Computational Linguistics

CSC 2501 / 485
Fall 2015

1

1. Introduction to computational linguistics

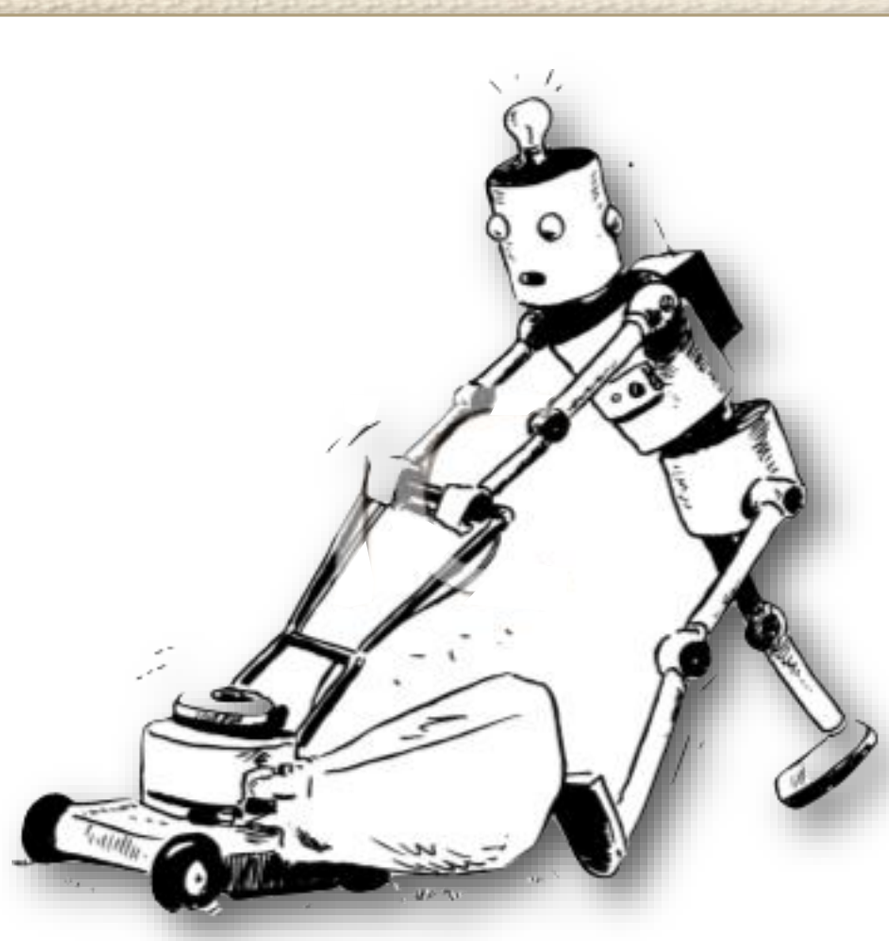
Frank Rudzicz
Toronto Rehabilitation Institute-UHN; and
Department of Computer Science, University of Toronto

Reading: Jurafsky & Martin: 1.
Bird et al: 1, [2.3, 4].

Why would a computer need to use natural language? Why would anyone want to talk to a computer?

Computer as autonomous agent.
Has to talk and understand like a human.

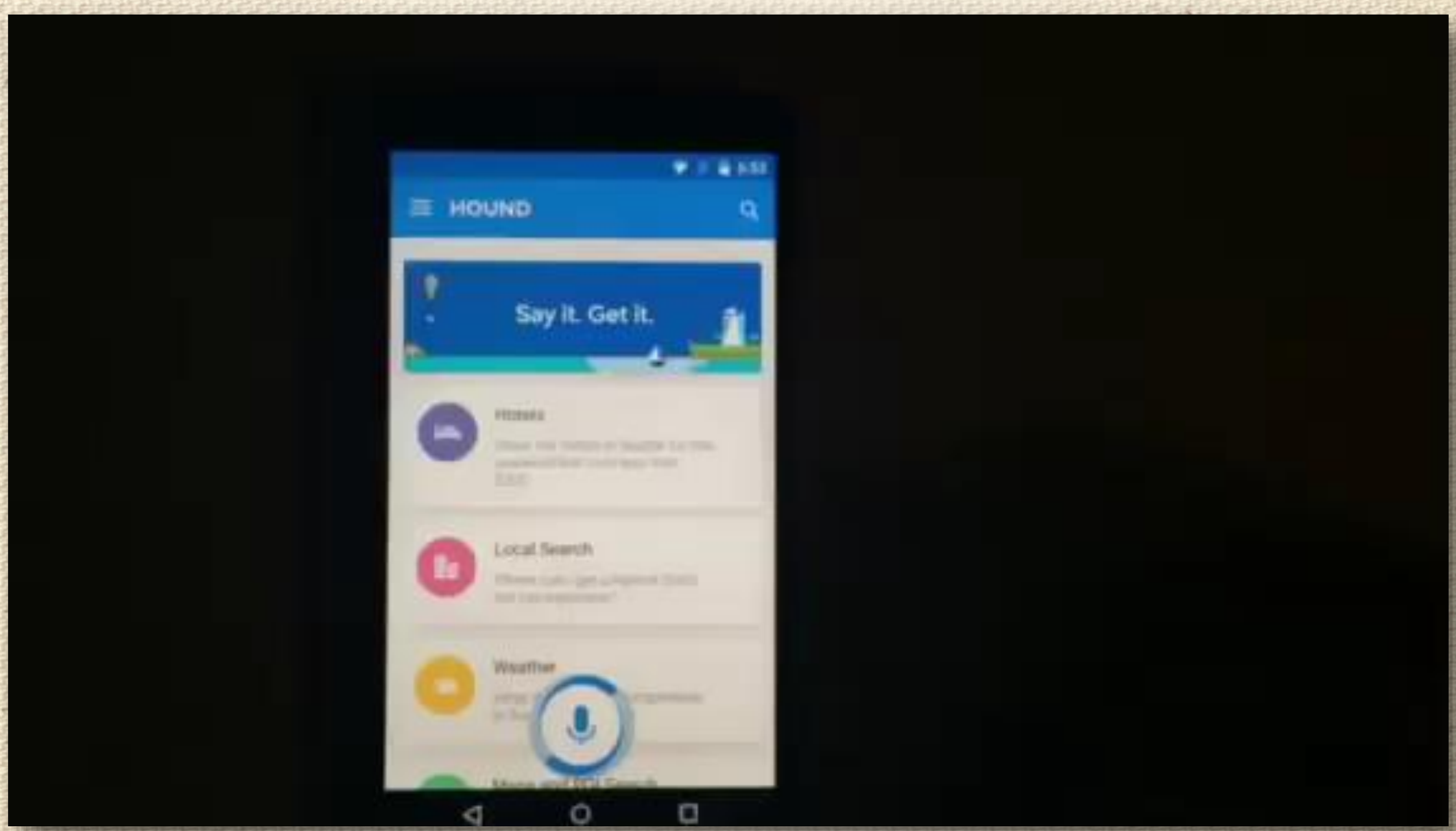
Computer as servant.
Has to take orders.



Computer as personal assistant.
Has to take orders.

Schedule a meeting tomorrow with George. **Book** me a flight to Vancouver for the conference. **Find** out *why* our sales have dropped in Lithuania. And **write** a thank-you note to my grandma for the birthday present.

Computer as personal assistant.
Has to take orders.



Computer as researcher.
Needs to read and listen to everything.

Computer as researcher.
Brings us the **information** we need.

Find me a well-rated **hotel** in or near **Stockholm** where they serve **vegetarian** food, but **not** one that has any **complaints** about noise.

Did people in 1878 really speak like the characters in *True Grit*?

