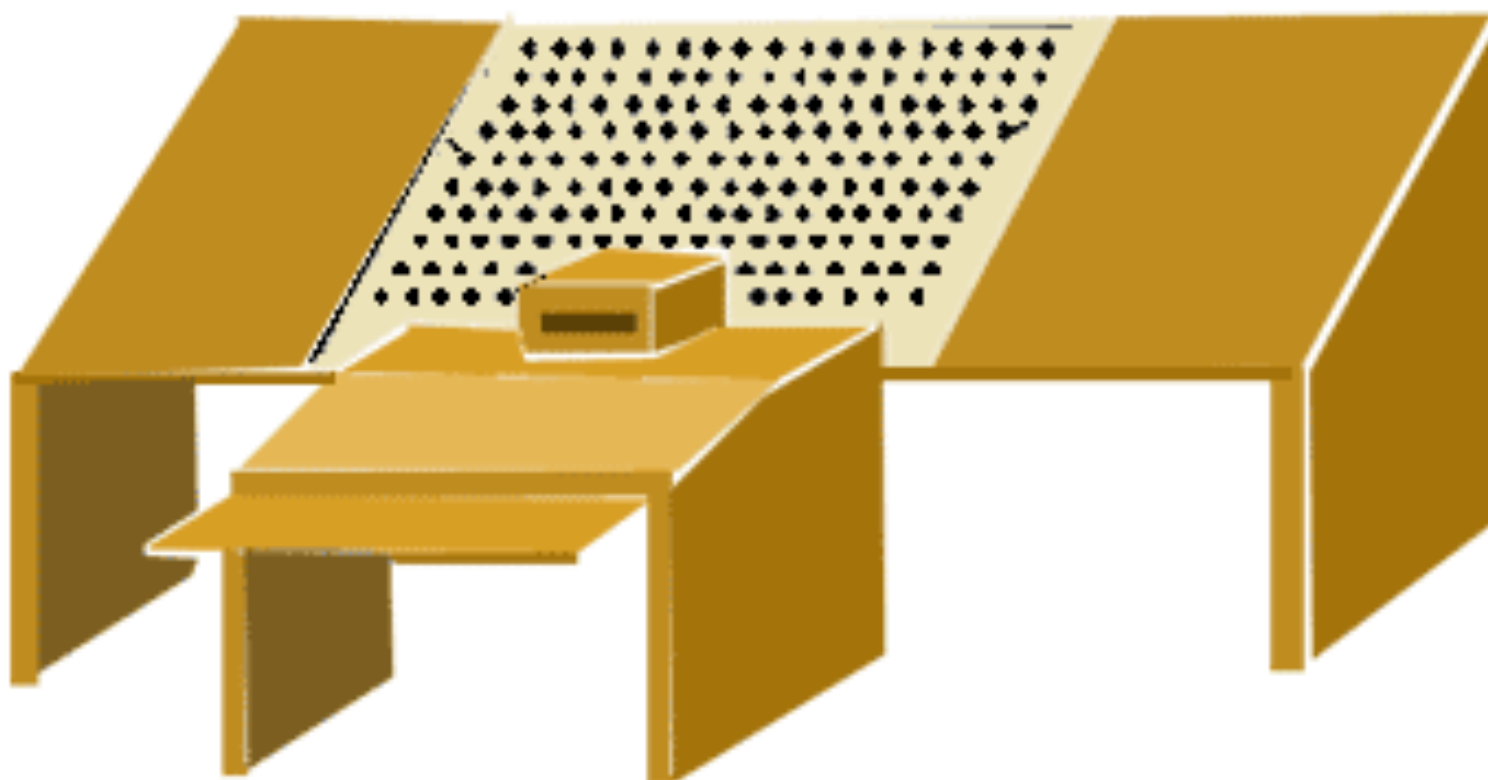
- A way of representing meaning
- Abstracts away from syntactic structure
- Example:
 - First-Order Logic: `watch(I, terrapin)`
 - Can be: "I watched the terrapin" or "The terrapin was watched by me"
- Real language is complex:
 - What did I watch?

History

Descartes (1629)

If a large dictionary is printed in all the languages in which one wants to be understood, where common characters are put for each primitive word corresponding to the sense and not to the syllables, such as the same character for *aymer*, *amare* and *filein*, those having the dictionary and knowing the grammar would be able, through seeking all these characters one after another, to understand in their language what is being written.

Petr Troyanskii (1933)



multilingual dictionary, stemmer,...

Background

- Development of formal language theory (Chomsky, Kleene, Backus)
 - Formal characterization of classes of grammar (context-free, regular)
 - Association with relevant automata
- Probability theory: language understanding as decoding through noisy channel (Shannon)
 - Use of information theoretic concepts like entropy to measure success of language models

Weaver (1947)

Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

History (1948)

- First recognizable NLP application was a **dictionary look-up system** developed at Birkbeck College, London

Turing Test (1950)

- Turing test: machine, human, and human judge
- Judge asks questions of computer and human
 - Machine's job is to act like a human; human's job is to convince judge that he's not the machine
 - Machine judged "intelligent" if it can fool judge
- Original: Male and the female try to convince the judge that they were female. If AI replaced male, would judge be more accurate in guessing who real female was?



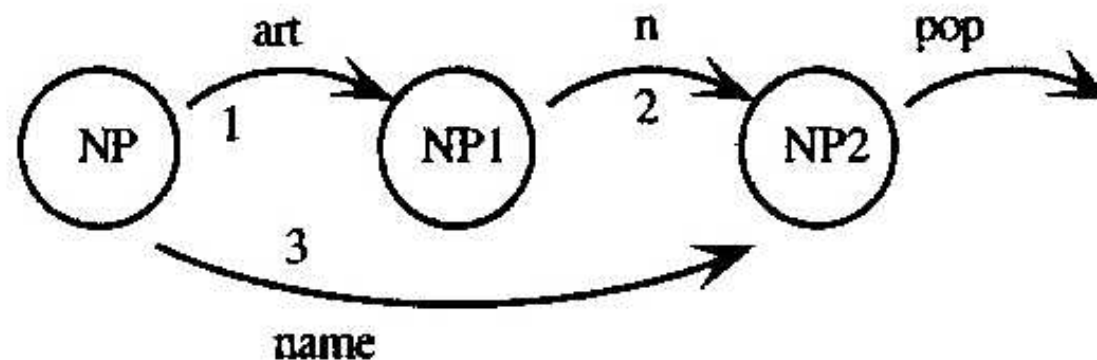
History (1940-1960)

- Symbolic approach
 - Generative syntax
 - AI
 - pattern matching
 - logic-based systems
 - special-purpose systems
- Stochastic approach
 - Bayesian methods

Symbolic vs. Stochastic

- Symbolic
 - Formal grammars as basis for NLP and learning systems. (Chomsky, Harris)
 - Logic and logic-based programming for characterizing syntactic/semantic inference (Kaplan, Kay, Pereira)
 - Toy natural-language understanding/generation systems (Woods, Minsky, Schank, Winograd, Colmerauer)
 - Discourse processing: Role of intention, focus (Grosz, Sidner, Hobbs)
- Stochastic Modeling
 - Probabilistic methods for early speech recognition, OCR (Bledsoe & Browning, Jelinek, Black, Mercer)

Augmented Transition Networks



Arc	Test	Actions
1	none	$DET := *$ $AGR := AGR \cup *$
2	$AGR \cap AGR \neq \emptyset$	$HEAD := *$ $AGR := AGR \cap AGR \neq \emptyset$
3	none	$NAME := *$ $AGR := AGR \cup *$

Early Excitement (1960s)

By 1966 US government had spent \$20M on machine translation alone

Critics:

- Bar Hillel - "no way to disambiguation without deep understanding"
- Pierce report (1966): "no way to justify work in terms of practical output"

ELIZA (1966)

- Remarkably simple "Rogerian Psychologist"
- Uses pattern patching to carry on limited form of conversation
- Seemed to pass the Turing Test! (*Machines Who Think*, McCorduck, 1979)
- Attacked (Weizenbaum)
- Demos:
 - <http://www.manifestation.com/neurotoys/eliza.php3>
 - http://www.lpa.co.uk/pws_dem4.htm

History (1970-1990)

- Stochastic
 - speech recognition and synthesis (Bell Labs)
- Logic-based
 - compositional semantics (Montague)
 - definite-clause grammars (Pereira & Warren)
- Ad-hoc
 - SHRDLU robot in blocks world (Winograd)
 - knowledge-representation systems (Shank)

SHRDLU (Winograd, 1970)

Answers questions, executes commands, and accepts information in an interactive English dialog... The system contains a parser, a recognition grammar of English, programs for semantic analysis, and a general problem solving system... It can remember and discuss its plans and actions as well as carrying them out... Knowledge in the system is represented in the form of procedures, rather than tables of rules or lists of patterns.

Dave McDonald

SHRDLU was a special program. Even today its parser would be competitive as an architecture. For a recursive descent algorithm it had some clever means of jumping to anticipated alternative analyses.... It defined the whole notion of **procedural semantics** (though **Bill Woods** tends to get the credit), and its grammar was the first instance of Systemic Functional Linguistics applied to language understanding and quite well done.

LIFER/LADDER (1978)

- Interface to db about Navy ships
- Semantic grammar
- User-friendly features (e.g. shortcuts)
- People adapted to system

History (1970-1990)

- Discourse modeling
 - anaphora
 - *If you need one, there's a towel in the top drawer.*
 - focus/topic (Grosz et al.)
 - conversational implicature (Grice)

History (1970-1990)

- Semantic representations
 - **Conceptual dependency** - method of expressing language in terms of semantic primitives (Schank et al.)
 - **Semantic network** (Quillian)
 - **Procedural semantics** - intermediate representation between NLP system and DB system (Woods)

Empiricism (1983-93)

- Use of stochastic techniques for part of speech tagging, parsing, word sense disambiguation, etc.
- Comparison of stochastic, symbolic, more or less powerful models for language understanding and learning tasks

Renaissance (1990-)

- Lessons from phonology & morphology successes:
 - finite-state models are very powerful
 - probabilistic models pervasive
 - Web creates new opportunities and challenges
 - practical applications driving the field again
- Multilinguality and Multimodality

2000s

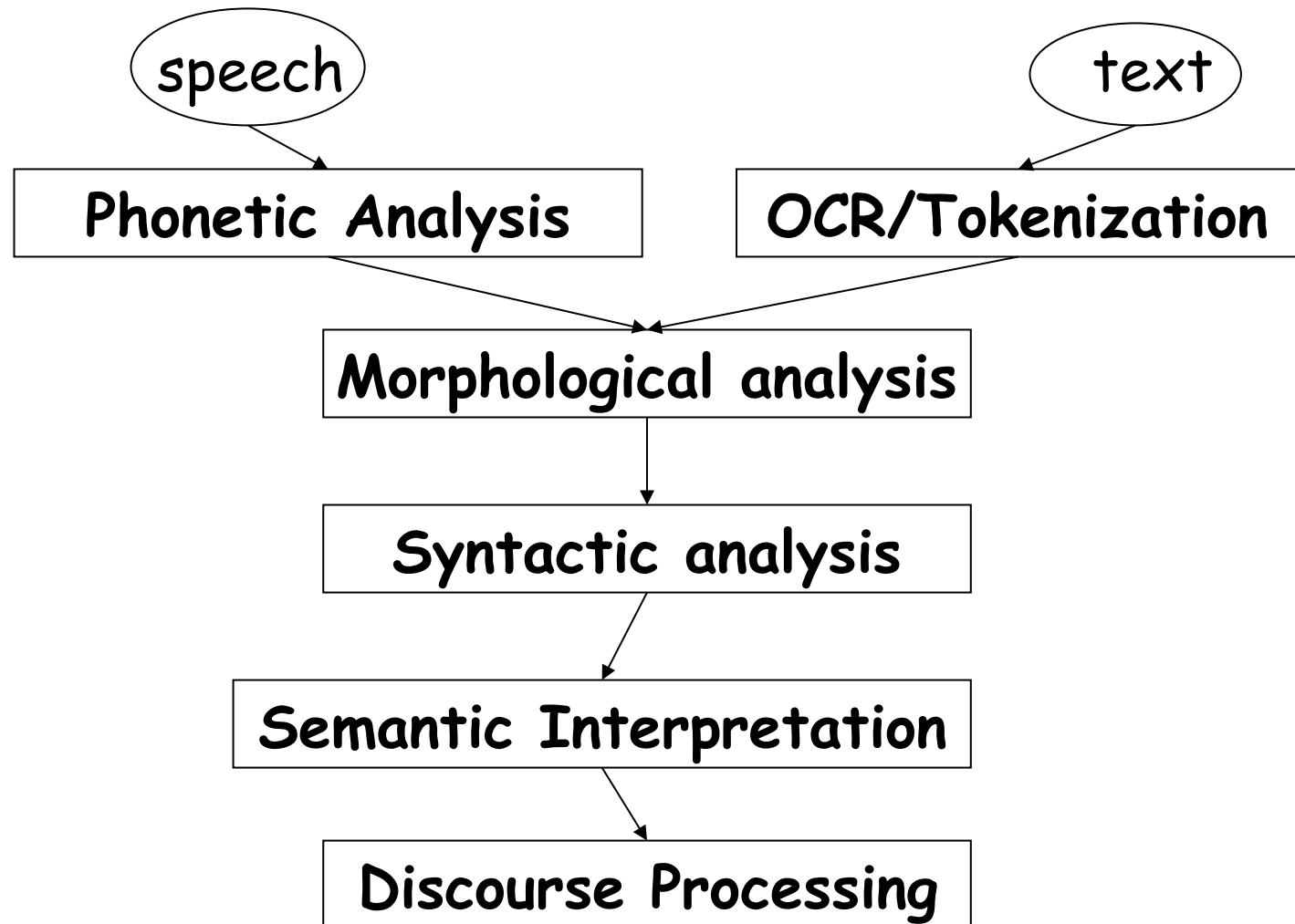
- MT: Speech-speech translation comes of age
- Corpus-based paradigm booms:
 - Example-based and statistical MT
 - Corpus-based statistical parsers (treebank trained)
 - Corpus-based fact extractors
- Speech recognition: now a commodity
- Linguistics: corpus-based revolution (e.g. Bresnan)
- IR boom: Google, translingual IR, summarization,...
- Return to basic unsolved problems contemplated:
 - Dialog: beyond "planners"
 - Definite reference resolution
 - Metaphor, Metonymy,...

