COMPUTATIONAL LINGUISTICS

Athanasios N. Karasimos

akarasimos@gmail.com

BA in Linguistics | National and Kapodistrian University of Athens

Lecture 1 | Wed 28 Feb 2018

LECTURE I:
COURSE AND HISTORICAL OVERVIEW

- Introduction | General Information | Outline
- Course Description | Course Overview
- Defining Computational (and Corpus) Linguistics
- Historical Overview of Computational Linguistics
- Computational Linguistics Applications and Tools: a quick-shot

INTRODUCTION

CONTACT INFO

- **Athanasios N. Karasimos**
  - *Researcher @PARTHENOS-EU Consortium Project*
  - akarasimos@gmail.com | akarasimos@academyofathens.gr | akarasimos@phil.uoa.gr
    - Email subject: ΕΚΠΑ ΥΠΓΛ | Τίτλος μηνύματος
  - https://www.linkedin.com/in/athanasios-karasimos-aa482a3b/
  - https://gamelinguist.wordpress.com/
  - 'Office' Hours: Wednesday 11:00-12.00 (office 903) or by (phone) appointment.

NETWORKING

- ‘Study’ Groups
  - Course Contact List (exclusion is an option)
- Group assignment/ project
- LinkedIn profiles
- LinguistList register
- A friend called Google

COURSE OVERVIEW

COURSE DESCRIPTION

The course "Computational Linguistics and Corpora" centers on the basic and fundamental concepts of this interdisciplinary area of Linguistics, Informatics and Cognitive Science.

Our purpose is to cover a wide range of theoretical and technical issues from Speech Recognition and Synthesis to Natural Language Processing and Machine Translation. Significant topics from Semantics, Syntax, Morphology and Phonology will be introduced through the dynamic prism of several state-of-the-art computing tools, applications, models and theories.

This course will use a methodology of empirical linguistic analysis and processing of natural language that includes regular expressions, language modeling, machine learning, morphological/syntactical parsing, corpus analysis and annotation and semantic analysis and representation.

COURSE DESCRIPTION

In particular, the main aim of the course is to familiarize students with significant and on-going research questions and theoretical approaches in this field and to provide them access to various tools and applications, while at the same time introducing them to language coding through programming.

Moreover, we will also focus on how linguistic theory is applied to the most up-to-date text processing techniques, word meaning and semantic interpretations.

Theoretical and technical issues such as n-grams models, neighborhood density, Context-free Grammars, morphosyntactic tagging, vector semantics, computing with word senses will be supported by exercises and mini-projects that will enable students to use practice tools, corpora and apply various semantic algorithms.

COURSE OBJECTIVES

- *understand* basic concepts of Computational Linguistics.
- *follow* the current trends of an ever-evolving scientific area.
- *recognize* mainstream linguistic theories in a more technical environment.
- (computationally) *analyze* the English and Greek language on different levels.
- *acquire* theoretical and computational skills in language processing.
- *interpret* various phenomena by approaching them computationally.
- *get* stimuli and motives for further studying this area.
- “*get*" their hands dirty by trying some basic and preliminary programming.

COURSE MATERIAL

- Notes and on-line available articles
- Presentations material
- Multimedia material
- Interactive quizzes and exercises
- Computational tools and applications
- Python programming tutorial (mock lab) – DON’T be scared!

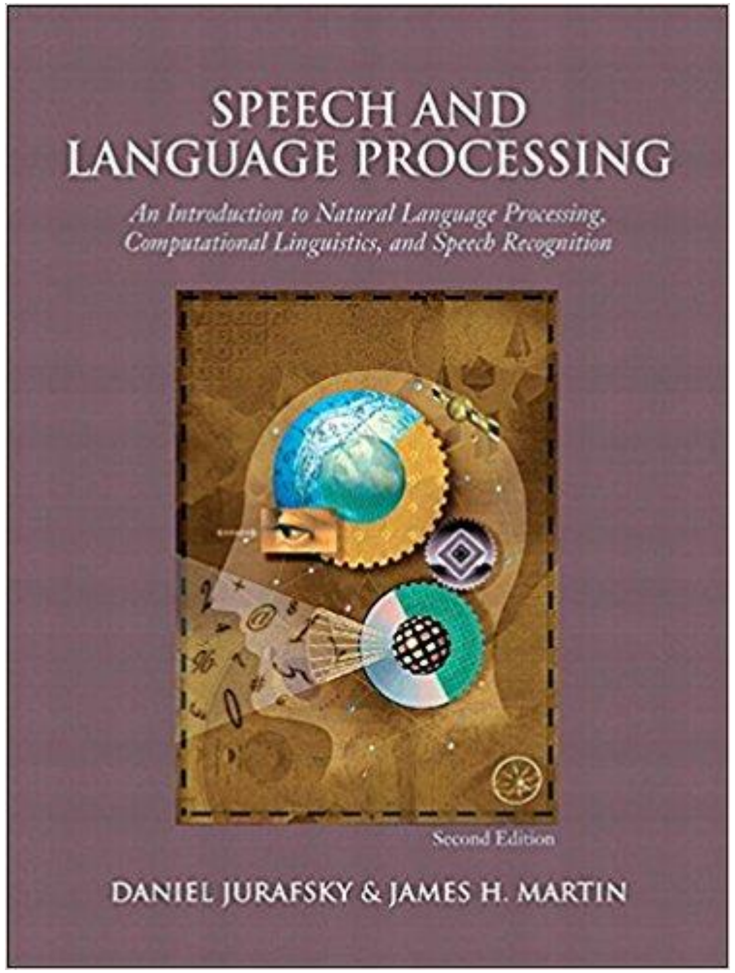
COURSEBOOK

SPEECH and LANGUAGE PROCESSING
An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (2nd Edition + 3rd Edition)

By Daniel Jurafsky & James H. Martin

- http://www.cs.colorado.edu/~martin/slp.html
- https://web.stanford.edu/~jurafsky/slp3/

The 'Bible' of the Field; it provides a holistic overview, analysis, explanations of techniques, theories, algorithms, etc.


SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

BIBLIOGRAPHY

- Roark B. & Sproat R. (2007). *Computational Approaches to Morphology and Syntax*. Oxford: Oxford University Press.
- Manning & Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Bird S., Klein E. & Loper E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O' Reilly Media.
- Καρασίμος, Α. (2011). *Υπολογιστική Επεξεργασία της Αλλομορφίας στην Παραγωγή Λέξεων της Νέας Ελληνικής*. Διδακτορική Διατριβή. σσ. 305. Πανεπιστήμιο Πατρών: Τμήμα Φιλολογίας. DOI: 10.13140/RG.2.1.1570.7926
- Additional books, articles, tools and applications via Zotero.

GRADES & EVALUATION

- (1.) 10% Participation {bonus}
- (2a.) 30% Assignments (multiple-choices, quizzes, article reviewing)
- (2b.) 30% Short Essay/ Mini-project OR ???
- (3.) 70% Exams
- Homework/ Assignments (Homework due every 5 weeks (2 total))
- Final Project/ Short Essay (due Week 13 (last class))
- Final Exam (Week 14 (almost nothing new))

CLARIFICATIONS & PLAGIARISM

- Late or missing Assignment/ Exercises/ Reading
  - You fall behind and have trouble keeping up, leading to lower mark on final exam and final project.
  - Absolution (no penalty)
    - You can skip or miss any assignment, exercise or even the project.
    - You cannot miss the exam (???)
  - Late submission (grade penalty)
    - Only if there is a serious reason.
  - Plagiarism and Collaboration
    - You may and you should discuss work with anyone, but your work should be your own (except the group projects).
    - PLAGIARISM is an act of fraud. It involves both stealing someone else's work and lying about it afterward. AVOID IT!

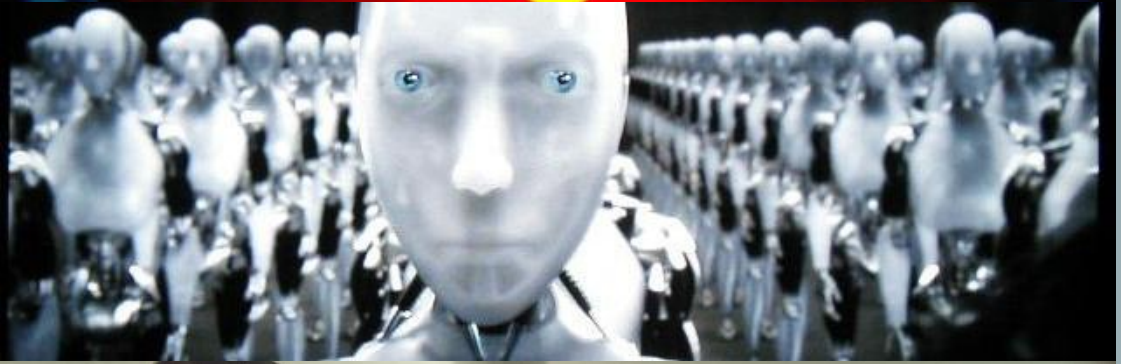
COURSE OUTLINE

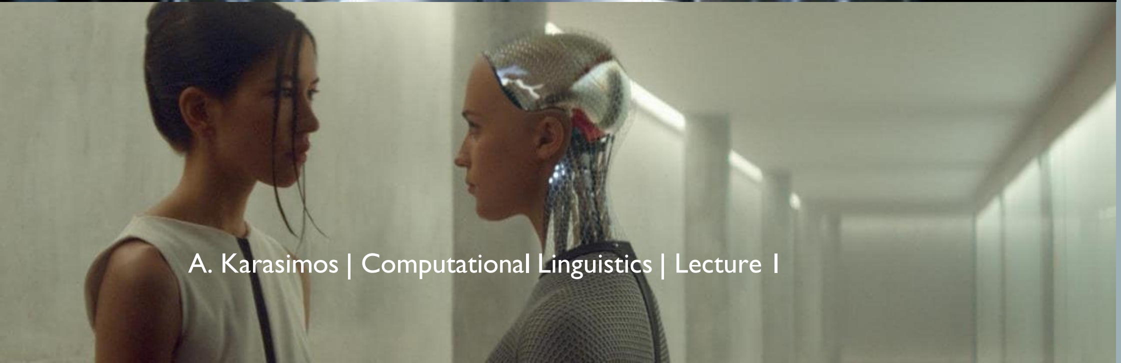
Week	Title
1.	Course and Historical Overview
2.	Regular Expressions
3.	Regular Languages & Finite-State Automata
4.	Language Modelling with N-Grams
5.	Tagging and Hidden Markov Models
6.	Parsing
7.	Logic
8.	Machine Learning
9.	Corpus Linguistics
10.	QA Systems & Dialogue Systems
11.	Machine Translation
12.	Speech Analysis and Synthesis
13.	Revision, Recap and Presentations

