

Computational Linguistics

CSC 485/2501

1

1. Introduction to computational linguistics

Gerald Penn

Department of Computer Science, University of Toronto
(many slides taken or adapted from others)

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

Copyright © 2021 Graeme
Hirst, Suzanne Stevenson
and Gerald Penn. All rights
reserved.

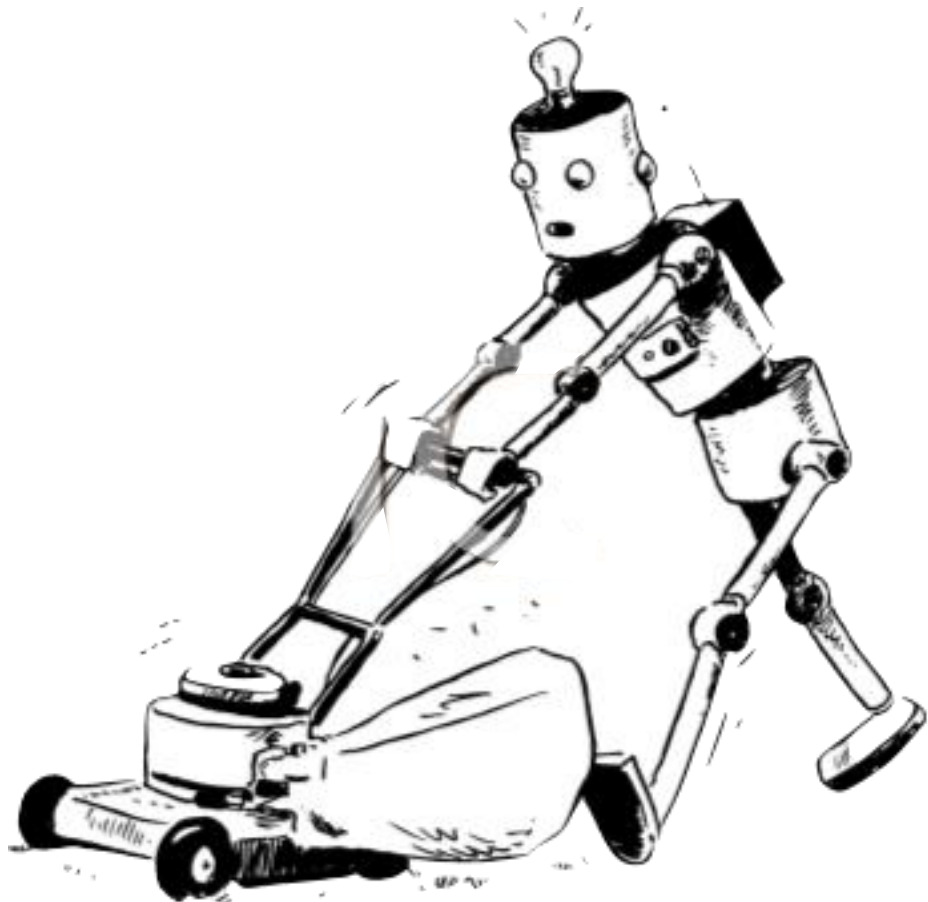
Why would a computer need
to use natural language?

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.




I lost my card last Friday and now there's this \$87 charge that I don't recognize?
Could you help me with that?

Intent_1: Lost_card

Value_date: 07-30-2021

Intent_2: Dispute_charge

Value_amount: \$87.00



Okay, I can help you with that.
I just need to confirm your details.