Present

- Advances in software and hardware create NLP needs for
 - information retrieval (web)
 - machine translation
 - spelling and grammar checking
 - speech recognition and synthesis
- Stochastic and symbolic methods combine for real-world applications

Issues

NLP Pipeline



NL Understanding

Query:

"Find me all articles concerning car accidents involving more than two cars in Malta during the first half of last year."

System must extract enough information to determine whether the article meets the criterion defined by the query

NLU

- Compute some representation of the information that can be used for later inference
- How much understanding is necessary to achieve the purpose of a given system?

Dialogue-based Applications

Human-machine communication

- NL database query systems
- Automated customer services
 - banking services
 - reservation systems
 - etc.

Multimodal Applications

Involve two or more modalities of communication

- Text
- Speech
- Gesture
- Image

Text → speech

Speech → text

- Multimodal document generation
- Spoken translation systems
- Spoken dialogue systems

Ambiguity

I made her duck

I made duckling for her

I made the duckling belonging to her

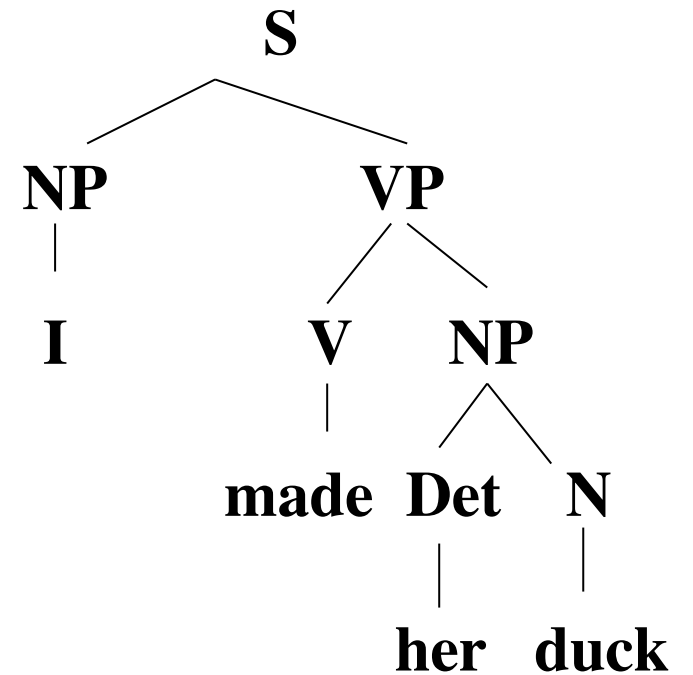
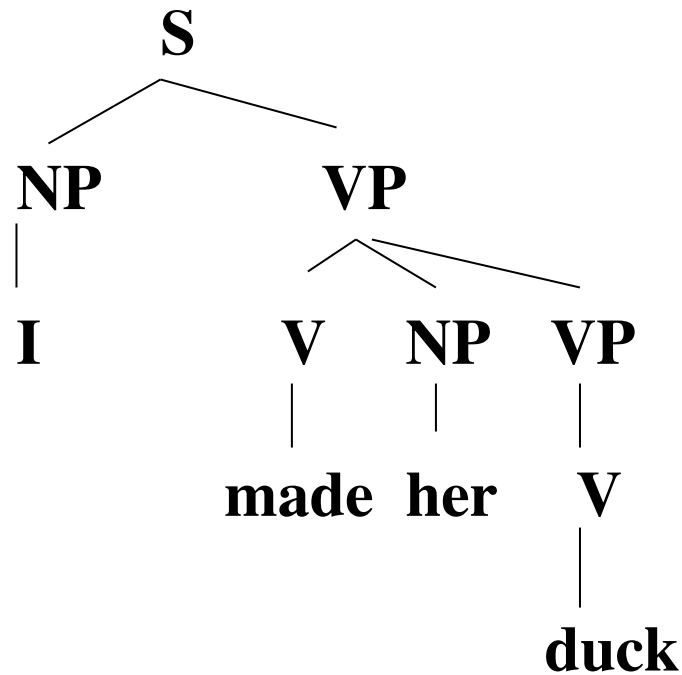
I created the duck she owns

I forced her to lower her head

By magic, I changed her into a duck

Syntactic Disambiguation

- Structural ambiguity:



Part of Speech (POS) Tagging and Word Sense Disambiguation (WSD)

[verb Duck] !

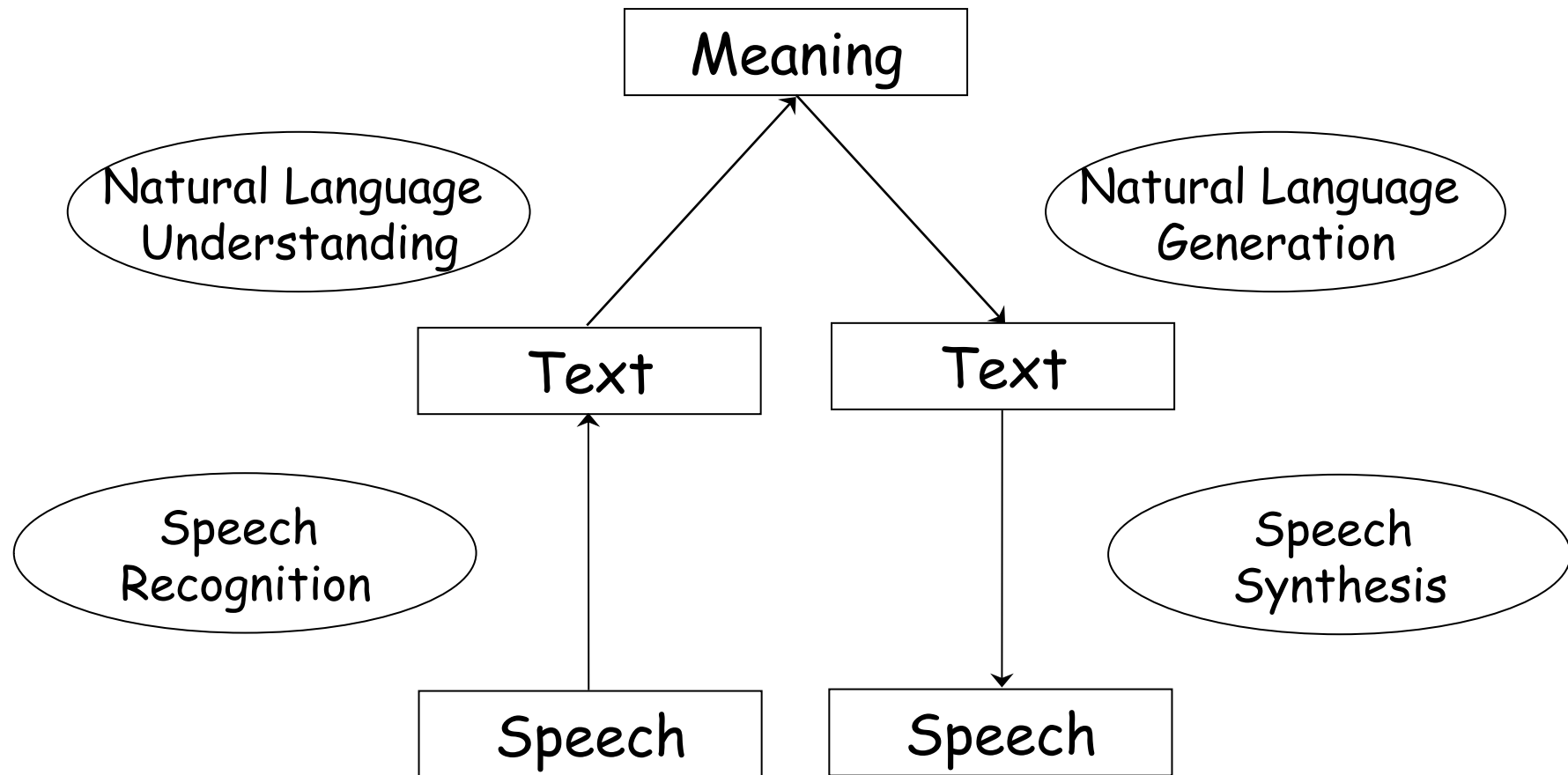
[noun Duck] is delicious for dinner

- I went to the **bank** to deposit my check.
- I went to the **bank** to look out at the river.
- I went to the **bank** of windows and chose the one dealing with last names beginning with "d".

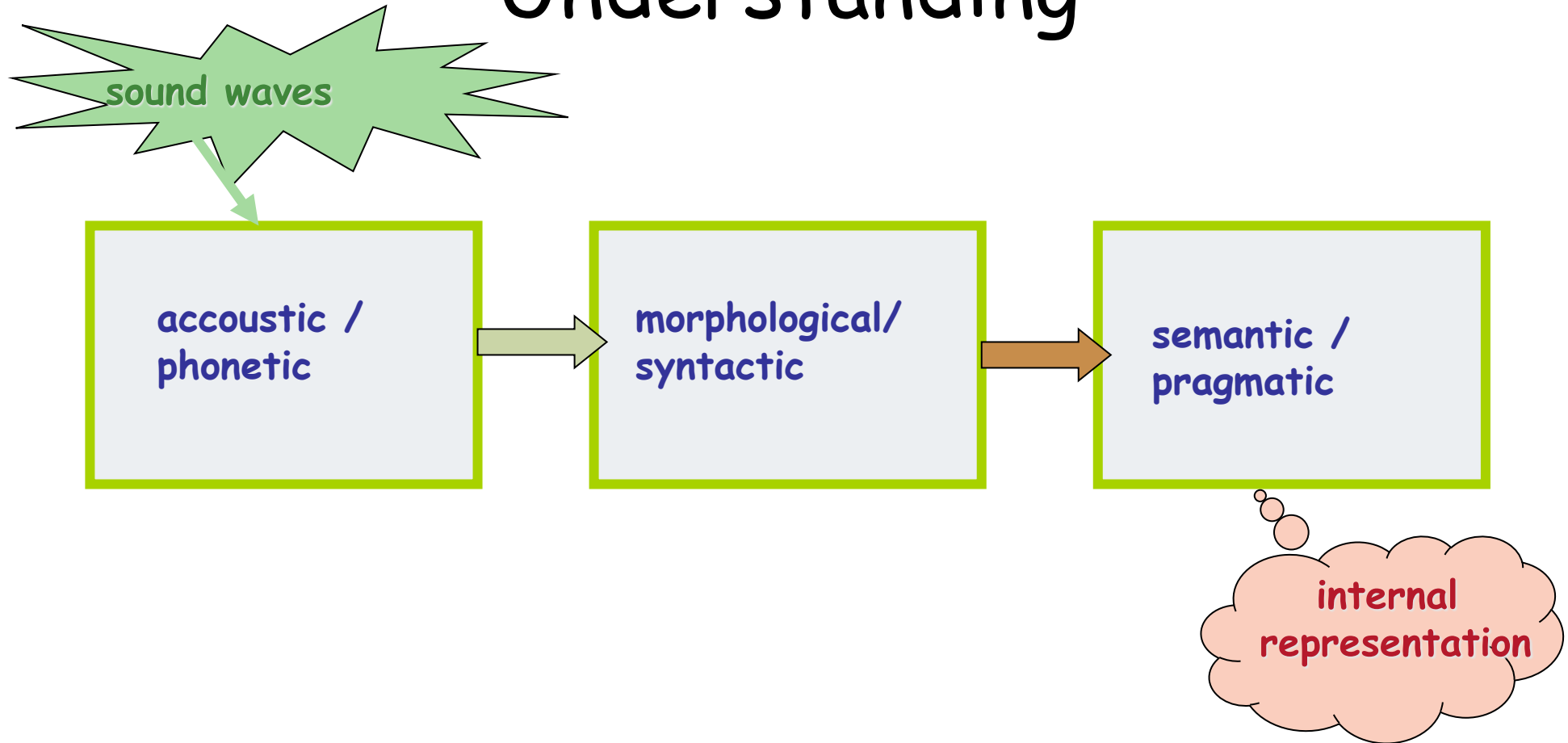
Resources

- Dictionary
- Morphology and Spelling Rules
- Grammar Rules
- Semantic Interpretation Rules
- Discourse Interpretation