A e-mail about Course Overview and Syllabus Outline

THE ENTERTAINMENT WORLD

A Computational Linguistics Movie

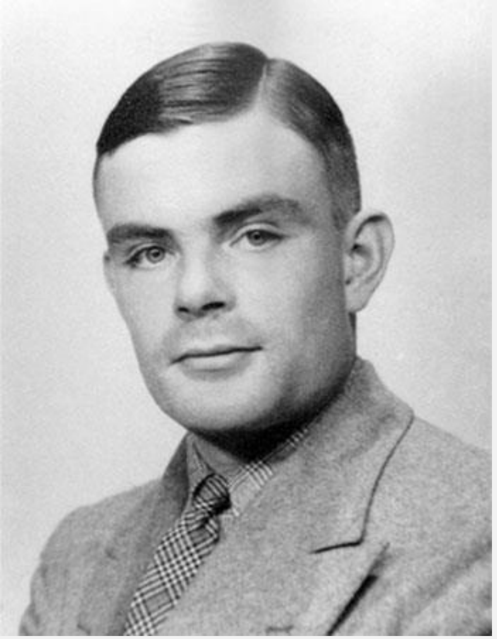




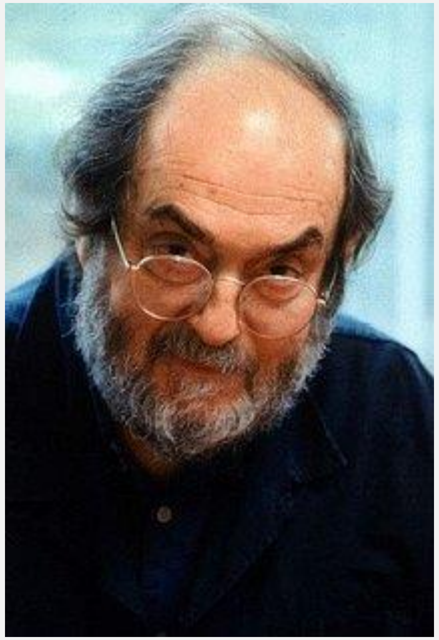
THE ART BEFORE THE SCIENCE

Lets guess:

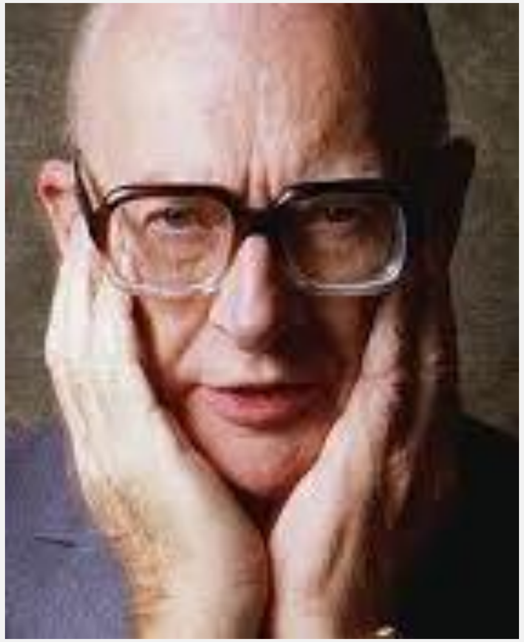
THE FANTASTIC 4: THE PIONEER, THE FILMAKER, THE FUTURIST & THE PRIEST



Alan Turing



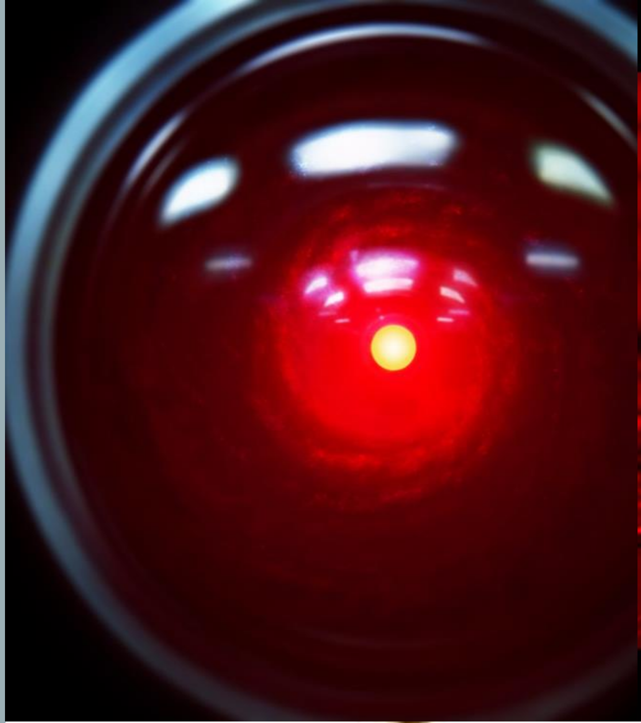
Stanley Kubrick

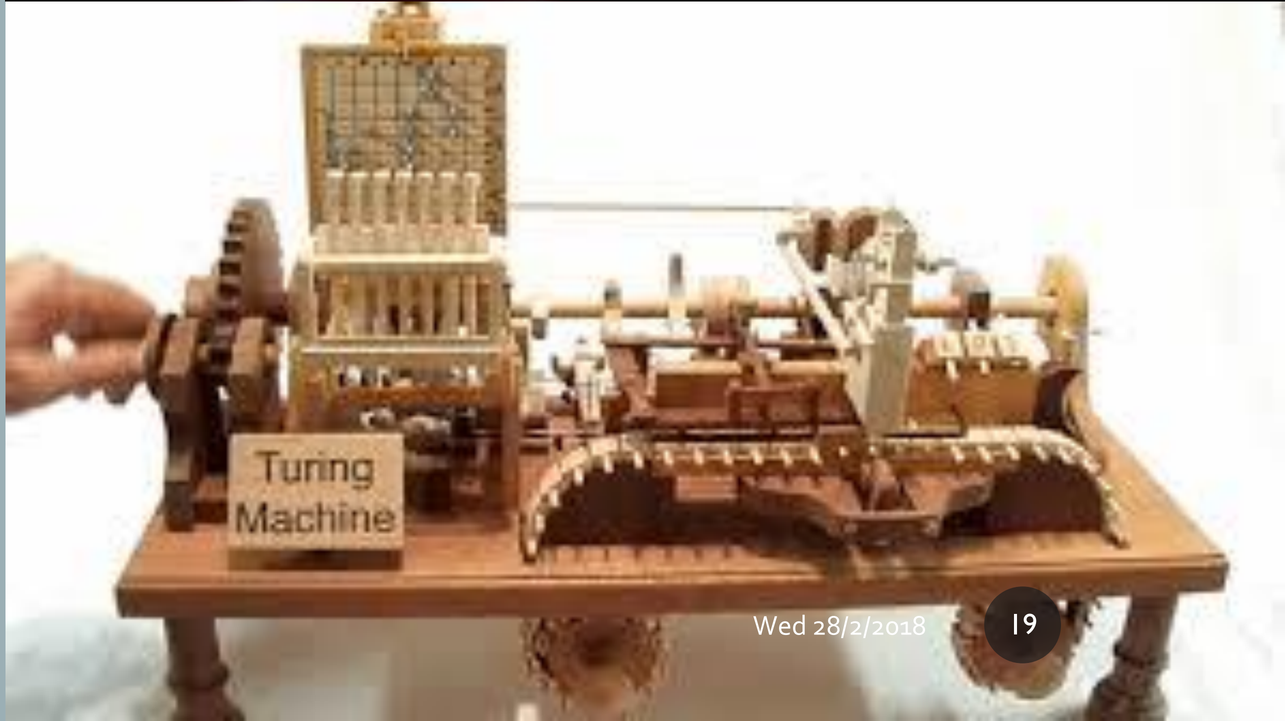


Arthur Clarke

Robert Busa

TURING MACHINE & HAL 9000





HAL 9000, A COMPLING ‘COMPANION’

HAL 9000’S CAPABILITIES

- Display graphics
- Play chess
- *Natural language production and understanding*
- Vision
- Planning
- Learning

HAL 9000 SPEAKING

David Bowman:

Open the pod bay doors, Hal.

HAL:

I’m sorry, Dave, I’m afraid I can’t do that.

David Bowman:

What are you talking about, Hal?

HAL:

I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

INTRODUCTION TO COMPUTATIONAL LINGUISTICS

The Definition Mode

COMPUTATIONAL LINGUISTICS I

- (a.k.a) Natural Language Processing (NLP), Language Engineering, Language Technology, […]
- Definition: *Study of how to solve problems computationally involving the interpretation and generation of human language text and speech.*
- Properties
  - As with applied science: the proof is in the pudding
  - Sometimes at odds with theoretical linguistics
    - No need of model human abilities and human methods
    - No need of correspond to published linguistic theories
    - But sometimes draws on one or both
  - Interdisciplinarity: linguists, computer scientists, experts in artificial intelligence, mathematicians, logicians, philosophers, cognitive scientists, cognitive psychologists, psycholinguists, anthropologists and neuroscientists, among others.

COMPUTATIONAL LINGUISTICS II

Natural Language Processing (NLP) studies how to get computers to do useful things with natural languages:

- Most commonly Natural Language Understanding
- The complementary task is Natural Language Generation

NLP draws on research in **Linguistics**, **Theoretical Computer Science**, Artificial Intelligence, Mathematics and Statistics, Psychology, **Cognitive Science**, etc.

COMPUTATIONAL LINGUISTICS III

Computational linguistics (CL) is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective, as well as the study of appropriate computational approaches to linguistic questions.

Traditionally, computational linguistics was performed by computer scientists who had specialized in the application of computers to the processing of a natural language.