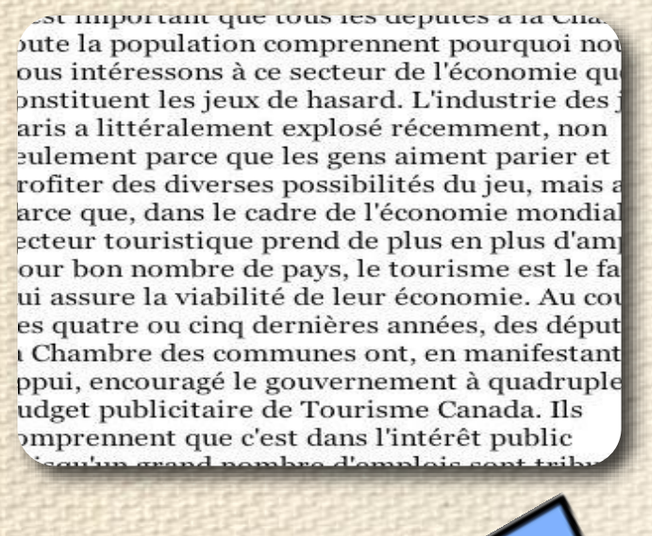
Are perfectly safe vaccines that save lives actually a government conspiracy?

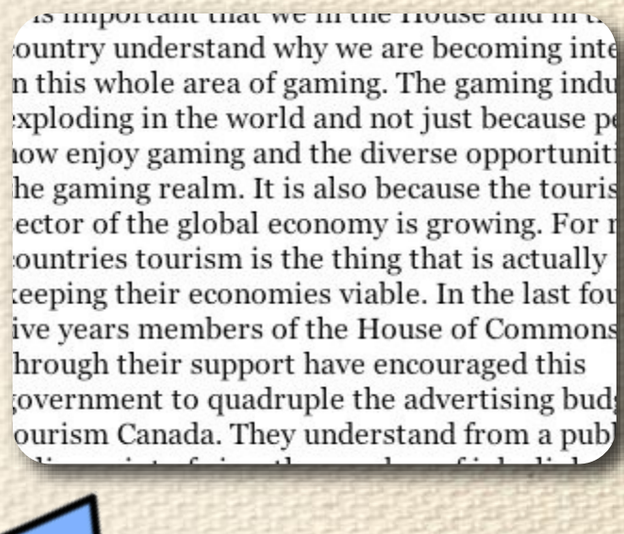
Computer as researcher.
Wins television game shows.

IBM's Watson on *Jeopardy!*, 16 February 2011

https://www.youtube.com/watch?v=yJptrlCVDHI
https://www.youtube.com/watch?v=Y2wQQ-xSE4s

Computer as language expert.
Translates our communications.





Input:

Spoken
Written

Output:

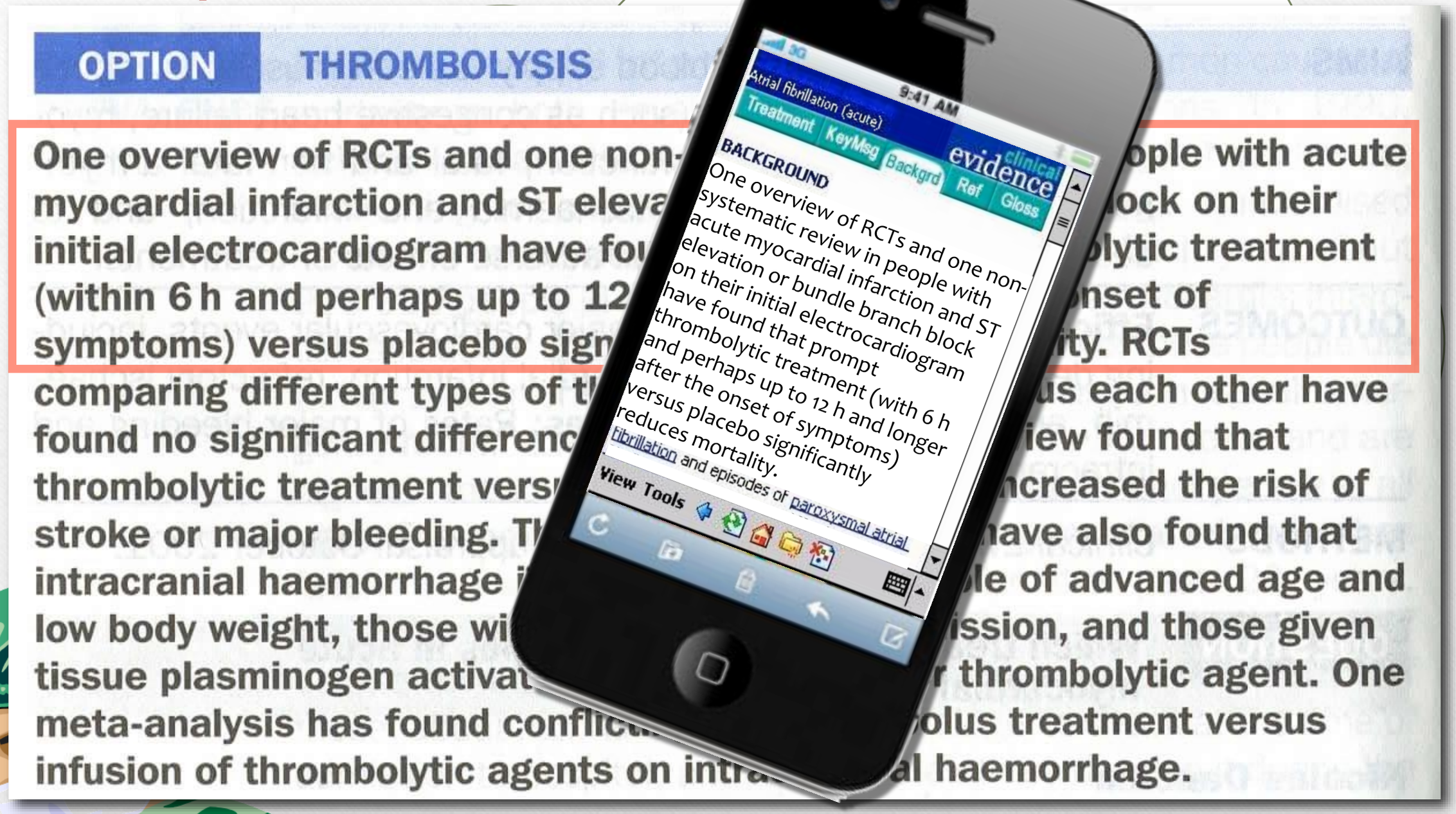
An action
A document or artifact
Some chosen text or speech
Some newly composed text or speech

Intelligent language processing

- Document applications
- Searching for documents by meaning
- Summarizing documents
- Answering questions
- Extracting information
- Content/authorship/sentiment analysis
- Helping language learners
- Helping people with disabilities

…

Example: Answering clinical questions at the point of care



Example: Early detection of Alzheimer's

- Look for deterioration in complexity of vocabulary and syntax.
- Study: Compare three British writers