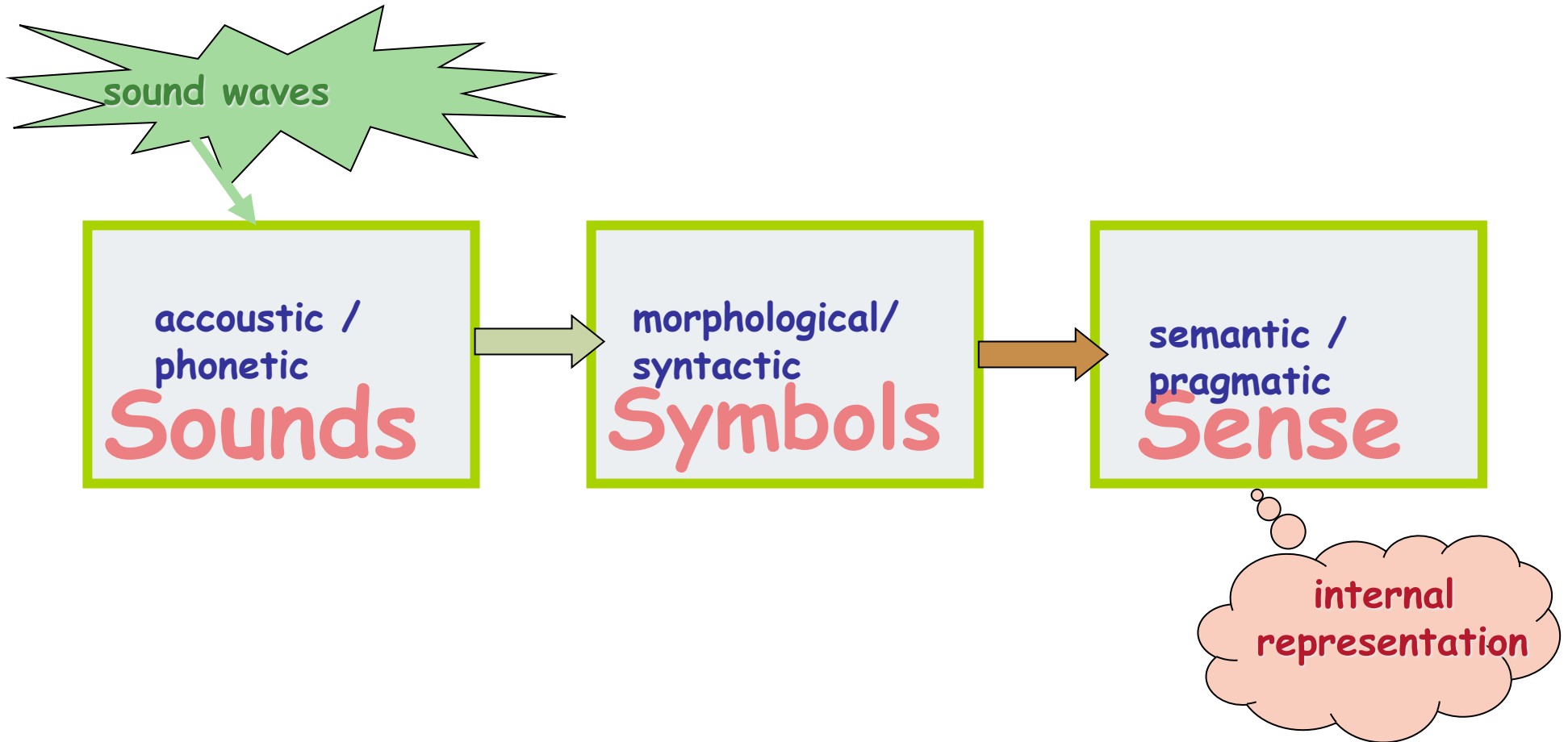
Language Technology



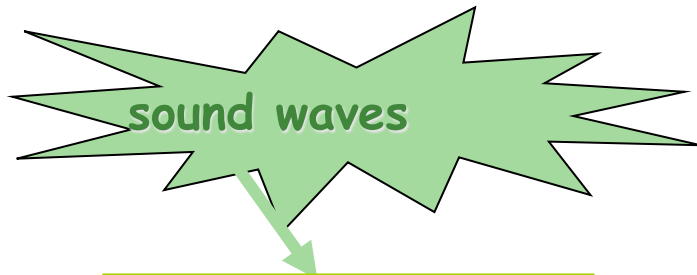
Natural Language Understanding



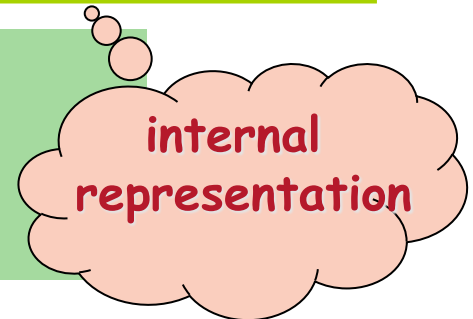
NLU



Where are the words?



- "How to recognize speech, not to wreck a nice beach"
- "The cat scares all the birds away"
- "The cat's cares are few"



- pauses in speech bear little relation to word breaks
- + intonation offers additional clues to meaning

Dissecting words/sentences

sound waves

accoustic /
phonetic

morphological/
syntactic

semantic /
pragmatic

- “I saw the Golden bridge flying into San Francisco”

- Word creation:

establish

establishment

the church of England as the official state church.

disestablishment

antidisestablishment

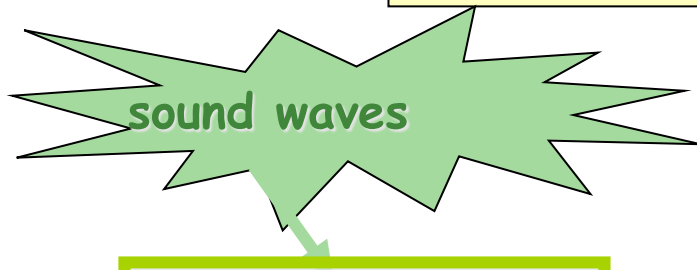
antidisestablishmentarian

antidisestablishmentarianism

is a political philosophy that is opposed to the separation of church and state.

internal
representation

What does it mean?



accoustic /
phonetic

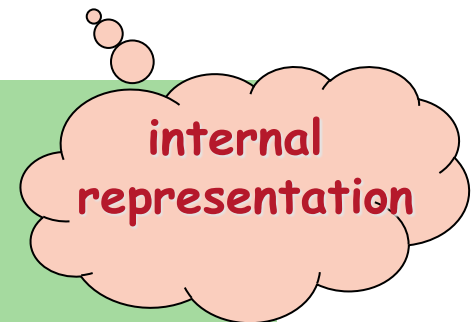


morphological/
syntactic

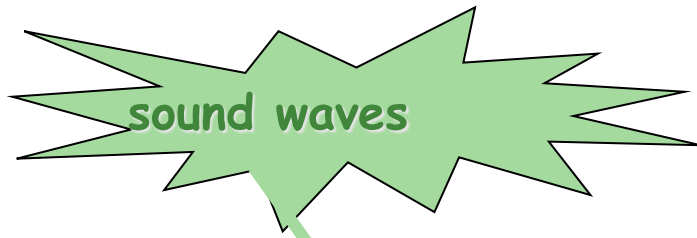


semantic /
pragmatic

- "I saw Pathfinder on Mars with a telescope"
- "Pathfinder photographed Mars"
- "The Pathfinder photograph from Ford has arrived"
- "When a Pathfinder fords a river it sometimes mars its paint job."



What does it mean?



acoustic /
phonetic

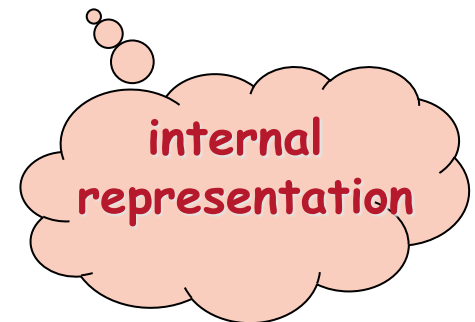


morphological/
syntactic



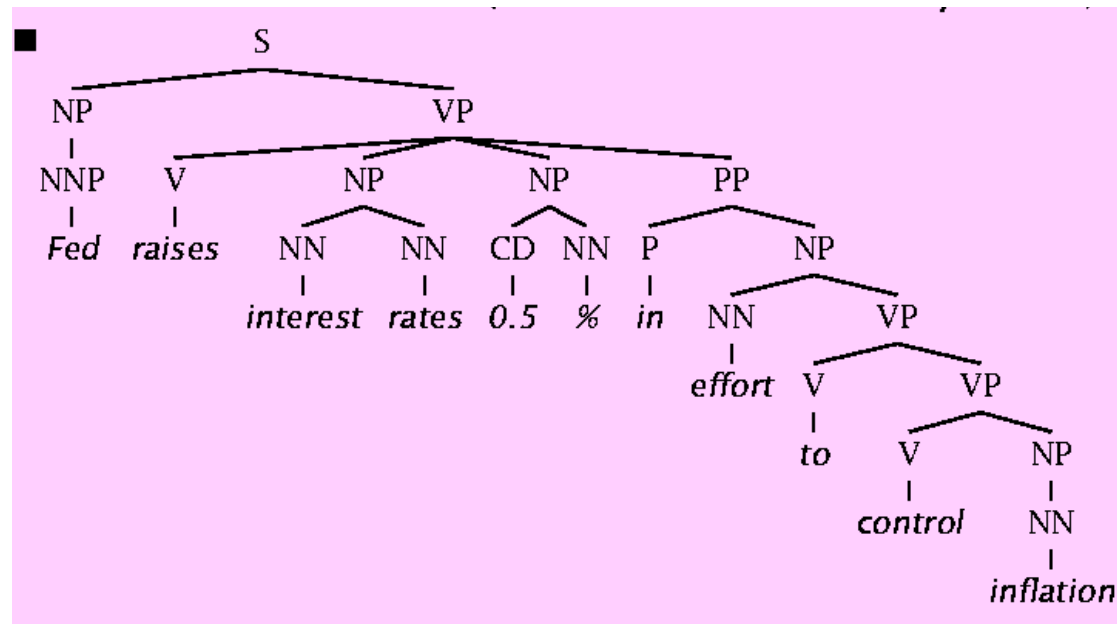
semantic /
pragmatic

- "Jack went to the store. **He** found the milk in aisle 3. **He** paid for **it** and left."
- "Surcharge for white orders."
- " Q: Did you read the report?
A: I read Bob's email."



Syntax

- The hidden structure of language is highly ambiguous
- Structures for: *Fed raises interest rates 0.5% in effort to control inflation* (NYT headline 5/17/00)



Where are the ambiguities?

Part of speech ambiguities

		VB				
	VBZ	VBP	VBZ			
NNP	NNS	NN	NNS	CD	NN	
<i>Fed</i>	<i>raises</i>	<i>interest</i>	<i>rates</i>	<i>0.5</i>	<i>%</i>	

Syntactic attachment ambiguities

*in effort
to control
inflation*

*Word sense ambiguities: Fed → “federal agent”
interest → a feeling of wanting to know or learn more*

POS Tags

- NN noun s
- NNP proper noun
- NNS noun plural
- VB v base form
- VBZ v 3p s present
- VBP v non-3p s present
- CD number

Lexical Semantics

The Terrapin, is whom I watched.
Watch the Terrapin is what I do best.

I = experiencer
Watch = predicate
Terrapin = patient

Logical Semantics

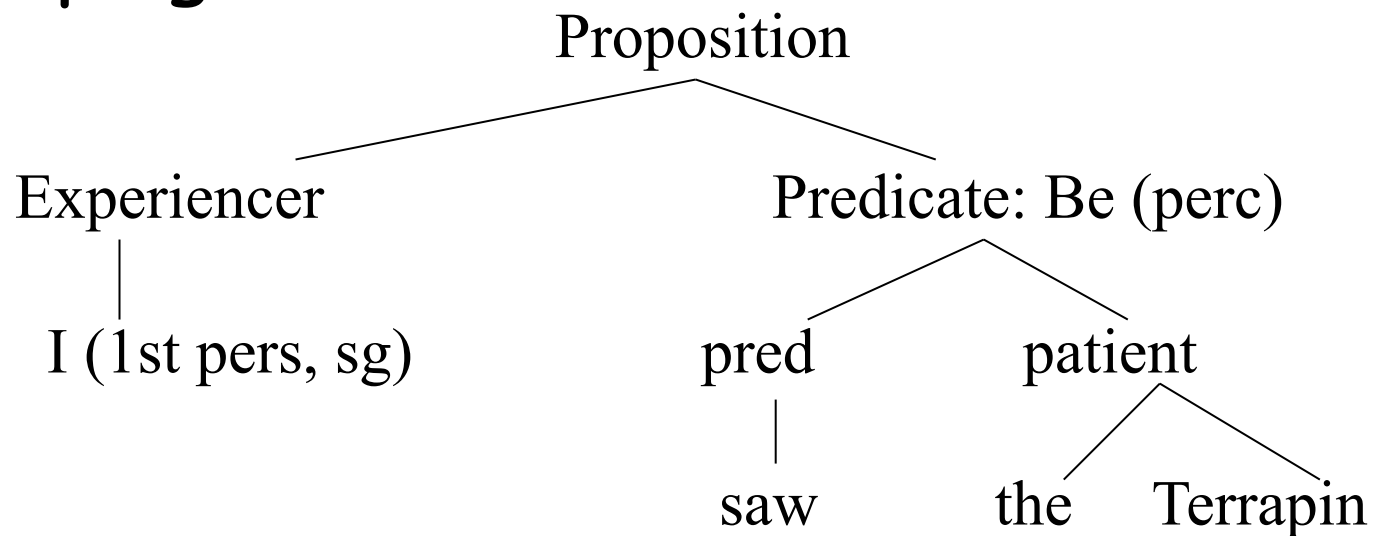
watch(I, terrapin)

The Terrapin was watched by me.