DREAMS AND REALITY OF COMPUTATIONAL LINGUISTICS

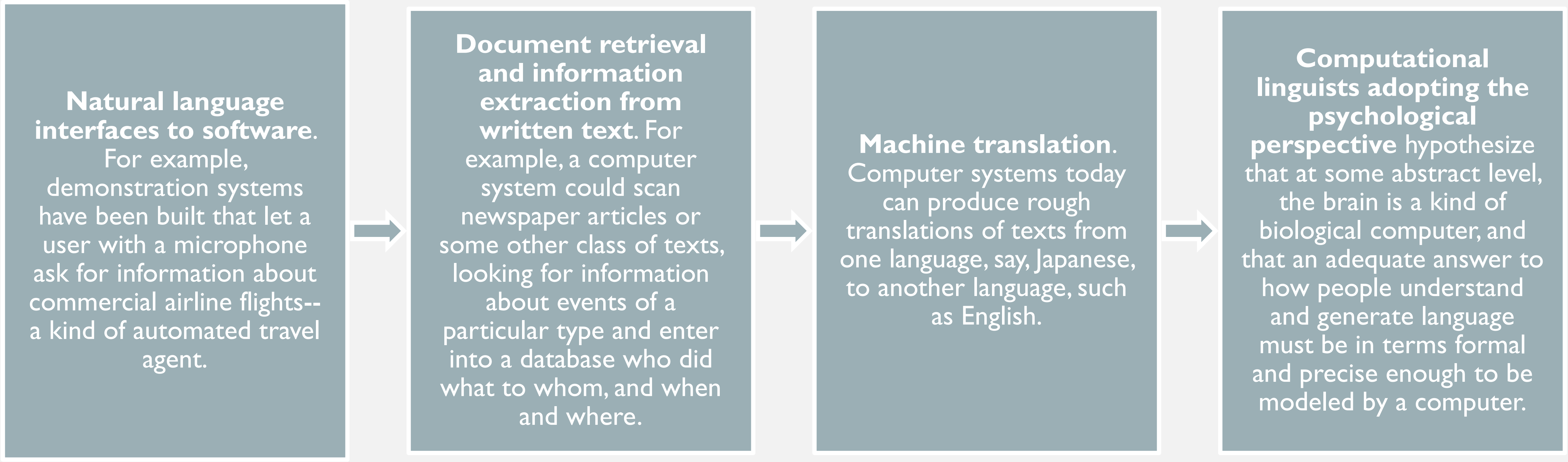
- Goals can be a very far-reaching dream
  - True text understanding
  - Reasoning and decision-making from text
  - Real-time spoken dialog
- Or more close to reality
  - Searching the Web
  - Context-sensitive spelling correction
  - Analyzing reading-level or authorship statistically
  - Extracting company names and locations from news articles.
- Nowadays, the later predominate focused on performing measurably useful tasks *now*.
- Although language is complex, and ambiguity is pervasive, NLP can also be surprisingly easy sometimes:
  - rough text features often do half the job

THE DARK SIDE OF NLP

Natural Language Computing is hard and difficult because of its actual subject (the natural language).

A Natural Language can be:

- Very rich and complicated at all linguistic levels.
- Highly ambiguous or semi-ambiguous.
- Fuzzy, unpredictable, probabilistic.
- Full dependent on context and co-text.
- Embedded a community of interacting people.



Natural language interfaces to software. For example, demonstration systems have been built that let a user with a microphone ask for information about commercial airline flights--a kind of automated travel agent.

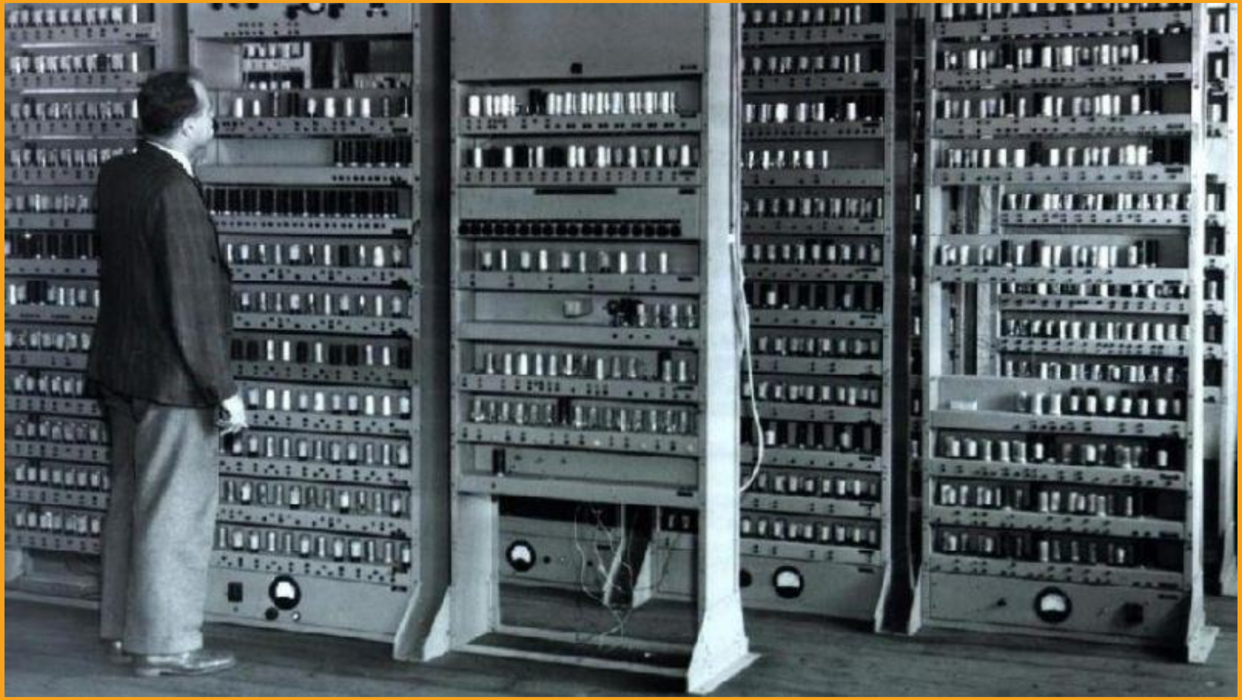
Document retrieval and information extraction from written text. For example, a computer system could scan newspaper articles or some other class of texts, looking for information about events of a particular type and enter into a database who did what to whom, and when and where.