Iris Murdoch

Died of Alzheimer's

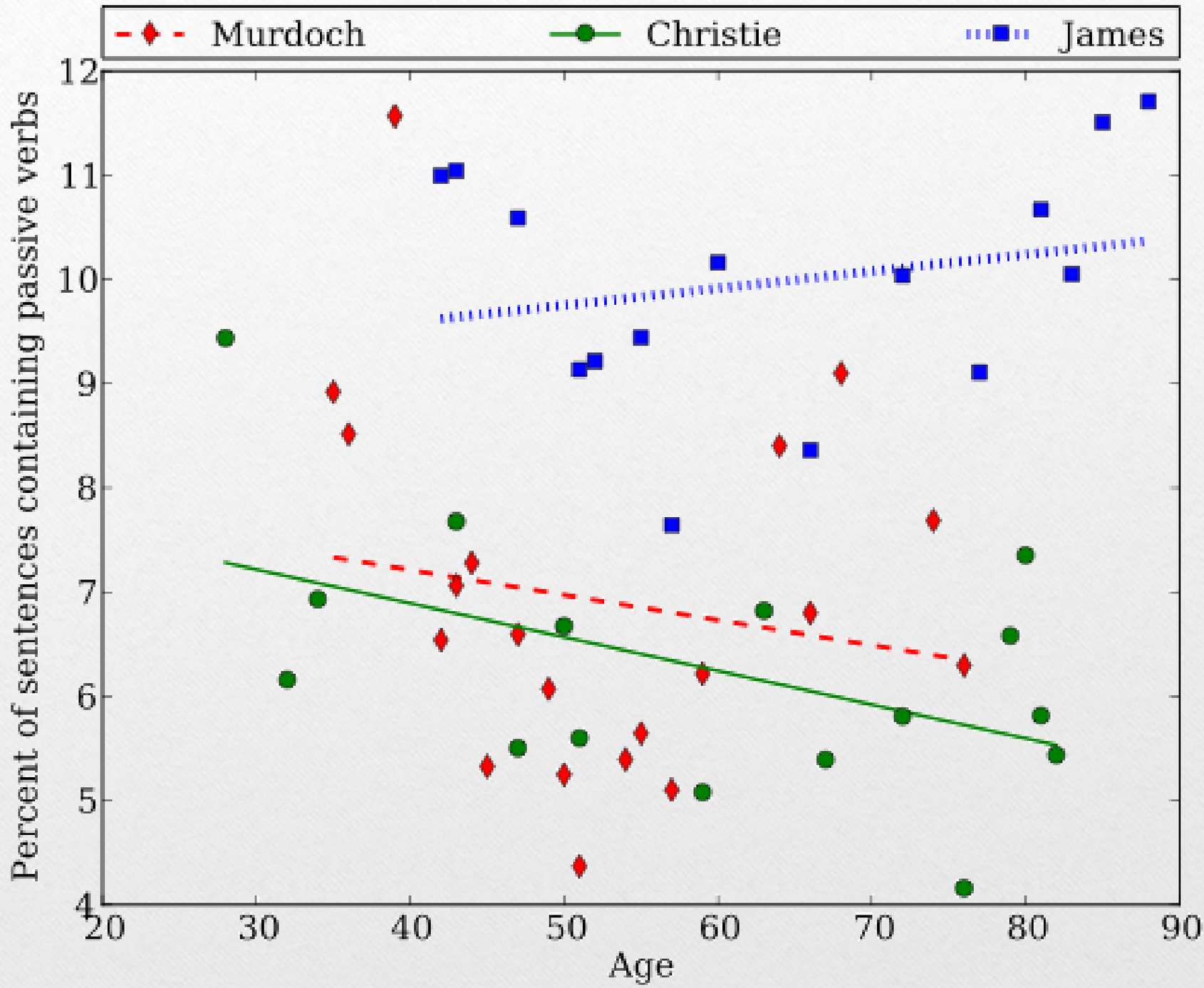
P.D. James

No Alzheimer's

Agatha Christie

Suspected Alzheimer's

Change in use of passive verbs

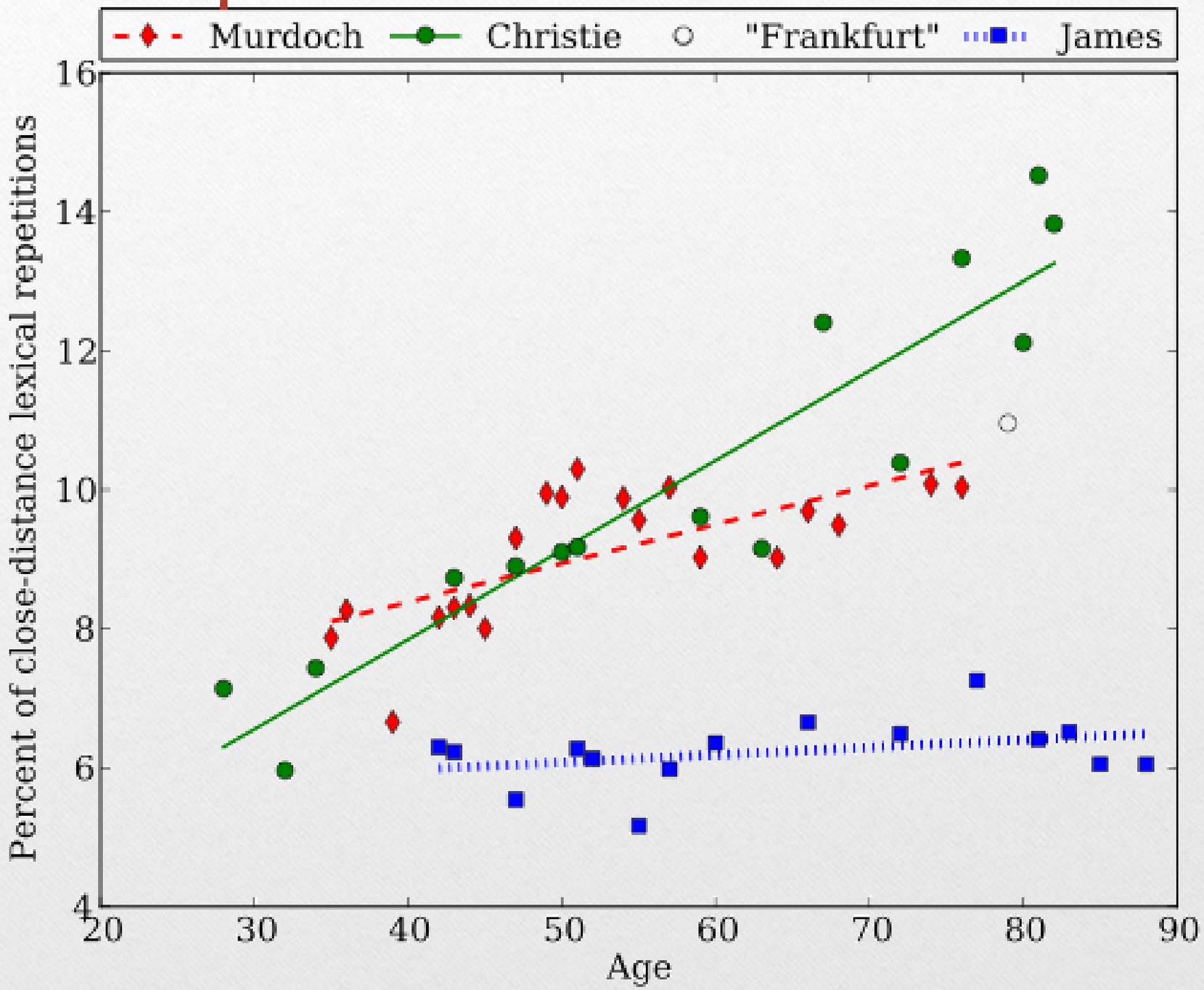


n.s.