I watched the Terrapin.

Compositional Semantics

- Association of parts of a proposition with semantic roles
- Scoping



Word-Governed Semantics

- Any verb can add "able" to form an adjective.
 - I taught the class . The class is teachable
 - I rejected the idea. The idea is rejectable.
- Association of particular words with specific semantic forms.
 - John (masculine)
 - The boys (masculine, plural, human)

Pragmatics

- Real world knowledge, speaker intention, goal of utterance.
- Related to sociology.
- Example 1:
 - Could you turn in your assignments now (command)
 - Could you finish the homework? (question, command)
- Example 2:
 - I couldn't decide how to catch the crook. Then I decided to spy on the crook with binoculars.
 - To my surprise, I found out he had them too. Then I knew to just follow the crook with binoculars.
 - [the crook [with binoculars]]
 - [the crook] [with binoculars]

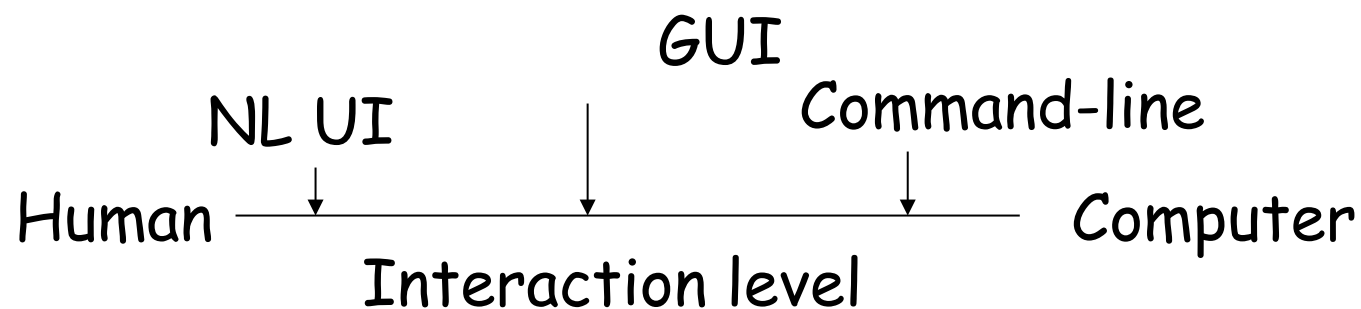
Discourse Analysis

- Discourse
 - Pronoun reference (anaphora)
The professor told the student to finish the assignment. He was pretty aggravated at how long it was taking to pass it in.
 - Multiple reference to same entity (co-reference)
George W. Bush, president of the U.S.
 - Relation between sentences:
John hit the man. He had stolen his bicycle.

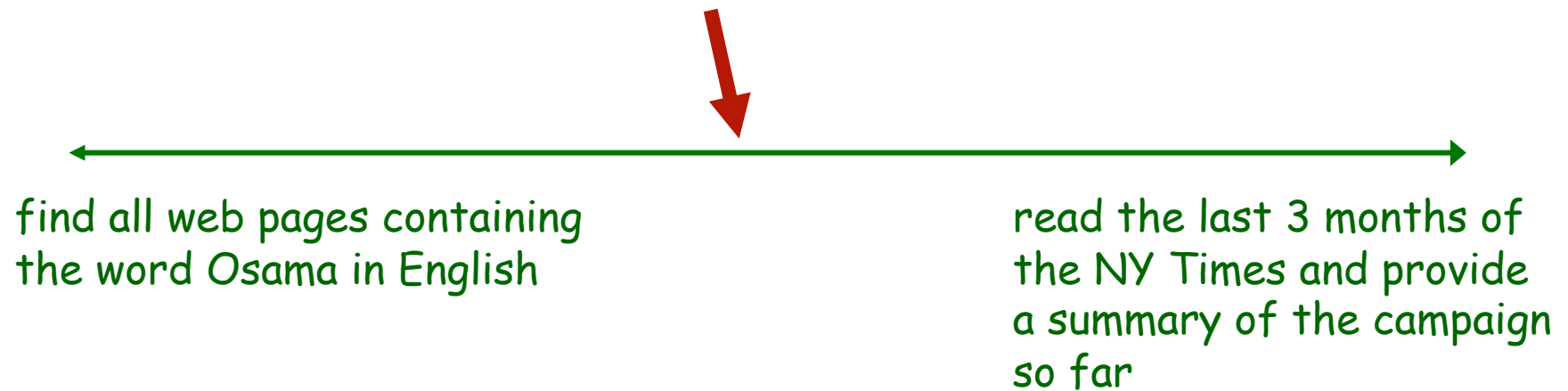
Applications

Interaction Level

- NL used to make interaction level more "human"



Spectrum



Some NLP Applications

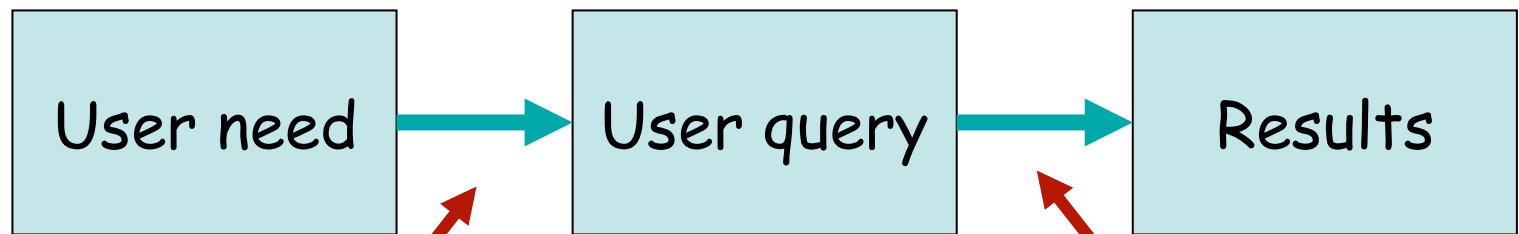
- Machine Translation—Babelfish (Alta Vista):
<http://babelfish.altavista.com/translate.dyn>
- Question Answering—Ask Jeeves (Ask Jeeves):
<http://www.ask.com/>
- Language Summarization—MEAD (U. Michigan):
<http://www.summarization.com/mead>
- Spoken Language Recognition—EduSpeak (SRI):
<http://www.eduspeak.com/>
- Automatic Essay evaluation—E-Rater (ETS):
<http://www.ets.org/research/erater.html>
- Information Retrieval and Extraction—NetOwl (SRA):
http://www.netowl.com/extractor_summary.html

Text-based Applications

Processing of written texts such as books, news, papers, reports:

- Finding appropriate documents on certain topics from a text database
- Extracting information from messages, articles, Web pages, etc.

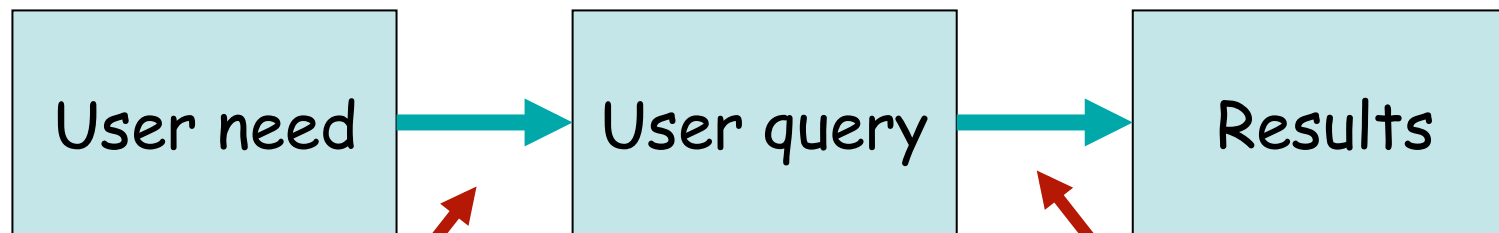
Translating user needs



For RDB, a lot of people know how to do this correctly, using SQL or a GUI tool

The answers coming out here will then be precisely what the user wanted

Translating user needs

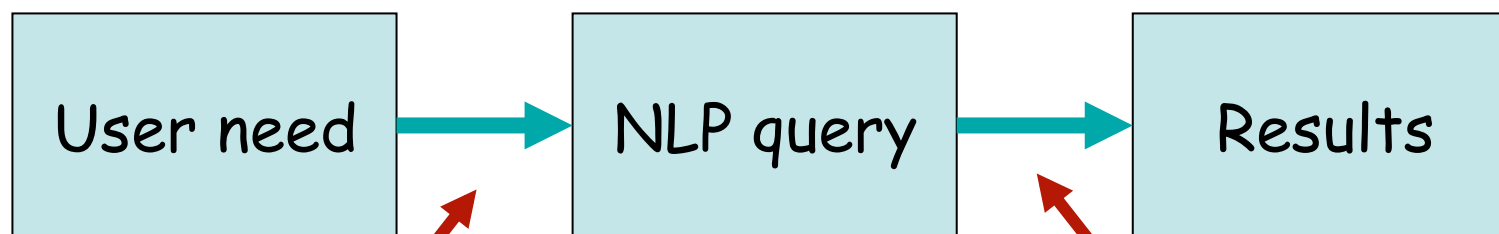


For meanings in text,
no IR-style query
gives one exactly
what one wants;
it only hints at it

The answers
coming out may
be roughly what
was wanted, or
can be refined

Sometimes!

Translating user needs



For a deeper NLP analysis system, the system subtly translates the user's language

If the answers coming back aren't what was wanted, the user frequently has *no idea* how to fix the problem

Risky!

Practical goals

Use language technology to add value to data by:

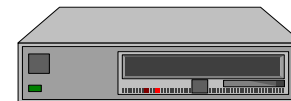
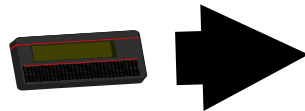
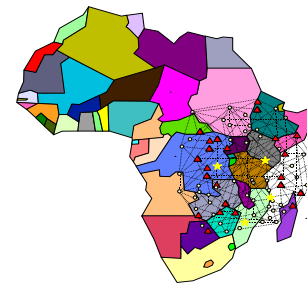
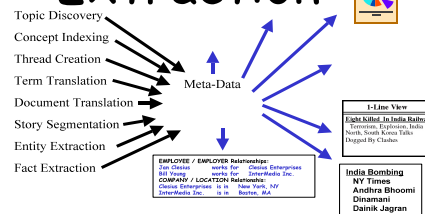
- interpretation
- transformation
- value filtering
- augmentation (providing metadata)

Two motivations:

- The amount of information in textual form
- Information integration needs NLP methods for coping with ambiguity and context

Knowledge Extraction Vision

Multi-dimensional Meta-data Extraction



Terms and technologies

- Text processing
 - Stuff like TextPad (Emacs, BBEdit), Perl, grep. Semantics and structure blind, but does what you tell it in a nice enough way. Still useful.
- Information Retrieval (IR)
 - Implies that the computer will try to find documents which are relevant to a user while understanding nothing (big collections)
- Intelligent Information Access (IIA)
 - Use of clever techniques to help users satisfy an information need (search or UI innovations)

Terms and technologies

- Locating small stuff. Useful nuggets of information that a user wants:
 - Information Extraction (IE): Database filling
 - The relevant bits of text will be found, and the computer will understand enough to satisfy the user's communicative goals
 - Wrapper Generation (WG) [or Wrapper Induction]
 - Producing filters so agents can "reverse engineer" web pages intended for humans back to the underlying structured data
 - Question Answering (QA) - NL querying
 - Thesaurus/key phrase/terminology generation

Terms and technologies

- Big Stuff. Overviews of data:
 - Summarization
 - Of one document or a collection of related documents (cross-document summarization)
 - Categorization (documents)
 - Including text filtering and routing
 - Clustering (collections)
- Text segmentation: subparts of big texts
- Topic detection and tracking
 - Combines IE, categorization, segmentation

Terms and technologies

- Digital libraries
- Text (Data) Mining (TDM)
 - Extracting nuggets from text. Opportunistic.
 - Unexpected connections that one can discover between bits of human recorded knowledge.
- Natural Language Understanding (NLU)
 - Implies an attempt to completely understand the text ...
- Machine translation (MT), OCR, Speech recognition, etc.
 - Now available wherever software is sold!

Database Interfaces

- Was going to be the big application of NLP in the 1980s
 - How many service calls did we receive from Europe last month?
 - I am listing the total service calls from Europe for November 2001.
 - The total for November 2001 was 1756.
- It has been recently integrated into MS SQL Server (English Query)
- Problems: need largely hand-built custom semantic support (improved wizards in new version!)
 - GUIs more tangible and effective?