Machine translation. Computer systems today can produce rough translations of texts from one language, say, Japanese, to another language, such as English.

Computational linguists adopting the psychological perspective hypothesize that at some abstract level, the brain is a kind of biological computer, and that an adequate answer to how people understand and generate language must be in terms formal and precise enough to be modeled by a computer.

From *Linguistic Association of America*

THE BRIGHT SIDE OF CL



HISTORICAL OVERVIEW OF COMPUTATIONAL LINGUISTICS

A short time-trip to the recent past

CL: A HISTORICAL REVIEW

- Phase 1: Late 1940s to Late 1960s
- Phase 2: Late 1960s to Late 1970s
- Phase 3: Late 1970s to Late 1980s
- Phase 4: Late 1980s Onward
- Phase 5: The 21st Century
- Where We Are Now

From Jones (1994), *Natural Language Processing: A Historical Review* &
Jurafsky & Martin (2008), *Chapter 1*.

PHASE 1: LATE 1940S TO LATE 1960S

- Early CL on machines less powerful than pocket calculators.
- Foundational work on automata, formal languages, probabilities and information theory.
- First speech systems (Davis et al, Bell Labs).
- MT heavily funded by military, but basically just word substitution programs.
- Little understanding of natural language syntax, semantics, pragmatics.

PHASE 2: LATE 1960S TO LATE 1970S

- Alvey report (1966) ends funding for MT in America - the lack of real results realized
- ELIZA (MIT): Fraudulent NLP in a simple pattern matcher psycholtherapist
  - It's true, I am unhappy.
  - *Do you think coming here will make you not to be unhappy?*
  - I need some help; that much is certain.
  - *What would it mean to you if you got some help?*
  - Perhaps I could earn to get along with my mother.
  - *Tell me more about your family.*
- Early corpora: Brown Corpus (Kudera and Francis)

PHASE 3: LATE 1970S TO LATE 1980S

- Winograd's SHRDLU (1971): existence proof of NLP (in tangled LISP code).
- Could interpret questions, statements commands.
  - –Which cube is sitting on the table?
  - – *The large green one which supports the red pyramid.*
  - – Is there a large block behind the pyramid?
  - – *Yes, three of them. A large red one, a large green cube, and the blue one.*
  - – Put a small one onto the green cube with supports a pyramid.
  - – *OK.*

PHASE 4: LATE 1980S

- Procedural --> Declarative (including logic programming)
- Separation of processing (parser) from description of linguistic knowledge.
- Representations of meaning: procedural semantics (SHRDLU), semantic nets (Schank), logic (perceived as answer; finally applicable to real languages (Montague)
- Perceived need for Knowledge Representation (Lenat and Cyc)
- Working MT in limited domains (METEO)

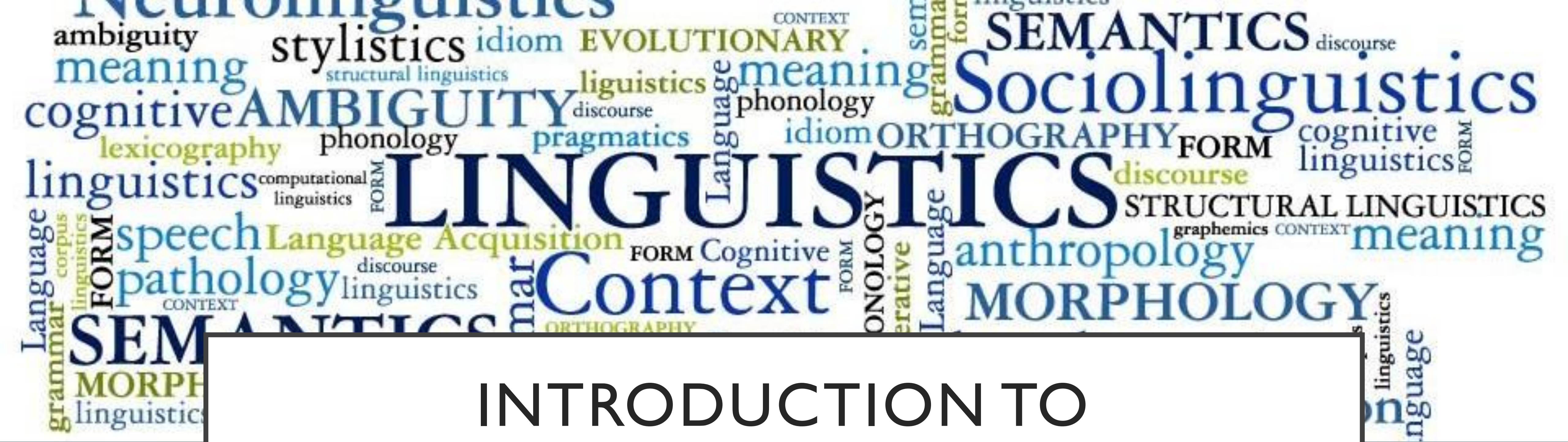
PHASE 4: LATE 1990S

- Resurgence of finite-state methods for NLP: in practice they are incredibly effective.
- Speech recognition becomes widely usable.
- Large amounts of digital text become widely available and reorient the field. The Web.
- Resurgence of probabilistic / statistical methods, led by a few centers, especially IBM (speech, parsing, Candide MT system), often replacing logic for reasoning.
- Recognition of *ambiguity* as key problem.
- Emphasis on machine learning methods.