n.s.; but deep trough in 40s–50s

Decline, $p = .054$

Increase in short-distance word repetition



Rise, $p < .01$

Rise, $p < .01$

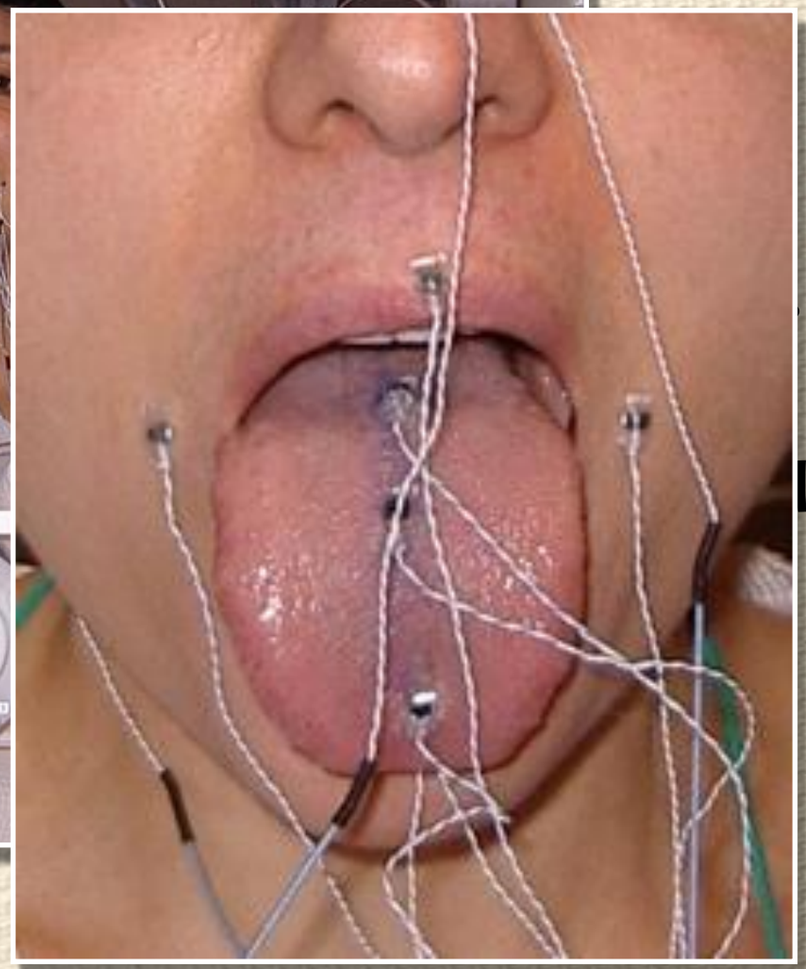
n.s.

Spoken documents

- **“Google for speech”**
  Search, indexing, and browsing through audio documents.
- **Speech summarization**
  Automatically select the 5–20% most important sentences of audio documents.

Speech recognition for

- Us ve
  sp
  sp
- Cr c
  sp dy

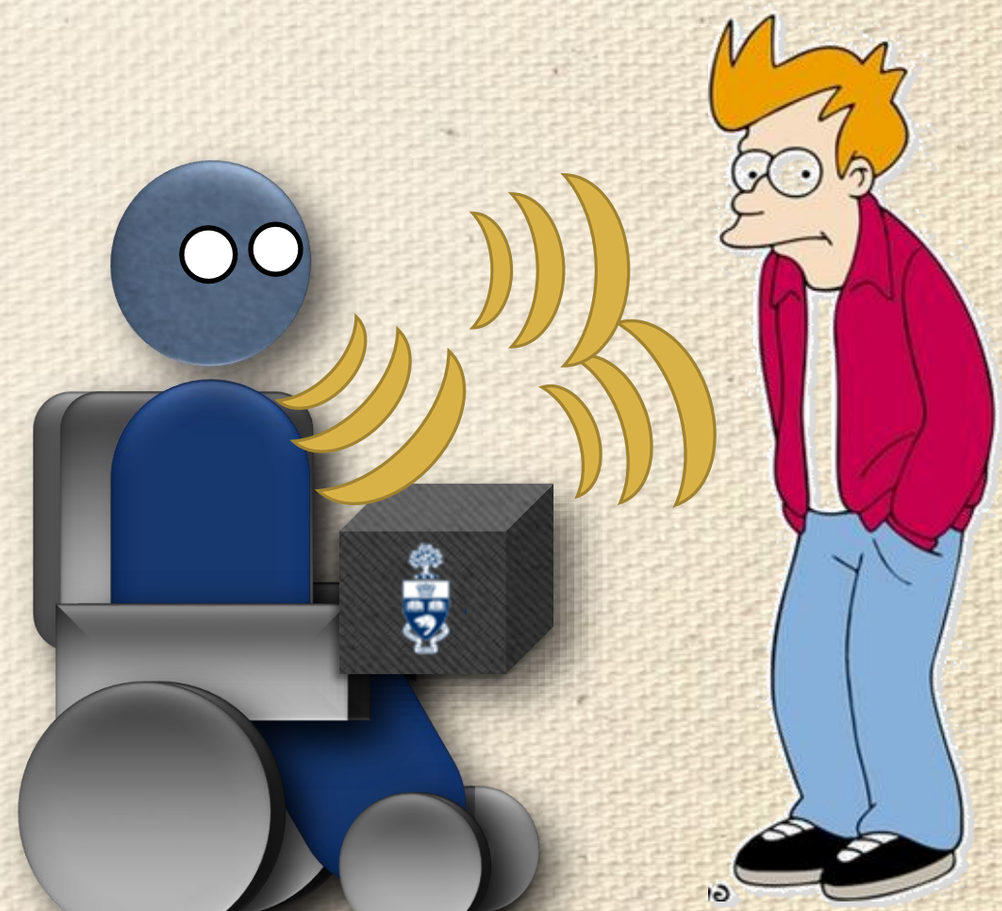


Bloorview KIDS REHAB

Speech-Language Pathology
UNIVERSITY OF TORONTO

Speech transformation for dysarthria

Transform dysarthric speech to improve comprehensibility



Quillsoft

Speech-Language Pathology
UNIVERSITY OF TORONTO

Models of human language processing

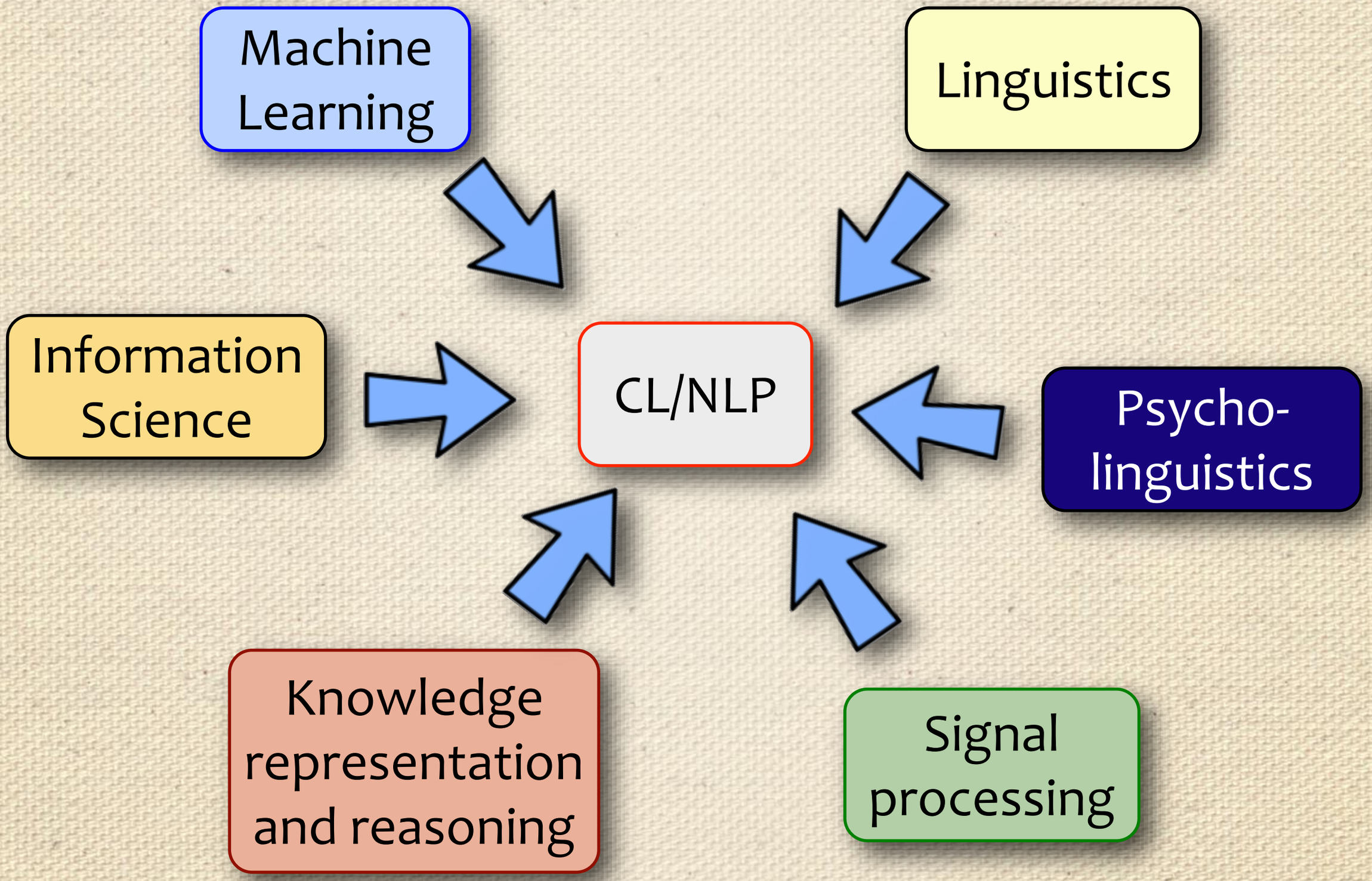
- Highly multidisciplinary approach
- Exploit the relation between **linguistic knowledge** and **statistical behaviour of words**