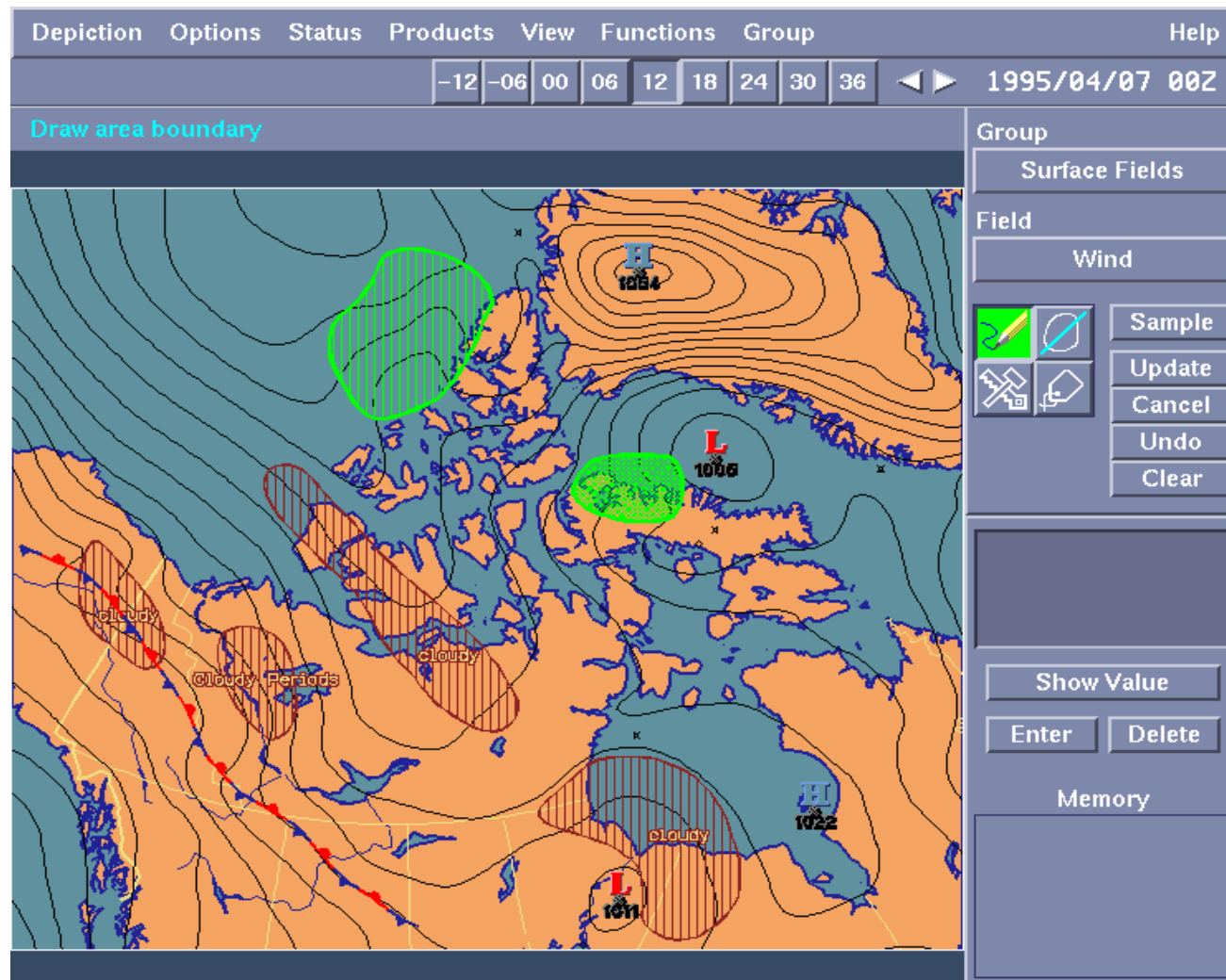
Lunar (Woods, 1971)

- Moon rock chemistry
- ATN and procedural semantics
 - sentences into procedure calls acting on db
- 78% of requests without error
- 90% when dictionary corrected
- Non-realistic numbers

Example : FoG

- Produces textual weather reports in English and French
- Input:
 - Graphical/numerical weather depiction
- User:
 - Environment Canada (Canadian Weather Service)

FoG: Input



FoG: Output

FPCN20 Status: CURRENT-NOT RELEASED

FPCN20 CWEG 152300
MARINE FORECASTS FOR ARCTIC WATERS ISSUED BY THE ARCTIC WEATHER CENTRE
OF ENVIRONMENT CANADA AT 05.00 PM MDT SATURDAY 15 APRIL 1995 FOR TONIGHT
AND SUNDAY WITH AN OUTLOOK FOR MONDAY.
THE NEXT SCHEDULED FORECAST WILL BE ISSUED AT 05.00 AM MDT.
WINDS ARE IN KNOTS.
FOG IMPLIES VISIBILITY LESS THAN 5/8 NM.
MIST IMPLIES VISIBILITY 5/8 TO 6 NM.

GREAT SLAVE LAKE.
WINDS LIGHT TONIGHT AND SUNDAY. SNOW ENDING NEAR MIDNIGHT. VISIBILITIES
NEAR 2 NM IN SNOW.
OUTLOOK FOR MONDAY... LIGHT WINDS.

GREAT BEAR LAKE.
FREEZING SPRAY WARNING ISSUED.
WINDS EAST 20 TO 25 TONIGHT AND SUNDAY. FREEZING SPRAY.
OUTLOOK FOR MONDAY... WINDS EASTERLY 20 TO 25.

MACKENZIE RIVER FROM MILE 0 TO MILE 100.
WINDS LIGHT TONIGHT AND SUNDAY. SNOW ENDING THIS EVENING. VISIBILITIES
NEAR 2 NM IN SNOW.
OUTLOOK FOR MONDAY... LIGHT WINDS.

MACKENZIE RIVER FROM MILE 100 TO MILE 300.
WINDS LIGHT STRENGTHENING TO SOUTHEAST 15 SUNDAY AFTERNOON. SNOW ENDING
EARLY THIS EVENING. VISIBILITIES NEAR 2 NM IN SNOW.
OUTLOOK FOR MONDAY... WINDS SOUTHEASTERLY 15.

Forecasts

-Marine--
* ARWC **
FPCN20
FPCN21
FPCN22/74
FPCN23/75
FPCN24/76
FPCN25/77
UL 22/83
-Public--
FPCN15

Set Element Priority ...

Set Active Areas ...

Source

☒ Working Version
☐ Official Release
☐ Forecast Rollup

Language

☒ English
☐ French

Generate

Update

Edit ...

Release

Print

Close

Help

Word Sense Disambiguation

- It's a no-brainer that NLP should be useful and used for web search (and IR in general):
 - Search for 'Jaguar'
 - the computer should know or ask whether you're interested in big cats [scarce on the web], cars, or, perhaps a molecule geometry and solvation energy package, or a package for fast network I/O in Java
 - Search for 'Michael Jordan'
 - The basketballer or the machine learning guy?
 - Search for laptop, don't find notebook
 - Google doesn't even *stem*:
 - Search for *probabilistic model*, and you don't even match pages with *probabilistic models*.

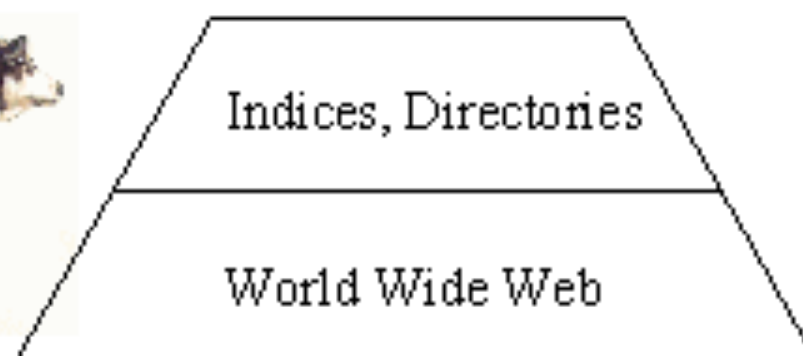
WSD

- Word sense disambiguation technology generally works well (like text categorization)
- Synonyms can be found or listed
- Lots of people have been into fixing this
 - e-Cyc had a beta version with Hotbot that disambiguated senses, and was going to go live in 2 months ... 14 months ago
 - Lots of startups:
 - LingoMotors
 - iPhrase "Traditional keyword search technology is hopelessly outdated"

Information Food Chain II



Personal
Assistants



Mass
Services

NLP for IR/web search?

- Methods which use of rich ontologies, etc., can work very well for intranet search within a customer's site (where anchor-text, link, and click patterns are much less relevant)
 - But don't really scale to the whole web
- *Moral: it's hard to beat keyword search for the task of general ad hoc document retrieval*
- *Conclusion: one should move up the food chain to tasks where finer grained understanding of meaning is needed*

Product information

CNET.com - Shopping - Search Results for "ibm x21" - Microsoft Internet Explorer

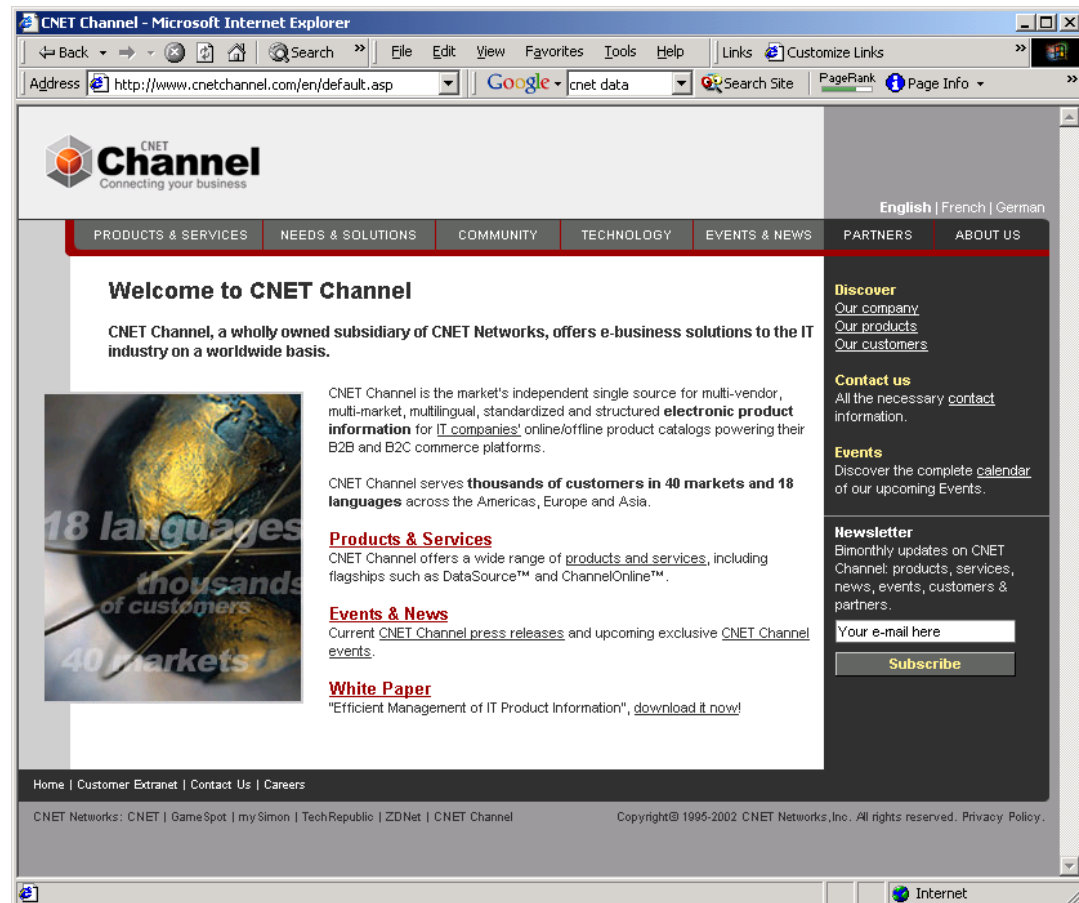
Found: 12 Displaying: 1-12

Resort by	Product Name	Manufacturer Name	CNET Review	Lowest Price
Most Popular	256MB PC100 IBM THINKPAD A20M A21 A22 X21 X20 T21 T22 390X 600X Add to my list	Crucial Technology		Check Latest Prices ▶ Price range: \$83.20-\$141.95
	THINKPAD X21 P3-700 20GB 128MB W2K 12-SVGA ENET INTEL 56K Add to my list	IBM Corp.	product info	Check Latest Prices ▶ Price range: \$1899.00-\$2923.36
	IBM Thinkpad X21 (Pentium III, 700 MHz, 128 MB, 20 GB) Add to my list	IBM Corp.	product info	Check Latest Prices ▶ Price range: \$2489.00-\$2996.47
	THINKPAD X21 P3-700 20GB 128MB 98 12-XGA ENET INTEL 56K Add to my list	IBM Corp.	product info	Check Latest Prices ▶ Price range: \$2048.95-\$2818.91
	IBM ThinkPad X21 (Pentium III 700MHz, 128MB RAM, 20GB) Add to my list	IBM Corp.	review 	Check Latest Prices ▶ Price range: \$2339.00-\$2818.91
	TP X21 NB P3/700 128MB 20GB 12.1 56K ETH W2K Add to my list	IBM Corp.	product info	Check Latest Prices ▶ Price range: \$2420.39-\$2925.21
	IBM ThinkPad X21 (Pentium III, 700 MHz, 128 MB, 20 GB) Add to my list	IBM Corp.	product info	Check Latest Prices ▶ Price range: \$2322.00-\$2892.02
	THINKPAD X21 P3-700 20GB 128MB 98 12 SVGA ENET	IBM Corp.		Check Latest Prices ▶ Price range: \$2378.96-\$2818.91

Done Internet

Product info

- C-net markets this information
- How do they get most of it?
 - Phone calls
 - Typing.