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



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

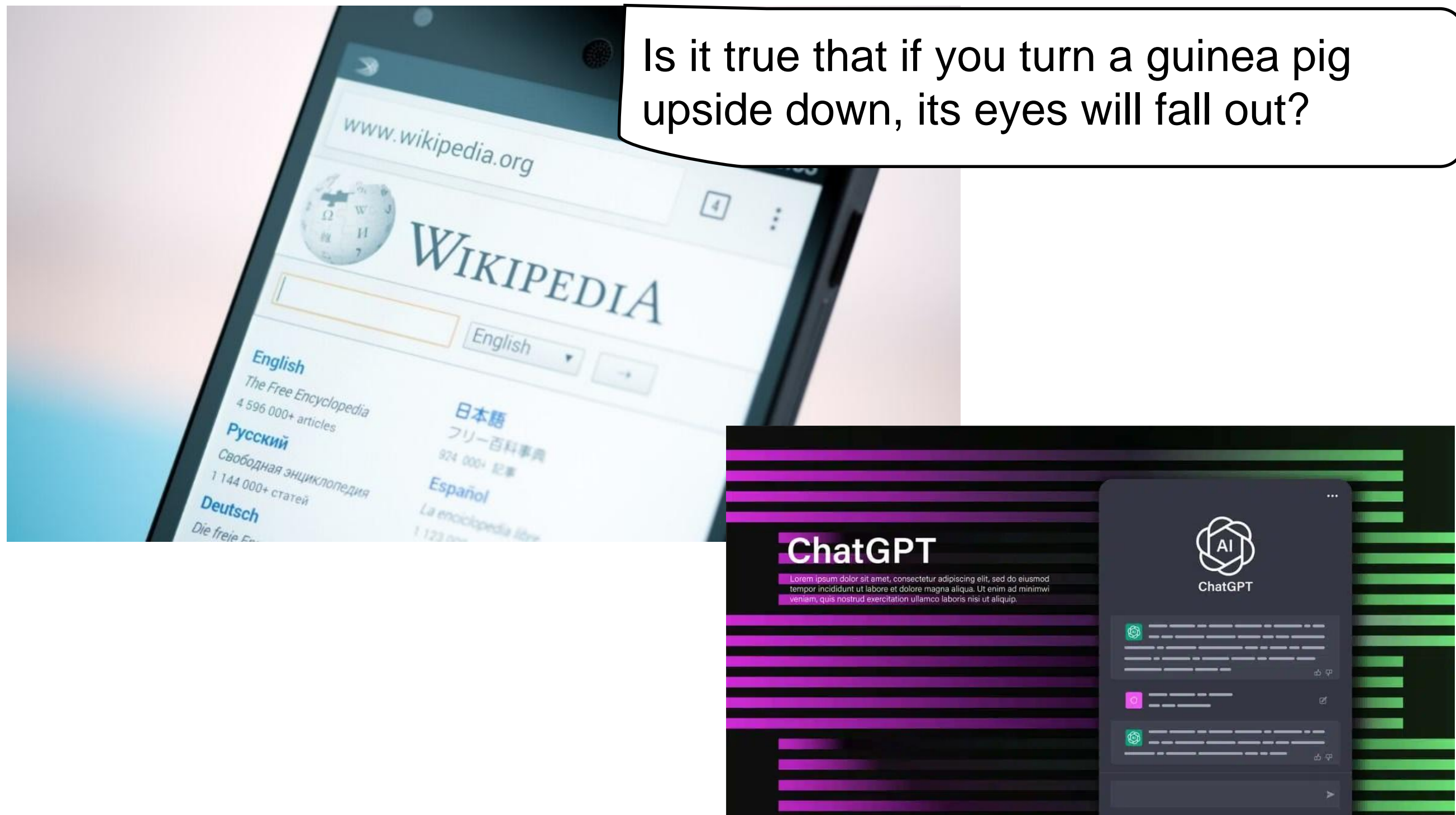


- Computer as researcher.
Brings us the information we need.



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

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



- Computer as researcher.
Organizes the information we need.



Please write a 500-word essay for me on “Why trees are important to our environment”.

And also write a thank-you note to my grandma for the birthday present.

- Computer as researcher.
Wins television game shows.



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

<https://www.youtube.com/watch?v=yJptrICVDHI>

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

- Computer as language expert.
Translates our communications.

est important que tous les députés à la Cha-
oute la population comprennent pourquoi nous
ous intéressons à ce secteur de l'économie qui
onstituent les jeux de hasard. L'industrie des j
aris a littéralement explosé récemment, non
eulement parce que les gens aiment parier et
rofitier des diverses possibilités du jeu, mais a
arce que, dans le cadre de l'économie mondial
ecteur touristique prend de plus en plus d'amp
our bon nombre de pays, le tourisme est le fa
ui assure la viabilité de leur économie. Au cou
es quatre ou cinq dernières années, des déput
Chambre des communes ont, en manifestant
appui, encouragé le gouvernement à quadruple
udget publicitaire de Tourisme Canada. Ils
omprennent que c'est dans l'intérêt public
un grand nombre d'emplois sont tribu

is important that we in the House and in t
country understand why we are becoming inte
n this whole area of gaming. The gaming indu
exploding in the world and not just because pe
now enjoy gaming and the diverse opportuniti
he gaming realm. It is also because the touris
ector of the global economy is growing. For r
ountries tourism is the thing that is actually
keeping their economies viable. In the last fou
ive years members of the House of Commons
hrough their support have encouraged this
government to quadruple the advertising budg
ourism Canada. They understand from a publ



- **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 and authorship analysis
 - Helping language learners
 - Helping people with disabilities
 - ...