Models of children's language acquisition

- Models of how children learn their language just from what they hear and observe
- Apply machine-learning techniques to show how children can learn:
  - to map **words in a sentence** to **real world objects**
  - the relation between **verbs** and their **arguments**

Mathematics of syntax and language

- Discrete mathematical models of sentence structure
  - Typed feature logic: algorithms for efficient lexicalized parsing
- Parsing in freer-word-order languages


Machine Learning
Linguistics
Information Science
CL/NLP
Psycho-linguistics
Knowledge representation and reasoning
Signal processing

Computational linguistics 1

- Anything that brings together **computers** and **human languages** …
  - … using knowledge about the **structure** and **meaning** of language (*i.e.*, not just string processing)
- *The dream:* "The linguistic computer"
  - Human-like competence in language

Computational linguistics 2

- The development of computational models with natural language as input and/or output.
- ***Goal:*** A **set of tools** for processing language (semi-) automatically:
  - To access linguistic information easily and to transform it — *e.g.*, summarize, translate, ….
  - To facilitate communication with a machine.
- "NLP": Natural language processing.

Computational linguistics 3

- Use of computational models in the study of natural language.
- ***Goal:*** A **scientific theory** of communication by language:
  - To understand the structure of language and its use as a complex computational system.
  - To develop the data structures and algorithms that can implement/approximate that system.

What does it mean to “understand” language?

The Turing Test

In the first line of your sonnet which reads “Shall I compare thee to a summer’s day,” would not “a spring day” do as well or better?

It wouldn’t scan.

How about “a winter’s day”? That would scan all right.

Yes, but nobody wants to be compared to a winter’s day.

Alan Turing, “Computing machinery and intelligence”, *Mind*, 59, 1950, 433–460.

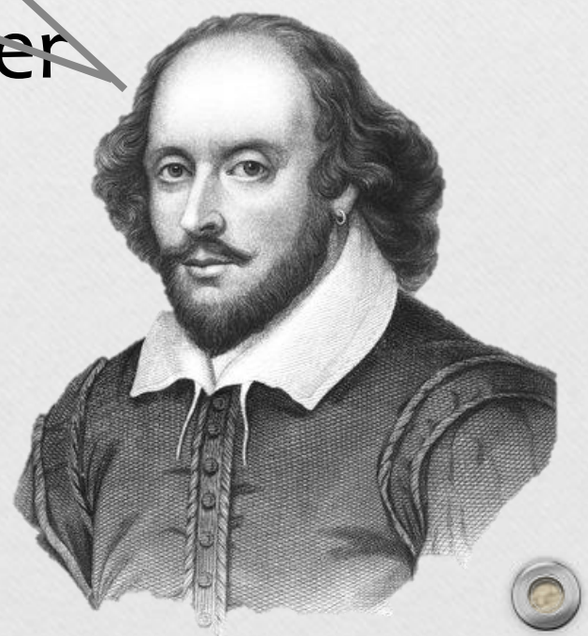
The Turing Test (cont.)

Would you say Mr Pickwick reminded you of Christmas?

In a way.

Yet Christmas is a winter's day, and I do not think Mr Pickwick would mind the comparison.

I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.



Alan Turing, "Computing machinery and intelligence", *Mind*, 59, 1950, 433–460.

Using language is not necessarily understanding language

~~Frontier~~ Computer Psychiatrist

Men are all alike.

In what way?

They're always bugging us about something or the other.

Can you think of a specific example?

Well, my boyfriend made me come here.

Your boyfriend made you come here.

He says I'm depressed.

Joseph Weizenbaum, *Computer Power and Human Reason*, W.H. Freeman, 1976.

~~Frontier~~ Computer Psychiatrist (cont.)

I'm sorry to hear you are depressed.

It's true. I am unhappy.

Do you think coming here will help you not to be unhappy?

Perhaps I could learn to get along with my mother.

Tell me more about your family.

Joseph Weizenbaum, *Computer Power and Human Reason*, W.H. Freeman, 1976.

Levels of understanding 1

- ***0. Keyword processing:***
  Limited knowledge of **particular words** or **phrases**, or their collocations.
  - Chatterbots.
  - Information retrieval.
  - Web searching.

Levels of understanding 2

- ***1. Limited linguistic ability:***
  Appropriate response to simple, highly constrained **sentences**.
  - Database queries in NL.
    "Show all sales staff who exceeded their quota in May."
  - Simple NL interfaces.
    "I want to fly from Toronto to Vancouver next Sunday."

Levels of understanding 3

- **2. *Full text comprehension:***
  Understanding of **multi-sentence** text and its relation to the "real world".
  - Conversational dialogue.
  - Automatic knowledge acquisition.

- ***3. Emotional understanding:***
  - Responding to **literature, poetry**.

Sources of knowledge

- Linguistics
  - Formal models of linguistic knowledge.
- Philosophy
  - Formal models of meaning, world knowledge.
- Psychology
  - Experiments on human linguistic processing.
- Information studies (cybernetics?)
  - Models of access and use of information.

The science of CL

- Formalisms: grammars, logics.
- Statistical and probabilistic modeling.
- Algorithms for combining the above.
- Automatic induction of linguistic information (machine learning).
- Cognitive modeling (two-way interaction between the fields).

Current research trends

- Emphasis on **large-scale** NLP applications.
  - *Combines:* language processing *and* machine learning.
- Availability of **large text corpora**, development of statistical methods.
  - *Combines:* grammatical theories *and* actual language use.
- Understanding the successes and limitations of statistical approaches.
  - *Combines:* statistical approaches *and* more-sophisticated linguistic knowledge.

Building blocks of CL systems 1

- Language interpretation, generation, and transfer (e.g., machine translation).
  - Part-of-speech (PoS) tagging.
  - Parsing and grammars.
  - Reference resolution.
  - Dialogue management.

Natural language interpretation

Does Flight AC2207 serve lunch?

YNQ (∃e SERVING(e) ∧ SERVER(e, flight-2207)
∧ SERVED(e, lunch))

Natural language generation

(spray-1 (OBJECT paint-1)
(PATH (path-1
(DESTINATION wall-1))))
(CAUSER sally-1)

Sally sprayed paint on the wall.

Machine translation

- Current systems based purely on statistical associations.
- Getting incrementally better as they learn from more data.
- Still very naïve linguistically.

Úvodní strana
Historie
Studium
Aktuální školní rok
Přijímací zkoušky
Maturity
Lidé
Předměty
Google Apps

Historie Gymnázia Duchcov

Všeobecná touha českých obyvatel po zřízení střední školy s českým vyučovacím jazykem byla naplněna až po první světové válce. Dne 6. října 1919 začalo české gymnázium prozatímně působit v části německého reálného gymnázia. Obrovský zájem o studium (české gymnázium v Duchcově bylo jediným pro teplický, duchcovský a bílinský okres) přiměl ředitelství ústavu otevřít několik tříd i mimo německé gymnázium. Toto a všeobecné přání získat vlastní objekt stálo u zrodu záměru postavit pro duchcovské české gymnázium účelnou a důstojnou budovu. Od přání z roku 1919 k realizaci uplynulo ještě dlouhých osm let. Novostavba byla předána do užívání v neděli 22. května 1927. V průběhu desetiletí se škola stala nejen významným vzdělávacím ústavem, ale i centrem českého menšinového života, řada vynikajících absolventů svědčí o úrovni výuky. Bylo vybudováno zázemí, vytvořeny tradice, škola se dostala do širokého povědomí.