PHASE 5: THE 21ST CENTURY

The first decade-step of a new era…

- Continued surge in probability, Bayesian methods of evidence combination, and joint inference.
- Emphasis on meaning and knowledge
- Representation.
- Emphasis on discourse and dialog.
- Strong integration of techniques, and levels:
  - brining together statistical NLP and sophisticated
  - linguistic representations.
- Increased emphasis on unsupervised learning.



INTRODUCTION TO COMPUTATIONAL LINGUISTICS

The Six Layers of Computational Linguistics

THE CL APPLICATIONS QUIZ

- TASK I:
  - Which daily CL/ NLP applications do you know?
- TASK II:
  - What is a conversational agent?
  - What is a QA system?
  - What is a tagger and parser?
- TASK III
  - Data Processing vs. Language Processing

LAYERS OF COMPUTATIONAL LINGUISTICS

1. Phonetics & Phonology
2. Morphology
3. Syntax
4. Semantics
5. Pragmatics
6. Discourse

LAYER 1: PHONETICS AND PHONOLOGY

- Phonetics and Phonology— knowledge about linguistic sounds
  - Phonetics: language sounds, how they are physically formed;
  - Phonology: systems of discrete sounds, e.g. languages' syllable structure.

- **processing** /ˈprəʊsesɪŋ/
- **language** /ˈlæŋ-ɡwɪdʒ/
- "It is easy to recognize speech." OR "It is easy to wreck a nice beach."

LAYER 2: MORPHOLOGY

- Morphology— knowledge of the meaningful components of words; the study of the sub-word units of meaning.

processing

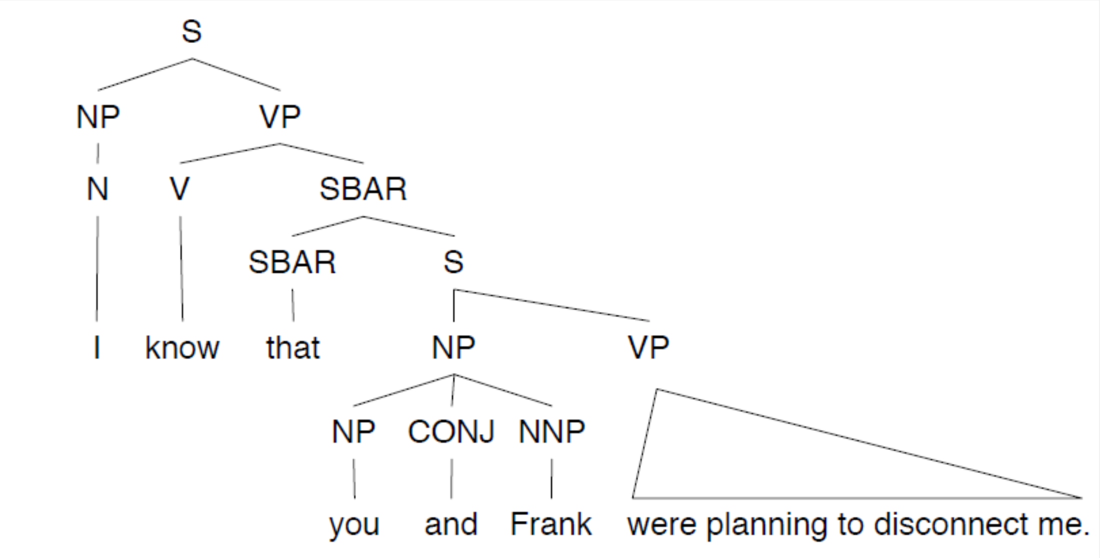
to handle systematically GERUND

But what about more morphologically-rich/ complex languages?

- *megszentségteleníthetetlenségeskedéseitekért*
- For your [plural] continued undesecratable behavior [Hungarian]
- *mempertidakmempertanggungjawabkannyalah*
- to (become clearly) to make not responsible (for a responsibility) of something that (does something) [Indonesian]
- *γαλαζοαίματος* vs. *χορευταρούλικο*

LAYER 3: SYNTAX

- Syntax— knowledge of the structural relationships between words; the study of the structural relationships between words.



- But what about ambiguity?

LAYER 4: SEMANTICS

- Semantics—knowledge of meaning; the study of the literal meaning.
- I know that you and Frank were planning to disconnect me.

- Roles
  - ACTION = disconnect
  - ACTOR = you
  - ACTOR = Frank
  - OBJECT = me.

LAYER 5: PRAGMATICS

- Pragmatics— knowledge of the relationship of meaning to the goals and intentions of the speaker; the study of how language is used to accomplish goals.
- What should you conclude from the fact I said something? How should you react?
- I'm sorry Dave, I'm afraid I can't do that.
- It's getting cold here.
- Cases of mis(pragmatically)interpretations?

LAYER 6: DISCOURSE

- Discourse— knowledge about linguistic units larger than a single utterance; the study of linguistic units larger than a single utterance.
- The structure of conversations: turn taking, thread of meaning, interactions.

David Bowman:

Open the pod bay doors, Hal.