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
Alzheimer's



P.D. James

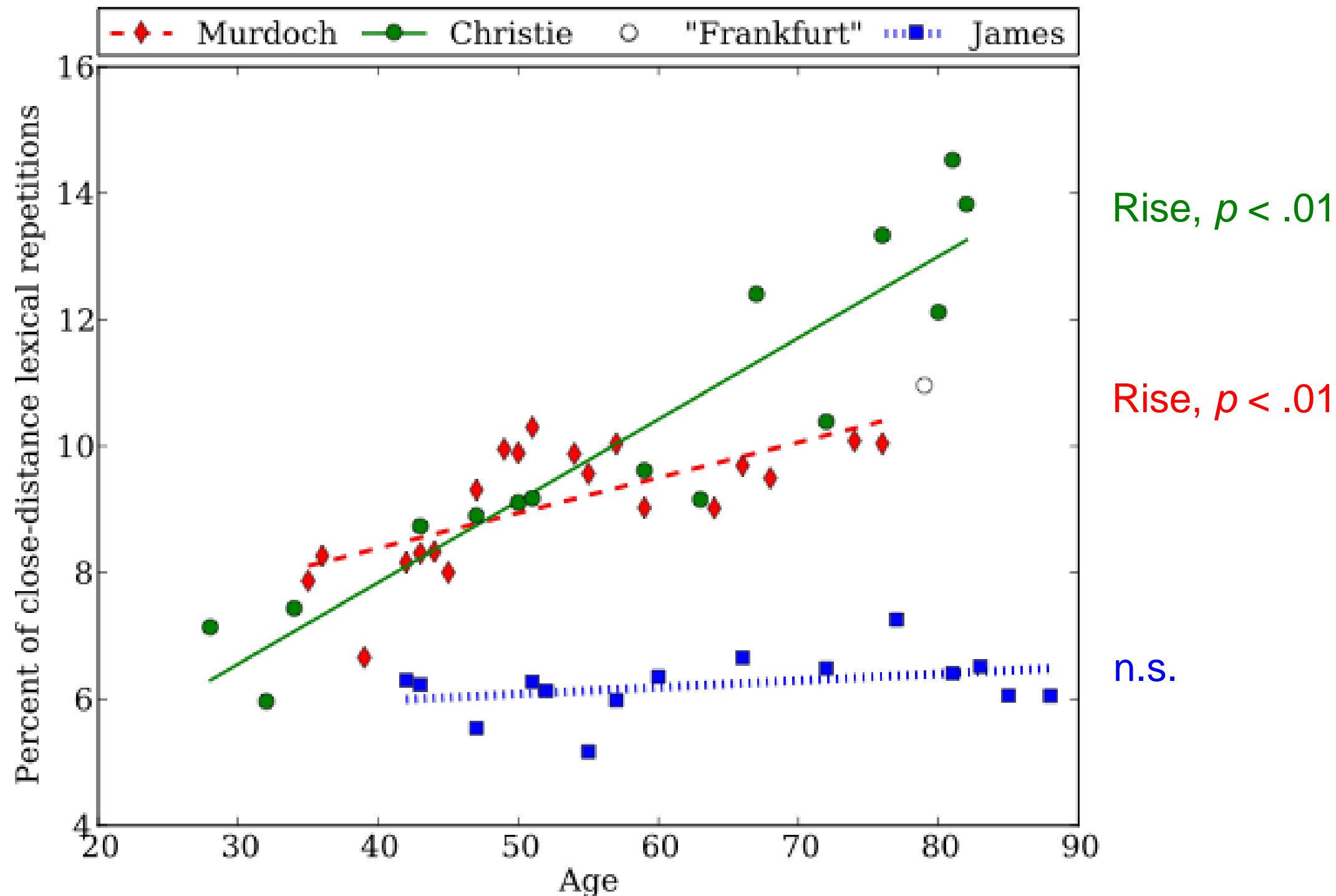
No Alzheimer's



Agatha Christie