Stejně jako celá naše země prožívala své slavné i méně slavné dny. Ty nejhorší snad za války, kdy byla výuka zcela přerušena. Z hlediska stavebního budova velmi utrpěla i v letech 1948 - 1953, v době, kdy bylo zrušeno reálné gymnázium a v objektu se střídaly různé druhy škol, které zde pro své provozní potřeby prováděly stavební úpravy.

http://www.duchcov.cz/gymnazium/

Home
History
Studied
Current school year
Admission
Maturity
People
Subjects
Google Apps

History Gymnasium Duchcov

General desire Czech population after the establishment of secondary schools and English language teaching has been filled up after the First World War. On 6 October 1919 became the Czech grammar school provisionally operate in the German real grammar school. Huge interest in the study (Czech grammar school in Duchcov was only for Teplice, duchcovský and bílinský District) led the Directorate of the Institute to open a few classes and outside the German grammar school. Such a universal desire to acquire its own building stood for the birth of intent to build a school for duchcovské Czech efficient and dignified building. From the desire of the year 1919 to be implemented yet expired long eight years. New was transferred to use on Sunday 22nd May 1927. Over the decades, the school became not only an important educational institution, but also the centre of minority life, a number of outstanding graduates shows the level of instruction. It was built facilities, established tradition, the school into a broad awareness.

Like our country prožívala its famous and less famous days. The worst, perhaps during the war, when it was completely interrupted instruction. From the perspective of building a building suffered in the years 1948 - 1953, when it was cancelled real school building to pieces and various types of schools, which are for their operational needs to carry out structural adjustments.

History Grammar Duchcov

General desire of the Czech population after the establishment of secondary schools in the Czech language of instruction was filled to the First World War. 6 October 1919 Czech Grammar School began to operate provisionally in the German grammar school. The enormous interest in the study (in Czech high school was only for Duchcov Teplice, Duchcovská and Bilina district) headquarters forced the institute to open a few classes outside of the German school. This and a general desire to get their own building was the genesis of the project to build a gymnasium for Duchcovská Czech efficient and dignified building. From the desire of the year 1919 to implement more long passed for eight years. New building was handed over for use on Sunday 22 May 1927. Over the decades, the school became not only an important educational institution, but also the center of the minority lives, many outstanding graduates shows the level of teaching. It was built facilities, created a tradition, the school turned into a broad awareness.

Like our country has endured its famous and less famous days. Perhaps the worst of the war, when teaching was completely interrupted. In terms of construction the building was seriously damaged in the years 1948 - 1953, when school was canceled and the real object is rotated different kinds of schools here for their operational needs to carry out construction work.

http://gymdux.sokolici.eu/index.php/informace/historie-koly
Translated by Google Translate, 3 August 2010.

History Gymnasium Duchcov

General desire Czech population after the establishment of secondary school teaching in Czech language was filled to the First World War. On 6 October 1919 began Czech grammar provisionally act in the German grammar school. The enormous interest in the study (Czech secondary school in Duchcově was only for Teplitz, and Duchcov Bílina district) headquarters induced Institute opened a few classes outside of German grammar. This and the general desire to obtain its object was the genesis of the project to build a high school for Duchcov Czech efficient and dignified building. From the desire of 1919 for implementation has passed even eight long years. The new building was handed over for use on Sunday 22 May 1927. Over the decades, the school has become not only a major educational institution, but also the center of the Czech minority, many outstanding graduates testifies to the level of instruction. It was built facilities created tradition, the school came into widespread awareness.

Like our entire country had its famous and less famous days. The worst perhaps during the war when teaching was completely interrupted. In terms of construction building was seriously damaged in the years 1948 - 1953, when it was abolished grammar school building and alternated different types of schools, which are for its operational needs in construction work.

http://gymdux.sokolici.eu/index.php/informace/historie-koly
Translated by Google Translate, 17 June 2013.

History Grammar School Duchcov

The general desire Czech population after the establishment of secondary school teaching in Czech language was filled to the First World War. On October 6, 1919 began Czech gymnasium provisionally act in the German grammar school. The huge interest in the study (Czech secondary school in Duchcově was only for Teplice, Bílina Duchcov and district) caused the Directorate of the Institute opened a few classes outside the German Gymnasium. This and a general desire to get your own object was the genesis of the project to build a high school for Duchcov Czech efficient and dignified building. From the desire of 1919 for implementation has passed even eight long years. The new building was handed over for use on Sunday 22 May 1927. During the decade, the school has not only become an important educational institution, but also the center of the minority, many outstanding graduates indicative of the level of education. It was built facilities, established tradition, the school came into widespread awareness.

Like our whole country had its famous and less famous days. The worst perhaps during the war when teaching was completely interrupted. In terms of construction and the building was seriously damaged in the years 1948 - 1953, when the grammar school was canceled in the building and alternated different types of schools, which are for its operational needs in construction work.

http://www.gspsd.cz/historie/historie-skoly
Translated by Google Translate, 26 May 2014.


to-speech system that—chinese text and converts it into
chinese language.
And then we've taken it out so up my own voice and we use
that to modulate the standard text-to-speech system so that it

Building blocks of CL systems 2

- Information extraction
  - Chunking (instead of parsing).
  - Template filling.
  - Named-entity recognition.

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 Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990."

Tie-up-1:	*Relation:*	Tie-up
	Entities:	Bridgestone Sports Co. a local concern a Japanese trading house
	Joint venture:	Bridgestone Sports Taiwan Co.
	Activity:	**Activity-1**
	Amount:	NT $ 20,000,000

Activity-1:	*Company:*	Bridgestone Sports Taiwan Co.
	Product:	golf clubs
	Start date:	January 1990