Inconsistency: digital cameras

- Image Capture Device: 1.68 million pixel 1/2-inch CCD sensor
- Image Capture Device Total Pixels Approx. 3.34 million Effective Pixels Approx. 3.24 million
- Image sensor Total Pixels: Approx. 2.11 million-pixel
- Imaging sensor Total Pixels: Approx. 2.11 million
1,688 (H) x 1,248 (V)
- CCD Total Pixels: Approx. 3,340,000 (2,140[H] x 1,560 [V])
 - Effective Pixels: Approx. 3,240,000 (2,088 [H] x 1,550 [V])
 - Recording Pixels: Approx. 3,145,000 (2,048 [H] x 1,536 [V])
- *These all came off the same manufacturer's website!!*
- And this is a very technical domain. Try sofa beds.



Comparison shopping

- Need to learn to extract info from online vendors
- Can exploit uniformity of layout, and (partial) knowledge of domain by querying with known products
- E.g., Jango Shopbot (Etzioni and Weld)
 - Gives convenient aggregation of online content
- Bug: not popular with vendors
 - A partial solution is for these tools to be personal agents rather than web services

Email handling

- Big point of pain for many people
- There just aren't enough hours in the day
 - even if you're not a customer service rep
- What kind of tools are there to provide an electronic secretary?
 - Negotiating routine correspondence
 - Scheduling meetings
 - Filtering junk
 - Summarizing content
- "The web's okay to use; it's my email that is out of control"

Text Categorization Uses

- Take a document and assign it a label representing its content (MeSH heading, ACM keyword, Yahoo category)
- Classic example: decide if a newspaper article is about politics, business, or sports?
- There are many other uses for the same technology:
 - Is this page a laser printer product page?
 - Does this company accept overseas orders?
 - What kind of job does this job posting describe?
 - What kind of position does this list of responsibilities describe?
 - What position does this this list of skills best fit?
 - Is this the "computer" or "harbor" sense of *port*?

Text Categorization

- Usually, simple machine learning algorithms are used.
- Examples: Naïve Bayes models, decision trees.
- Very robust, very re-usable, very fast.
- Recently, slightly better performance from better algorithms
 - e.g., use of support vector machines, nearest neighbor methods, boosting
- Accuracy is more dependent on:
 - Naturalness of classes.
 - Quality of features extracted and amount of training data available.
- Accuracy typically ranges from 65% to 97% depending on the situation
 - Note particularly performance on rare classes

Email response: "eCRM"

- Automated systems which attempt to categorize incoming email, and to automatically respond to users with standard, or frequently seen questions
- Most but not all are more sophisticated than just keyword matching
- Generally use text classification techniques
 - E.g., Echomail, Kana Classify, Banter
 - More linguistic analysis: YY software
- Can save real money by doing 50% of the task close to 100% right

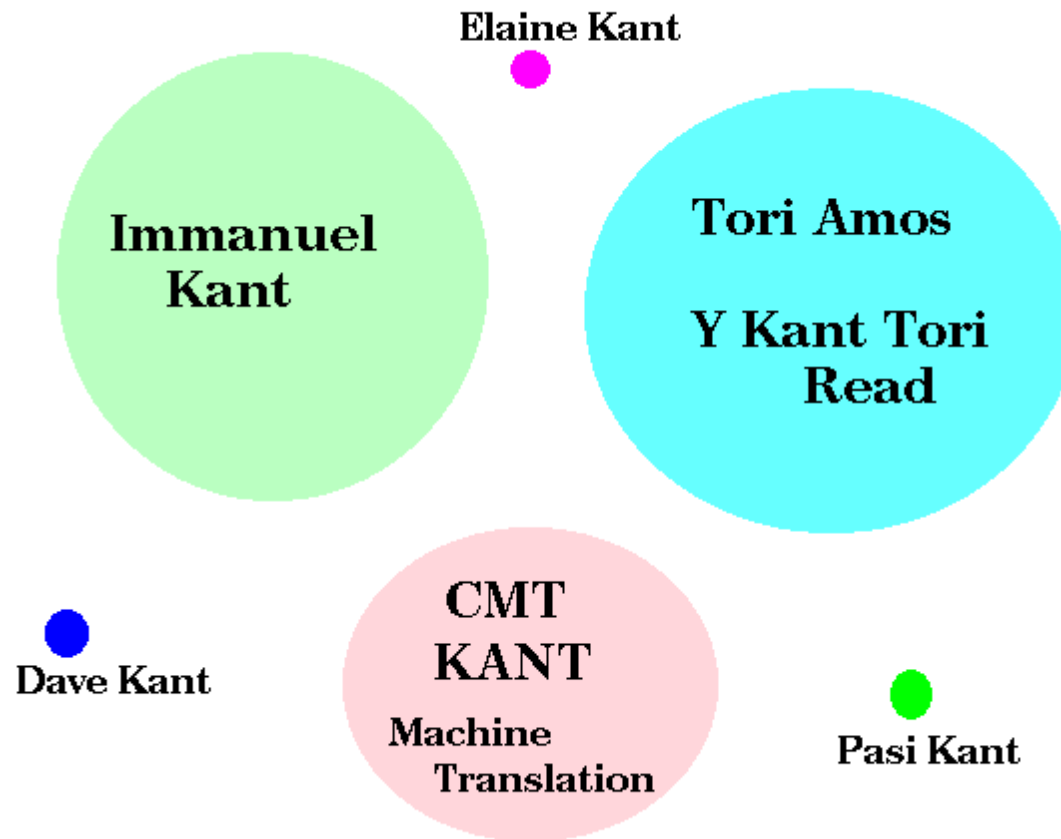
Financial markets

- Quantitative data are (relatively) easily and rapidly processed by computer systems, and consequently many numerical tools are available to stock market analysts
 - However, a lot of these are in the form of (widely derided) technical analysis
 - It's meant to be *information* that moves markets
- Financial market players are overloaded with qualitative information - mainly news articles - with few tools to help them (beyond people)
 - Need tools to identify, summarize, and partition information, and to generate meaningful links

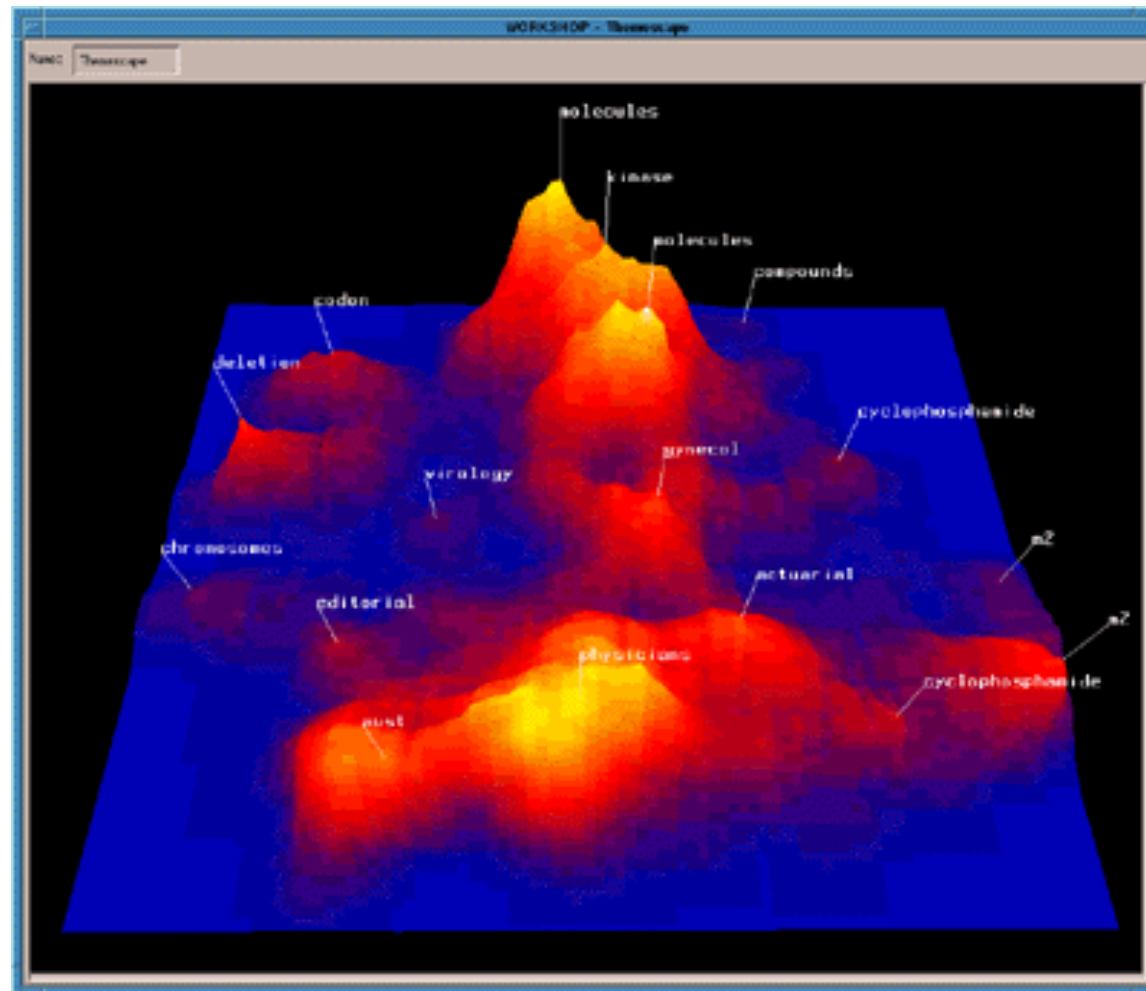
Text Clustering

- Scatter/Gather Clustering
 - Cutting, Pedersen, Karger, Tukey '92, '93
- Cluster sets of documents into general "themes", like a table of contents
- Display the contents of the clusters by showing topical terms and typical titles
- User chooses subsets of the clusters and re-clusters the documents within them
- Resulting new groups have different "themes"

Clustering (of query *Kant*)



Clustering a Multi-Dimensional Document Space



Clustering

- June 11, 2001: The latest KDnuggets Poll asked: What types of analysis did you do in the past 12 months.
 - The results, multiple choices allowed, indicate that a wide variety of tasks is performed by data miners. Clustering was by far the most frequent (22%), followed by Direct Marketing (14%), and Cross-Sell Models (12%)
- Clustering of results can work well in certain domains (e.g., biomedical literature)
- But it doesn't seem compelling for the average user, it appears (Altavista, Northern Light, Vivisimo)

CiteSeer/ResearchIndex

- An online repository of papers, with citations, etc. Specialized search with semantics in it
- Great product; research people love it
- However it's fairly low tech. NLP could improve on it:
 - Better parsing of bibliographic entries
 - Better linking from author names to web pages
 - Better resolution of cases of name identity
 - E.g., by also using the paper content
 - Cf. Cora, which did some of these tasks better

Chats, Forums, etc.

- Many of these are public on the web
- The signal to noise ratio is very low
- But there's still lots of good information there
- Some of it has commercial value
 - What problems have users had with your product?
 - Why did people end up buying product X rather than your product Y
- Some of it is time sensitive
 - Rumors on chat rooms can affect stockprice
 - Regardless of whether they are factual or not

Small devices

- With a big monitor, humans can scan for the right information
- On a small screen, there's *hugely* more value from a system that can show you what you want:
 - phone number
 - business hours
 - email summary
 - "Call me at 11 to finalize this"



Machine translation

- High quality MT is still a distant goal
- But MT is effective for scanning content
- And for machine-assisted human translation
- Dictionary use accounts for about half of a traditional translator's time.
- Printed lexical resources are not up-to-date
- Electronic lexical resources ease access to terminological data.
- "Translation memory" systems: remember previously translated documents, allowing automatic recycling of translations

Jackson & Moulinier

The Web really changed everything, because there was suddenly a pressing need to process large amounts of text, and there was also a ready-made vehicle for delivering it to the world. Technologies such as information retrieval, information extraction, and text categorization no longer seemed quite so arcane to upper management. The applications were, in some cases, obvious to anyone with half a brain; all one needed to do was demonstrate that they could be built and made to work, which we proceeded to do."