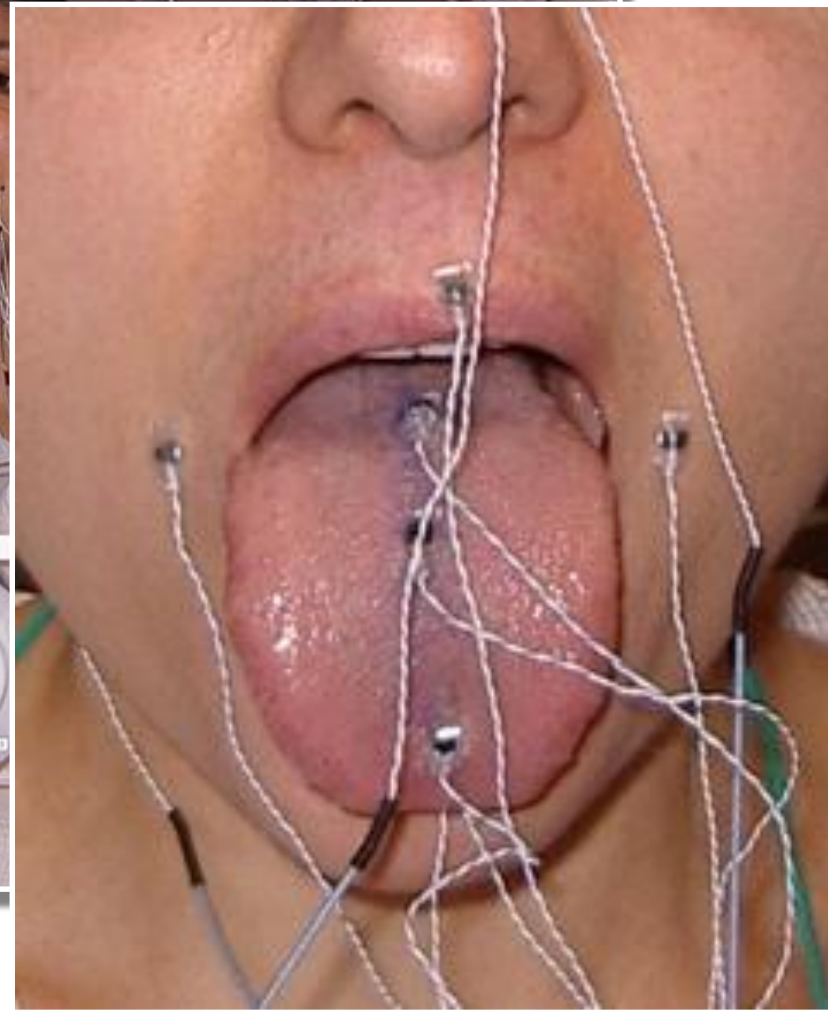
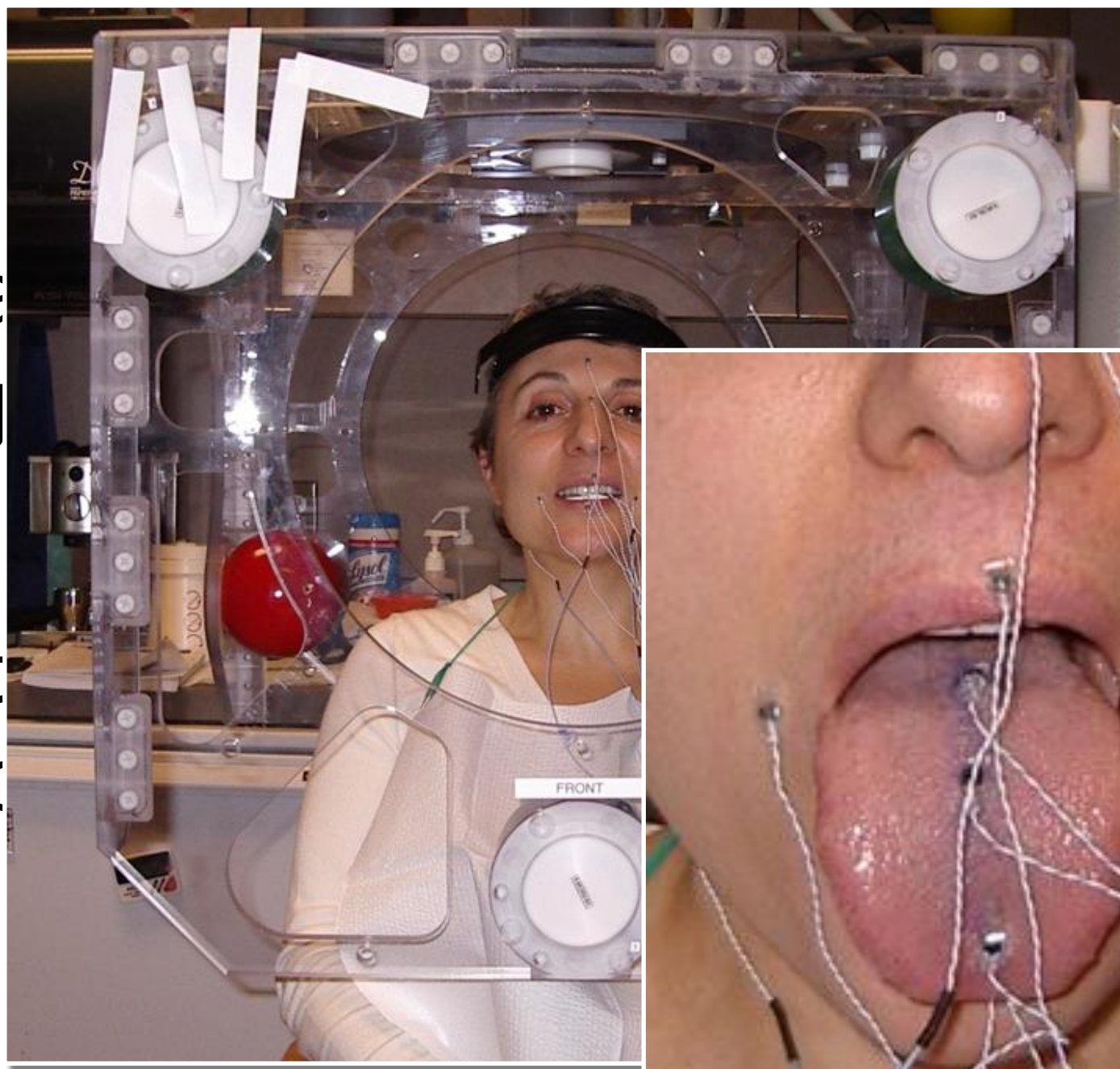
Suspected