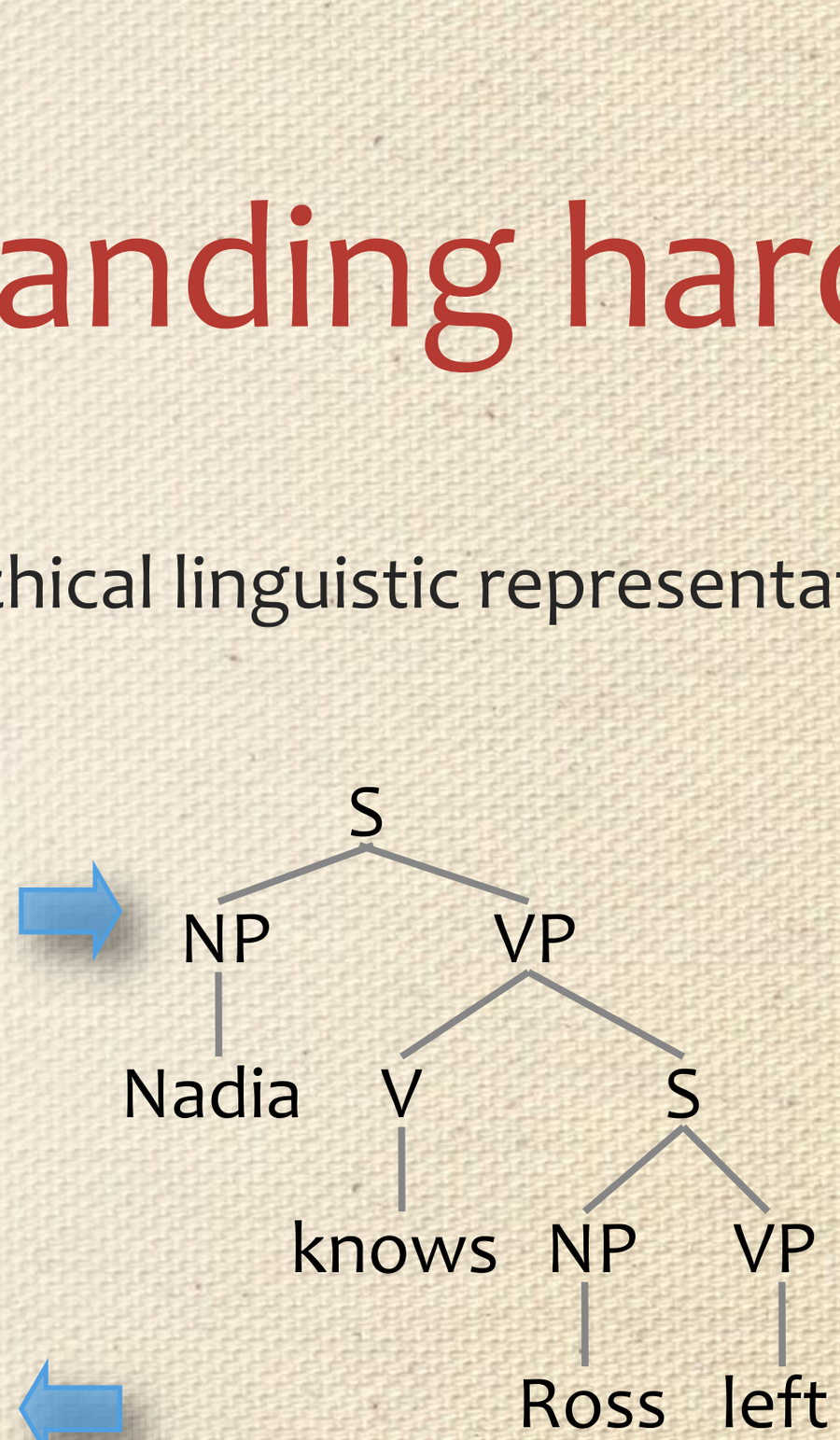
Building blocks of CL systems 3

- Lexical semantics
  - Word sense disambiguation (WSD).
  - Taxonomies of word senses.
  - Analysis of verbs and other predicates.
- Computational morphology

Why is understanding hard?

- Mapping of string of words to hierarchical linguistic representation.

Nadia knows Ross left.



KNOWS(Nadia, LEFT(Ross))

Why is understanding hard?

- Mapping from ***surface-form*** to meaning is many-to-one: Expressiveness.

Nadia kisses Ross.

Ross is kissed by Nadia.

KISS (Nadia, Ross)

Nadia gave Ross a kiss.

Nadia gave a kiss to Ross.

Why is understanding hard?

- Mapping is one-to-many: Ambiguity at all levels.
  - Lexical
  - Syntactic
  - Semantic
  - Pragmatic

Lexical ambiguity

The lawyer walked to the bar and addressed the jury.

The lawyer walked to the bar and ordered a beer.

You held your breath and the door for me. (Alanis Morissette)

Earl of Sandwich: *You will die either of the pox or on the gallows.*

John Wilkes: *That will depend on whether I embrace your mistress or your principles.*

"zeugma"

- Computational issues
  - Representing the possible **meanings** of words, and their **frequencies** and their **indications**.
  - Representing **semantic relations** between words.
  - Maintaining adequate **context**.

used to strain microscopic plant life from the
zonal distribution of plant life .
close-up studies of plant life and natural
too rapid growth of aquatic plant life in water
the proliferation of plant and animal life
establishment phase of the plant virus life cycle
that divide life into plant and animal kingdom
many dangers to plant and animal life
mammals . Animal and plant life are delicately

automated manufacturing plant in Fremont
vast manufacturing plant and distribution
chemical manufacturing plant , producing viscose
keep a manufacturing plant profitable without
computer manufacturing plant and adjacent
discovered at a St. Louis plant manufacturing
copper manufacturing plant found that they
copper wire manufacturing plant , for example
‘s cement manufacturing plant in Alpena

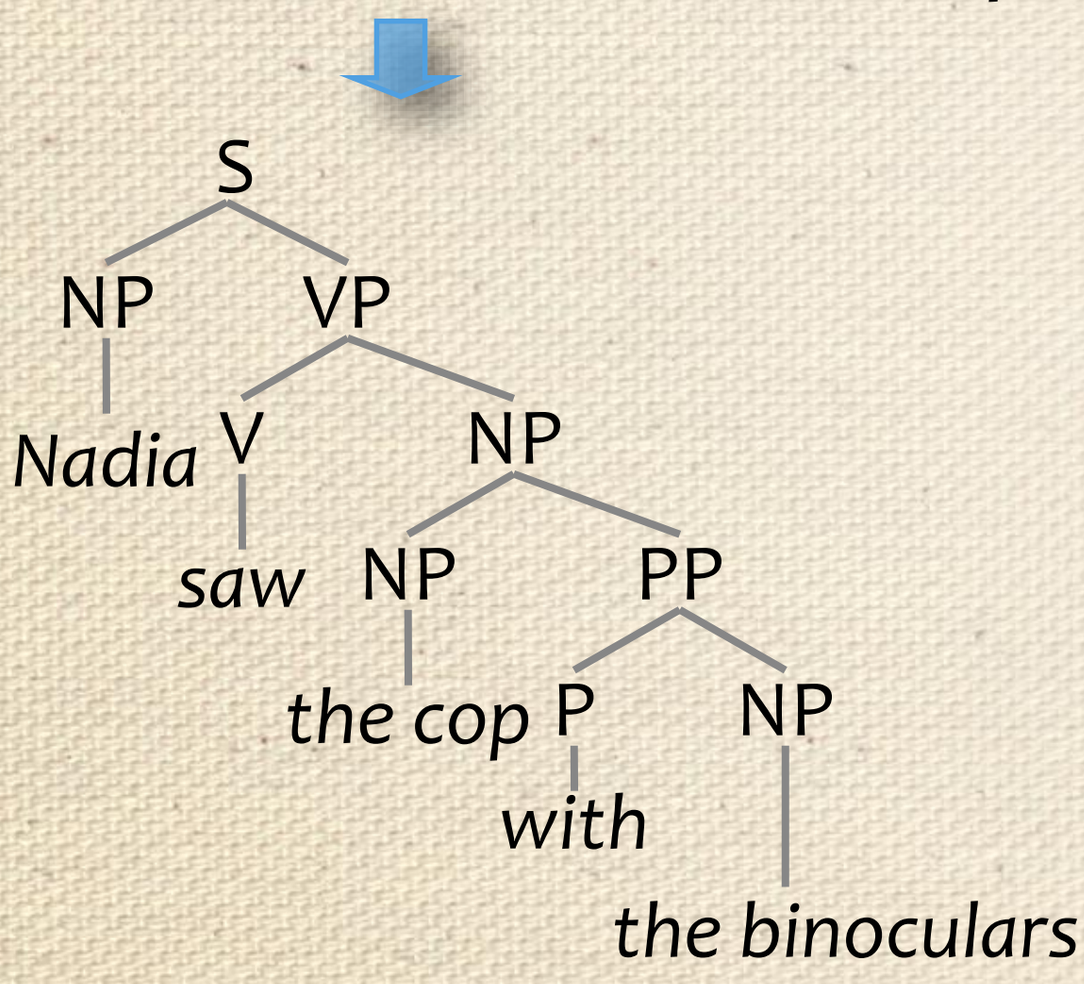
vinyl chloride monomer plant , which is
molecules found in plant and animal tissue
Nissan car and truck plant in Japan is
and Golgi apparatus of plant and animal cells
union responses to plant closures .
cell types found in the plant kingdom are
company said the plant is still operating
Although thousands of plant and animal species
animal rather than plant tissues can be