Information Extraction

Suppositions:

- A lot of information that *could* be represented in a structured semantically clear format isn't
- It may be costly, not desired, or not in one's control (screen scraping) to change this.
- Goal: being able to answer semantic queries using "unstructured" natural language sources

Information Extraction

- Information extraction systems
 - Find and understand relevant parts of texts.
 - Produce a structured representation of the relevant information: *relations* (in the DB sense)
 - Combine knowledge about language and the application domain
 - Automatically extract the desired information
- When is IE appropriate?
 - Clear, factual information (who did what to whom and when?)
 - *Only a small portion of the text is relevant.*
 - ***Some errors can be tolerated***

Wrapper Induction

- Wrapper Induction
 - Sometimes, the relations are structural.
 - Web pages generated by a database.
 - Tables, lists, etc.
 - Wrapper induction is usually regular relations which can be expressed by the *structure* of the document:
 - the item in bold in the 3rd column of the table is the price
- Handcoding a wrapper in Perl isn't very viable
 - sites are numerous, and their surface structure mutates rapidly
- Wrapper induction techniques can also learn:
 - If there is a page about a research project X and there is a link near the word 'people' to a page that is about a person Y then Y is a member of the project X.
 - [e.g, Tom Mitchell's Web→KB project]

Existing IE Systems

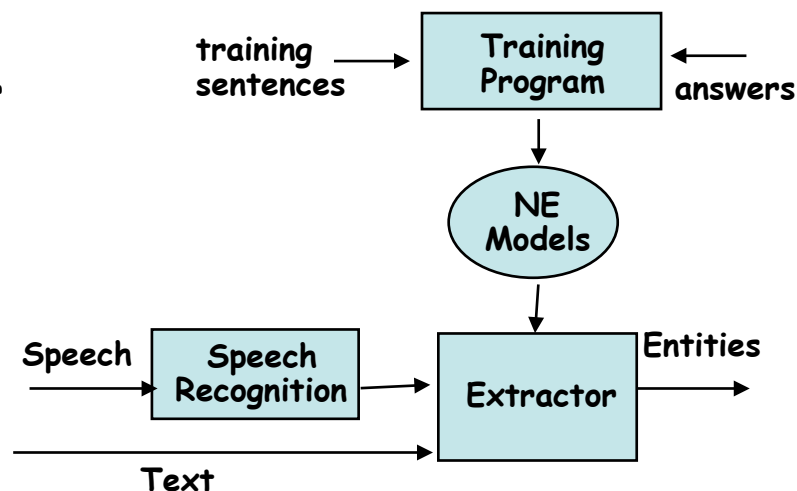
- Systems to summarize medical patient records by extracting diagnoses, symptoms, physical findings, test results, and therapeutic treatments.
- Gathering earnings, profits, board members, etc. from company reports
- Verification of construction industry specifications documents (are the quantities correct/reasonable?)
- Real estate advertisements
- Building job databases from textual job vacancy postings
- Extraction of company take-over events
- Extracting gene locations from biomed texts

3 generations of IE systems

- Hand-Built Systems - Knowledge Engineering [1980s-]
 - Rules written by hand
 - Require experts who understand both the systems and the domain
 - Iterative guess-test-tweak-repeat cycle
- Automatic, Trainable Rule-Extraction Systems [1990s-]
 - Rules discovered automatically using predefined templates, using methods like ILP
 - Require huge, labeled corpora (effort is just moved!)
- Statistical Generative Models [1997 -]
 - One decodes the statistical model to find which bits of the text were relevant, using HMMs or statistical parsers
 - Learning usually supervised; may be partially unsupervised

Name Extraction via HMMs

The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.



The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.

- Prior to 1997 - no learning approach competitive with hand-built rule systems
- Since 1997 - Statistical approaches (BBN, NYU, MITRE, CMU/JustSystems) achieve state-of-the-art performance

- Locations
- Persons
- Organizations

Classified Advertisements

Background:

- Advertisements are plain text
- Lowest common denominator: only thing that 70+ newspapers with 20+ publishing systems can all handle

```
<ADNUM>2067206v1</ADNUM>
<DATE>March 02, 1998</DATE>
<ADTITLE>MADDINGTON $89,000</
ADTITLE>
<ADTEXT>
OPEN 1.00 - 1.45<BR>
U 11 / 10 BERTRAM ST<BR>
NEW TO MARKET Beautiful<BR>
3 brm freestanding<BR>
villa, close to shops & bus<BR>
Owner moved to Melbourne<BR>
ideally suit 1st home buyer,<BR>
investor & 55 and over.<BR>
Brian Hazelden 0418 958 996<BR>
R WHITE LEEMING 9332 3477
</ADTEXT>
```

news real estate

--Please Choose--

- [New Search](#)
- [Return to Listing](#)
- [Guided Tour](#)

MEMBER LOGIN

Username

Password

ENTER

Press to
fill in an
Online
Application

PROPERTYMAP

Use Navigation Aids
to change chosen
area

ZOOM IN

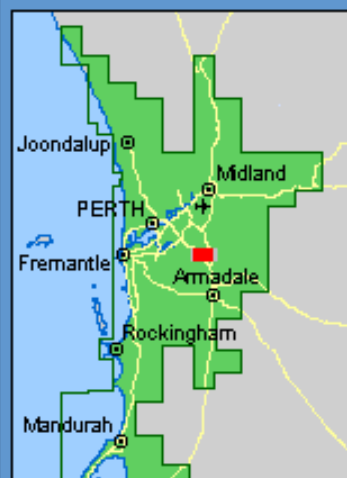
ZOOM OUT

UBD Reference:

"332 D10"



The Exact location was successfully mapped [0]



[Add to Inspection List](#) [Show More Detail](#)

Property Details

Address: 10 BERTRAM ST

Suburb: MADDINGTON

State: WA

Why doesn't text search (IR) work?

What you search for in real estate advertisements:

- Suburbs. You might think easy, but:
 - Real estate agents: Coldwell Banker, Mosman
 - Phrases: Only 45 minutes from Parramatta
 - Multiple property ads have different suburbs
- Money: want a range not a textual match
 - Multiple amounts: was \$155K, now \$145K
 - Variations: offers in the high 700s [*but not* rents for \$270]
- Bedrooms: similar issues (br, bdr, beds, B/R)

Machine learning

- To keep up with and exploit the web, you need to be able to *learn*
 - Discovery: How do you find new information sources S ?
 - Extraction: How can you access and parse the information in S ?
 - Semantics: How does one understand and link up the information in contained in S ?
 - Pragmatics: What is the accuracy, reliability, and scope of information in S ?
- Hand-coding just doesn't scale

Question Answering

- TREC 8/9 QA competition: an idea originating from the IR community
- With massive collections of on-line documents, manual translation of knowledge is impractical: we want answers from textbases [cf. bioinformatics]
- Evaluated output is 5 answers of 50/250 byte snippets of text drawn from a 3 Gb text collection, and required to contain at least one concept of the semantic category of the expected answer type. (IR think. Suggests the use of named entity recognizers.)
- Get reciprocal points for highest correct answer.

Pasca and Harabagiu (2001)

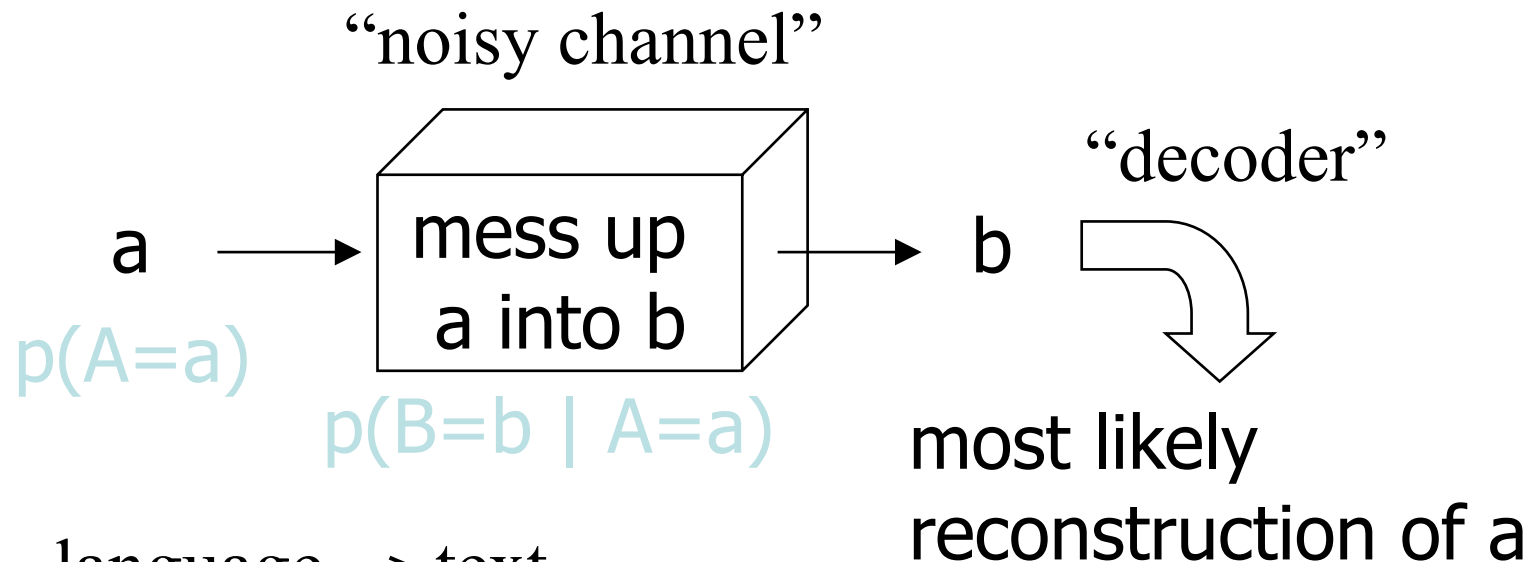
- Good IR is needed: paragraph retrieval based on SMART
- Large taxonomy of question types and expected answer types is crucial
- Statistical parser (modeled on Collins 1997) used to parse questions and relevant text for answers, and to build knowledge base
- Controlled query expansion loops (morphological, lexical synonyms, and semantic relations) are all important
- Answer ranking by simple ML method

Question-Answering

- How hot does the inside of an active volcano get?
- `get(TEMPERATURE, inside(volcano(active)))`
- "lava fragments belched out of the mountain were as hot as 300 degrees Fahrenheit"
- `fragments(lava, TEMPERATURE(degrees(300)),
belched(out, mountain))`
 - volcano ISA mountain
 - lava ISPARTOF volcano • lava inside volcano
 - fragments of lava HAVEPROPERTIESOF lava
- The needed semantic information is in WordNet definitions, and was successfully translated into a form that can be used for rough 'proofs'

Machine Translation

Noisy Channel



language $\rightarrow$ text
 text $\rightarrow$ speech
 spelled $\rightarrow$ misspelled
 English $\rightarrow$ French

$$\begin{aligned}
 &\text{maximize } p(A=a | B=b) \\
 &= p(A=a) p(B=b | A=a) / (B=b) \\
 &= p(A=a) p(B=b | A=a) \\
 &\quad / \sum p(A=a') p(B=b | A=a')
 \end{aligned}$$

Machine Translation

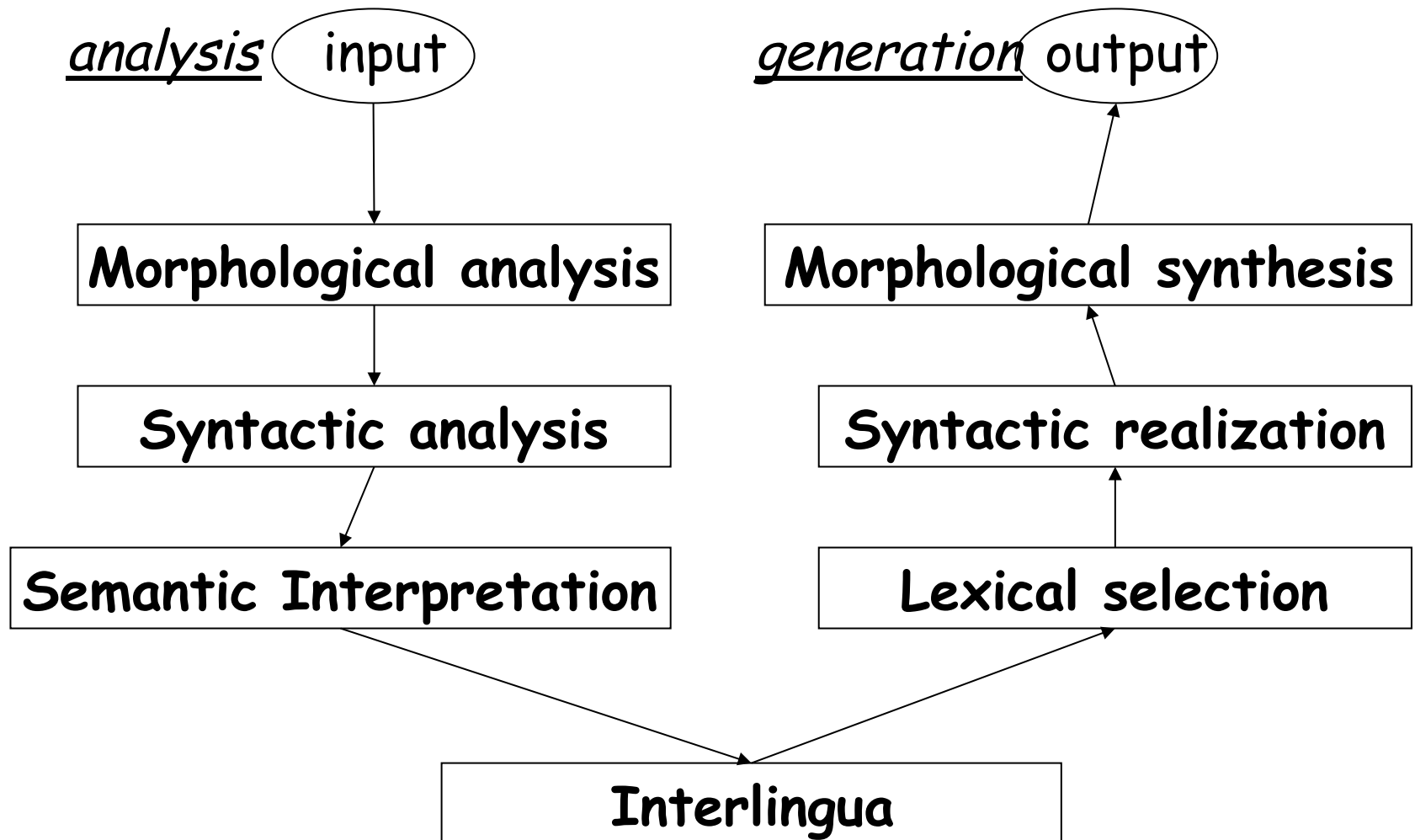
- Translation from one natural language to another by means of a computerized system
- Early failures
- Later: varying degrees of success