Increase in short-distance word repetition



Speech recognition for

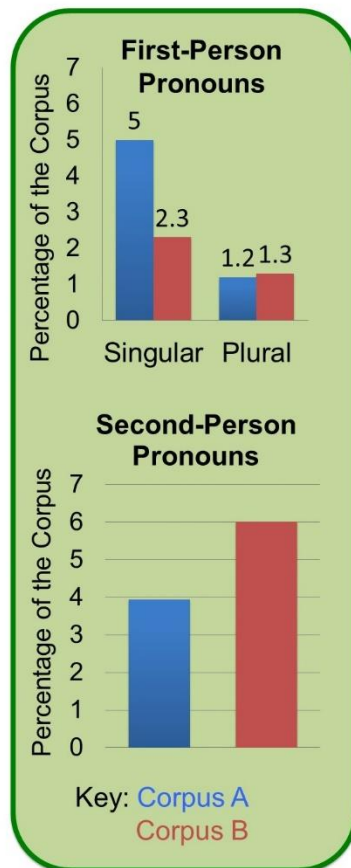
- Use a speech recognition device
- Create a speech recognition system



Language Change through Time

Results:

Pronouns:



Conclusion

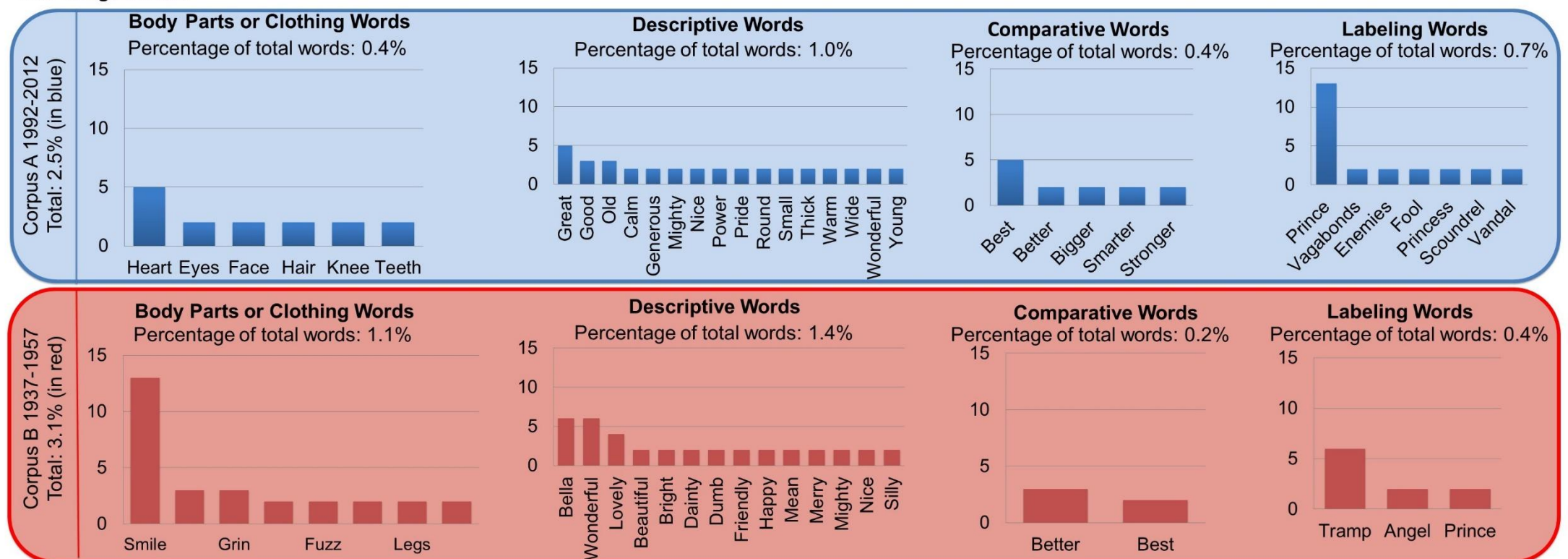
S:Pronoun Usage

- Over time, content has become more focused on oneself than on others.
- The trend in pronoun use supports the theory that cultural products match cultural changes. In this study, the song lyrics, a cultural product, have become more focused on self as culture becomes more individualistic.