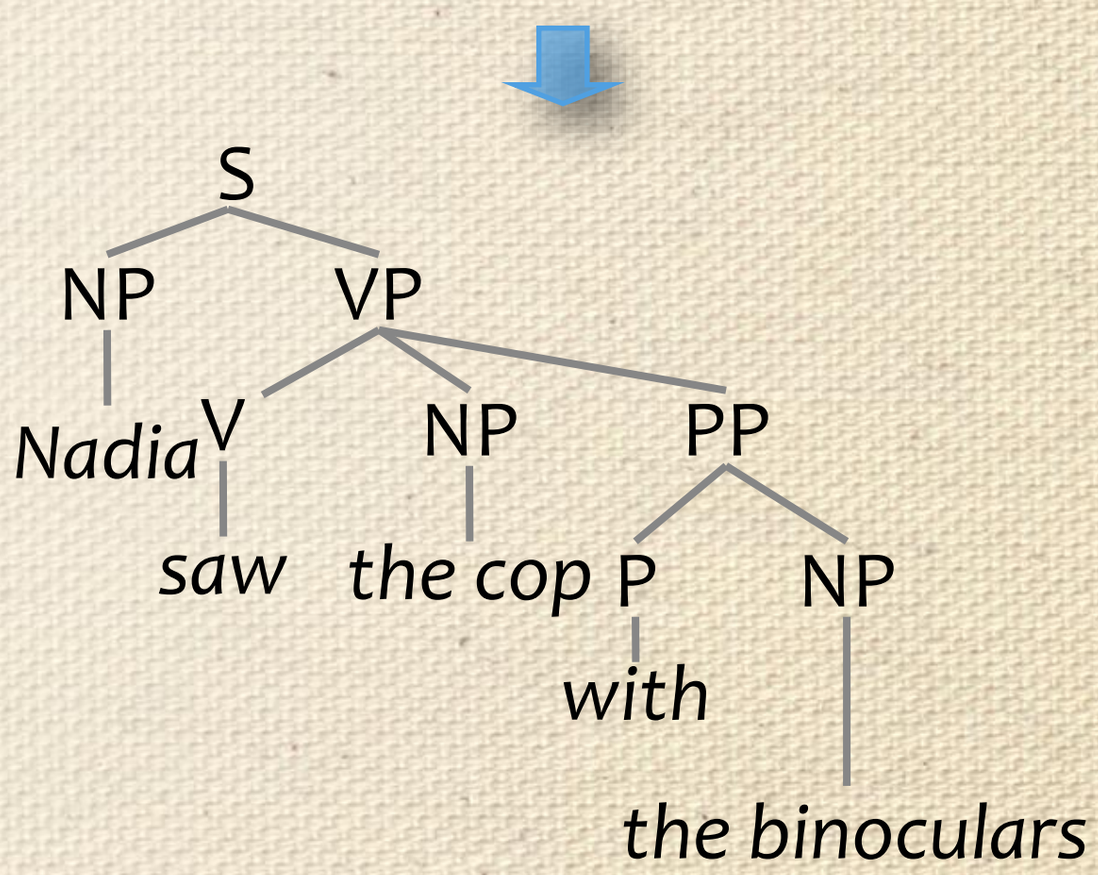
Decision for *plant*

LogL	Collocation	Sense
8.10	*plant life*	→ A
7.58	*manufacturing plant*	→ B
7.39	*life* (within ±2-10 words)	→ A
7.20	*manufacturing* (in ±2-10 words)	→ B
6.27	*animal* (within ±2-10 words)	→ A
4.70	*equipment* (within ±2-10 words)	→ B
4.39	*employee* (within ±2-10 words)	→ B
4.30	*assembly plant*	→ B
4.10	*plant closure*	→ B
3.52	*plant species*	→ A
3.48	*automate* (within ±2-10 words)	→ B
3.45	*microscopic plant*	→ A
	…	

Syntactic ambiguity

Nadia saw the cop with the binoculars.





Syntactic ambiguity 2

Put the book in the box on the table.

Put the book in the red book box.

Visiting relatives can be trying.

Noun phrase

Adj Noun

Verb Noun

Verb phrase

Syntactic ambiguity 3

- Most syntactic ambiguity is **local** — resolved by syntactic or semantic context.

  Visiting relatives is trying.
  Visiting relatives are trying.
  Nadia saw the cop with the gun.

- Sometimes, resolution comes too fast!

  [ ][ ][[????
  The cotton clothing is made from comes from Mississippi.
  [[ ][ ]][[ ]]

  "Garden-path" sentences.

Syntactic ambiguity 4

- Computational issues
  - Representing the possible **combinatorial structure** of words.
  - Capturing syntactic **preferences** and frequencies.
  - Devising **incremental parsing** algorithms.

Semantic ambiguity

- Sentences can have more than one meaning, *even when* the words and structure are agreed on.

Nadia wants a dog like Ross's.

Everyone here speaks two languages.

Pragmatic ambiguity

- A sample dialogue
  - Nadia: *Do you know who's going to the party?*
  - Emily: *Who?*
  - Nadia: *I don't know.*
  - Emily: *Oh … I think Carol and Amy will be there.*
- Computational issues
  - Representing **intentions** and **beliefs**.
  - **Planning** and **plan recognition**.
  - **Inferencing** and diagnosis.

Need for domain knowledge 1

Derivatization of the carboxyl function of retinoic acid by fluorescent or electroactive reagents prior to liquid chromatography was studied. Ferrocenylethylamine was synthesized and could be coupled to retinoic acid. The coupling reaction involved activation by diphenylphosphinyl chloride. The reaction was carried out at ambient temperature in 50 min with a yield of ca. 95%. The derivative can be detected by coulometric reduction (+100 mV) after on-line coulometric oxidation (+400 mV). The limit of detection was 1 pmol of derivative on-column, injected in a volume of 10µl, but the limit of quantification was 10 pmol of retinoic acid.

S. El Mansouri, M. Tod, M. Leclercq, M. Porthault, J. Chalom, "Precolumn derivatization of retinoic acid for liquid chromatography with fluorescence and coulometric detection." *Analytica Chimica Acta*, 293(3), 29 July 1994, 245–250.

Need for domain knowledge 2

In doing sociology, lay and professional, every reference to the "real world", even where the reference is to physical or biological events, is a reference to the organized activities of everyday life. Thereby, in contrast to certain versions of Durkheim that teach that the objective reality of social facts is sociology's fundamental principle, the lesson is taken instead, and used as a study policy, that the objective reality of social facts *as* an ongoing accomplishment of the concerted activities of daily life, with the ordinary, artful ways of that accomplishment being by members known, used, and taken for granted is, for members doing sociology, a fundamental phenomenon.

Harold Garfinkel, Preface, *Studies in Ethnomethodology*, Prentice-Hall, 1967, page vii.

Levels of linguistic structure and analysis 1

- Phonology
  - The **sound system** of a language.
- Morphology
  - The **minimal meaningful pieces** of language (root of a word; suffixes and prefixes), and how they combine.
- Lexicon
  - The semantic and syntactic **properties of words.**

Levels of linguistic structure and analysis 2

- Syntax
  - The structure of a sentence: **how words can combine**, and the relation to meaning.
- Semantics
  - The **meaning** of a sentence (a logic statement).
- Pragmatics
  - The **use** of a sentence: pronoun referents; intentions; multi-sentence structure.

Focus of this course 1

- Grammars and parsing.
- Resolving syntactic ambiguities.
- Determining semantic relationships.
- Lexical semantics, resolving word-sense ambiguities.
- Understanding pronouns.

Focus of this course 2

- Current methodologies
  - Integrating statistical knowledge into grammars and parsing algorithms.
  - Using text corpora as sources of linguistic knowledge.

Not included

- Machine-learning, data-intensive methods *§
  - Statistical models, text classification, …
- Machine translation *
- Speech recognition and synthesis *¶
- Cognitive science–based methods
- Understanding dialogues and conversations

* See CSC 401 / 2511. ¶ See CSC 2518. § See CSC 2540.