An Old Example

The spirit is willing but the flesh is weak

The vodka is good but the meat is rotten

Machine Translation



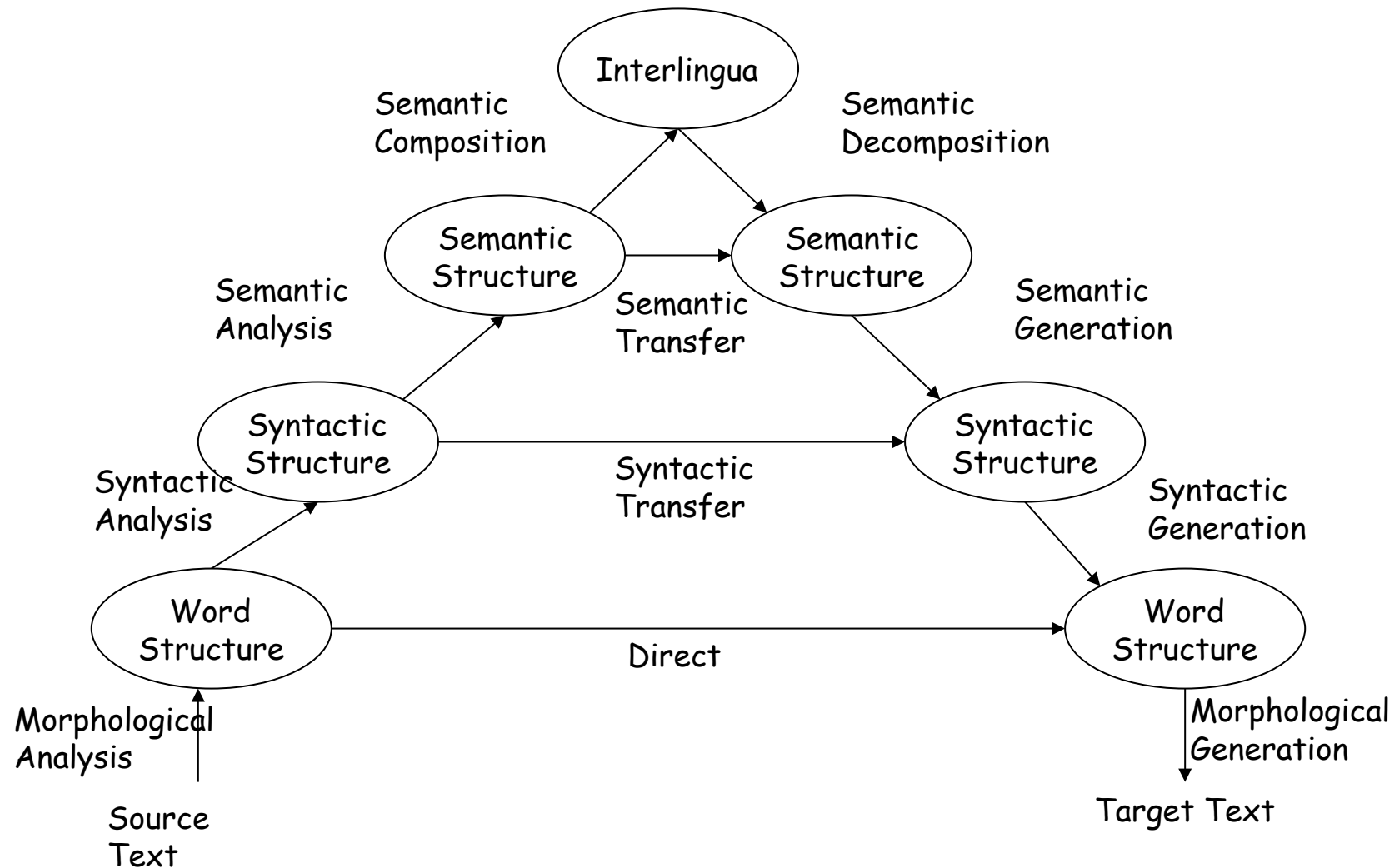
Machine Translation History

- 1950's: Intensive research activity in MT
- 1960's: Direct word-for-word replacement
- 1966 (ALPAC): NRC Report on MT
- Conclusion: MT no longer worthy of serious scientific investigation.
- 1966-1975: Recovery period
- 1975-1985: Resurgence (Europe, Japan)
- 1985-present: Resurgence (US)
- Internet

What happened?

- Need for MT and other NLP applications confirmed
- Change in expectations
- Computers have become faster, more powerful
- WWW
- Political state of the world
- Maturation of Linguistics
- Development of hybrid statistical/symbolic approaches

MT Approaches



Three Approaches

- Direct:
 - I checked his answers against those of the teacher →
Yo comparé sus respuestas a las de la profesora
 - Rule: [check X against Y] → [comparar X a Y]
- Transfer:
 - Ich habe ihn gesehen → I have seen him
 - Rule: [clause agt aux obj pred] → [clause agt aux pred obj]
- Interlingual:
 - I like Mary → Mary me gusta a mí
 - Rep: [Be_{Ident} (I [AT_{Ident} (I, Mary)] Like+ingly)]

Direct

- Pros
 - Fast
 - Simple
 - Inexpensive
- Cons
 - Unreliable
 - Not powerful
 - Rule proliferation
 - Requires too much context
 - Major restructuring after lexical substitution

Transfer

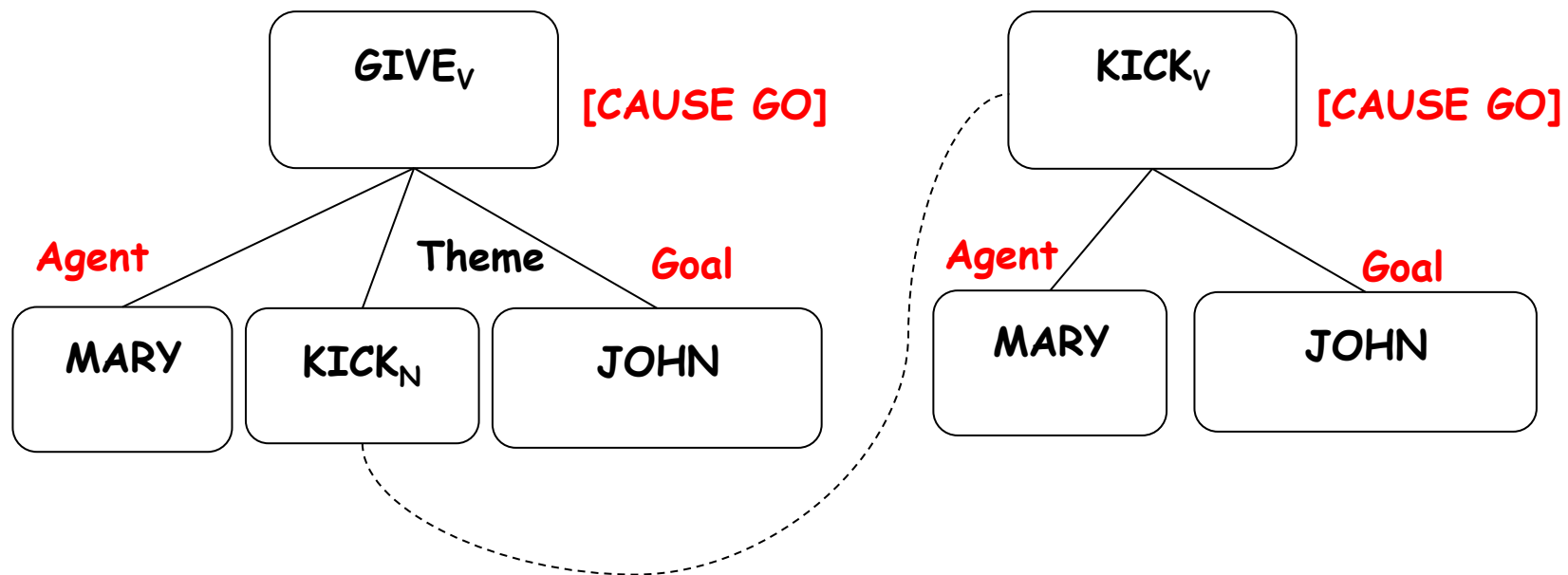
- Pros
 - Don't need to find language-neutral rep
 - No translation rules hidden in lexicon
 - Relatively fast
- Cons
 - N^2 sets of transfer rules: Difficult to extend
 - Proliferation of language-specific rules in lexicon and syntax
 - Cross-language generalizations lost

Interlingual

- Pros
 - Portable (avoids N^2 problem)
 - Lexical rules and structural transformations stated more simply on normalized representation
 - Explanatory Adequacy
- Cons
 - Difficult to deal with terms on primitive level: universals?
 - Must decompose and reassemble concepts
 - Useful information lost (paraphrase)

Mapping from Input Dependency to English Dependency Tree

Mary le dio patadas a John → *Mary kicked John*



Knowledge Resources in English only: (LVD; Dorr, 2001).

Statistical Extraction

Mary kicked John . [0.670270]
Mary gave a kick at John . [-2.175831]
Mary gave the kick at John . [-3.969686]
Mary gave an kick at John . [-4.489933]
Mary gave a kick by John . [-4.803054]
Mary gave a kick to John . [-5.045810]
Mary gave a kick into John . [-5.810673]
Mary gave a kick through John . [-5.836419]
Mary gave a foot wound by John . [-6.041891]
Mary gave John a foot wound . [-6.212851]

MT Challenges: Ambiguity

- Syntactic Ambiguity
I saw the man on the hill with the telescope
- Lexical Ambiguity
E: book
S: libro, reservar
- Semantic Ambiguity
 - Homography:
ball(E) = pelota, baile(S)
 - Polysemy:
kill(E), matar, acabar (S)
 - Semantic granularity
esperar(S) = wait, expect, hope (E)
be(E) = ser, estar(S)
fish(E) = pez, pescado(S)

How do we evaluate MT?

- Human-based Metrics
 - Semantic Invariance
 - Pragmatic Invariance
 - Lexical Invariance
 - Structural Invariance
 - Spatial Invariance
 - Fluency
 - Accuracy
 - "Do you get it?"
- Automatic Metrics: Bleu

Bleu Comparison

Chinese-English Translation Example:

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Modified Unigram Precision: Candidate #1

It(1) is(1) a(1) guide(1) to(1) action(1) which(1)
ensures(1) that(2) the(4) military(1) always(1) obeys
(0) **the** commands(1) of(1) **the** party(1)

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

What's the answer??????

17/1

8

Modified Unigram Precision: Candidate #2

It(1) is(1) to(1) insure(0) the(4) troops(0) forever(1)
hearing(0) **the** activity(0) guidebook(0) that(2) party
(1) direct(0)

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

What's the answer??????

8/1

4

Modified Bigram Precision:

Candidate #1

It is(1) is a(1) a guide(1) guide to(1) to action(1) action which
(0) which ensures(0) ensures that(1) that the(1) the military(1)
military always(0) always obeys(0) obeys the(0) the commands
(0) commands of(0) of the(1) the party(1)

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

What's the answer?????

10/17

Modified Bigram Precision: Candidate #2

It is(1) is to(0) to insure(0) insure the(0) the troops
(0) troops forever(0) forever hearing(0) hearing the
(0) the activity(0) activity guidebook(0) guidebook
that(0) that party(0) party direct(0)

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Catching Cheaters

the(2) the the the(0) the(0) the(0) the
(0)

Reference 1: The cat is on the mat

Reference 2: There is a cat on the mat

What's the unigram answer? 2/7

What's the bigram answer? 0/7

Conclusion

Controversial questions

- Language organ
- Universal grammar
- Single dramatic mutation or gradual adaptation?

Many Links

<http://nlp.stanford.edu/links/statnlp.html>

Course Site

<http://www.cs.tau.ac.il/~nachum/NLP>

Thanks...

- Bob Berwick (MIT)
- Chris Manning (Stanford)
- Bonnie Dorr, Nizar Habash (Maryland)
- Peter Hancox (Birmingham)
- Paula Matuszek, Mary-Angela Papalaskari (Villanova)
- Hesham Feili (Sharif)
- Ehud Reiter (Aberdeen)
- Mona Diab, Alon Itai, UN, Ethnologue