Self-Image Words

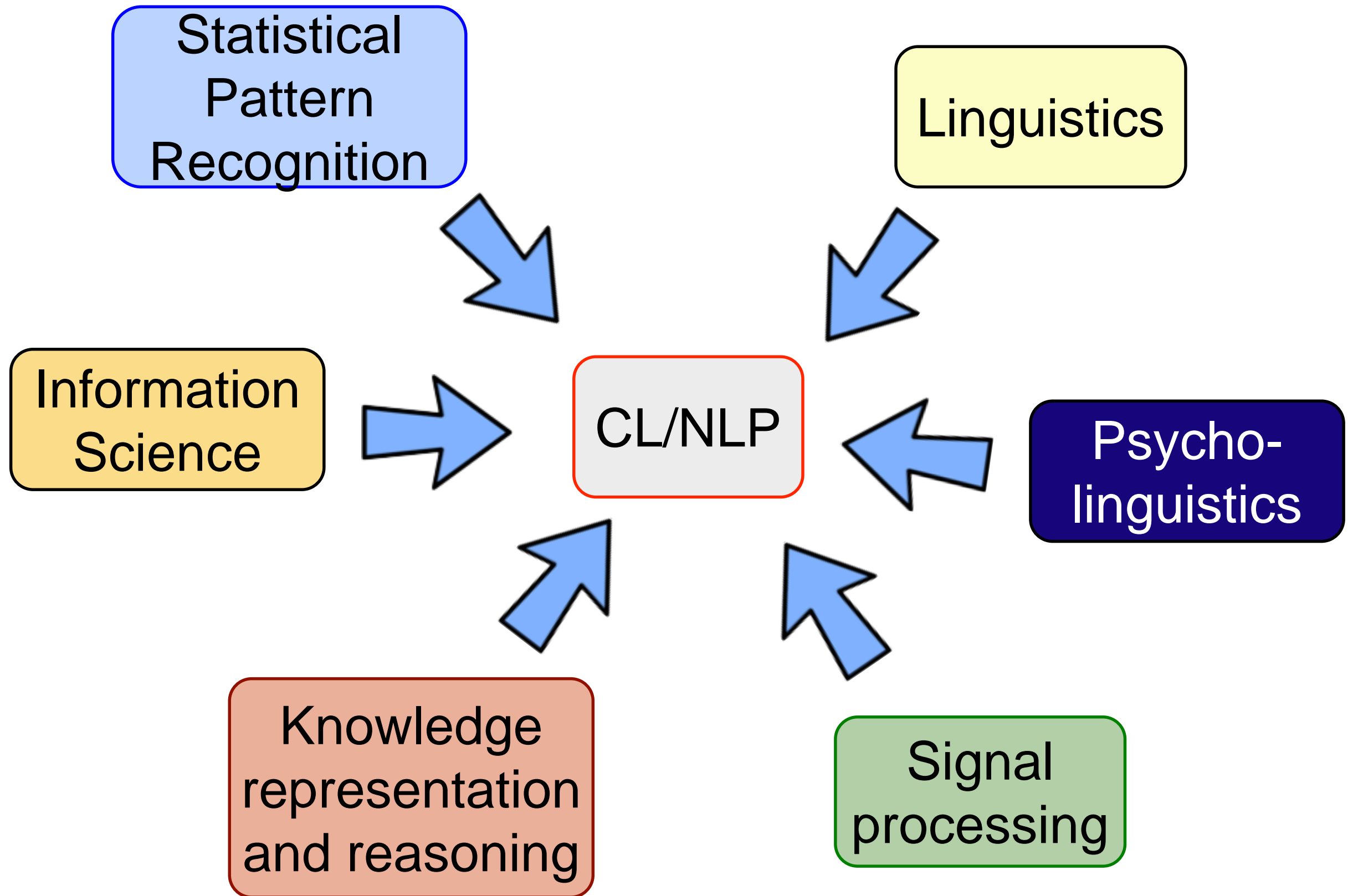
- Self-image has not been a prominent theme in Disney lyrics.
- The low frequency of self-image words in the results show that the messages are not conveyed through song lyrics. However, self-image messages may be conveyed through some other aspect of the movies.

Self-Image Words:



Mathematics of syntax and language

- Corlett complexity (2021): analyses parsing complexity as a function of language model entropy
- Fowler's algorithm (2009): first quasi-polynomial time algorithm for parsing with Lambek categorial grammars
- McDonald's algorithm (2005): novel dependency-grammar parsing algorithm based upon minimum spanning trees



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.

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.
- Embedding structure into known problem spaces (especially with neural networks).
 - *Combines:* statistical pattern recognition *and* some relatively simple linguistic knowledge.

Focus of this course 1

- “Grammars”
- “Parsing”
- “Language Models”
- Resolving ambiguity
- Determining “argument structure”
- Lexical semantics, word sense
- “Compositional” semantics
- Question Answering
- Understanding pronoun reference

Focus of this course 2

- Current methods
 - Integrating statistical knowledge into grammars and parsing algorithms.
 - Using text corpora as sources of linguistic knowledge.
- What about deep learning?
 - There will be plenty of that.
 - But keep in mind that much of it proceeds from some assumptions about representational adequacy that haven't been born out empirically.
 - This class will teach you how to use the terminology and do the evaluations so that you don't make the same mistakes.

Not included

- Machine translation, text classification, information retrieval...*
- Graph-theoretic and spectral methods%
- Speech recognition and synthesis*¶
- Cognitively based methods\$^
- Semantic inference,% semantic change/drift^
- Understanding dialogues and conversations¶
- Bias, fake news detection, ethics in NLP\$

* CSC 401 / 2511. % CSC 2517. ¶ CSC 2518. § CSC 2540. ^ CSC 2611. \$CSC 2528.

Practice Quiz

- Which of the following language technologies was traditionally considered to have been the “holy grail” of computational linguistics?
 - a) Machine translation systems
 - b) Question answering systems
 - c) Dialogue systems
 - d) Parsers