HAL:

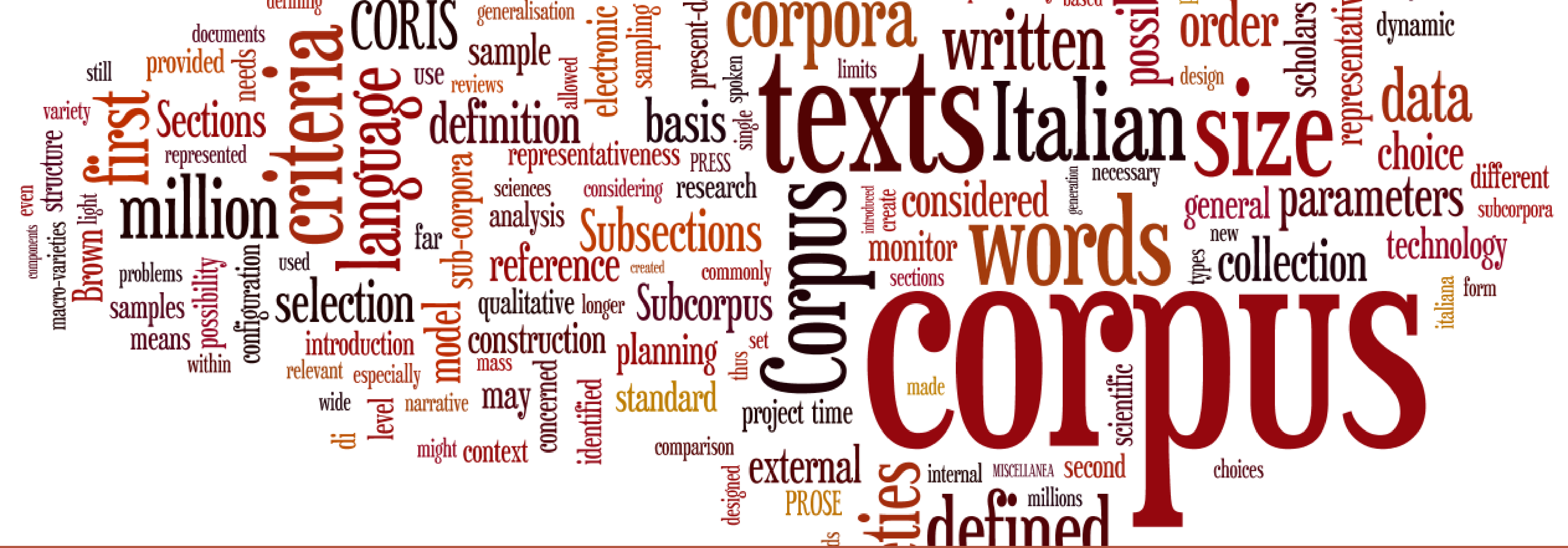
I'm sorry, Dave, I'm afraid I can't do that.

David Bowman:

What are you talking about, Hal?

HAL*:*

I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.



INTRODUCTION TO CORPUS LINGUISTICS

WHAT IS CORPUS LINGUISTICS?

- Corpus linguistics is the analysis of naturally occurring language on the basis of computerized corpora.
- Corpus: a large principled (<u>and systematic</u>) collection of naturally occurring examples of language stored electronically.
- There are many types of corpora, which can be used for different kinds of analyses.

WHAT CORPUS LINGUISTICS IS NOT

- Corpus Linguistics is NOT:
  - able to provide negative evidence.
  - able to explain why.
  - able to provide all possible languages at one time.

TYPES OF CORPORA

- general/reference corpora (vs. specialized corpora) (e.g. BNC = British National Corpus, or Bank of English): aim at representing a language or variety as a whole (contain both spoken and written language, different text types etc.)
- historical corpora (vs. corpora of present-day language) (e.g. Helsinki Corpus, ARCHER) aim at representing an earlier stage or earlier stages of a language
- regional corpora (vs. corpora containing more than one variety) (e.g. WCNZE = Wellington Corpus of Written New Zealand English) aim at representing one regional variety of a language
- learner corpora (vs. native speaker corpora) (e.g. ICLE = International Corpus of Learner English) aim at representing the language as produced by learners of this language
- multilingual corpora (vs. one-language corpora) aim at representing several, at least two, different languages, often with the same text types (for contrastive analyses)
- spoken (vs. written vs. mixed corpora) (e.g. LLC = London-Lund Corpus of Spoken English) aim at representing spoken language

CORPUS SOFTWARE: TYPES

Two types of software for corpus analysis can be distinguished in principle:

- software that is tailored to one specific corpus
  - Examples of this type are two software programs that have been tailored to the British National Corpus, namely SARA and BNCWeb.
  - A further example is ICE-CUP, which has been tailored to ICE-GB (such as the online search facilities provided for the BNC and the Collins Wordbanks Online English)
- software that can be used with almost any kind of corpus.
  - MonoConc Pro (demo available at http://www.camsoftpartners.co.uk/monoconc.htm) and
  - WordSmith Tools, which is probably the most widely used corpus software.

WHAT CAN CORPUS SOFTWARE DO?

- While there are many differences between the software packages designed for corpus analysis, certain basic functions can be performed by practically all the available software.
  - the occurrence of certain strings (i.e. words or phrases) a.k.a. concordance-lines
  - sorting (for example according to the word to the right or left of the search term)
  - “thinning” the results (i.e. the removal of irrelevant instances)
  - search for words or phrases occurring within a certain distance of each other and also usually allows the use of wildcards in searches.

The Ambiguous Cesar and the the Problematix Gauls

AMBIGUITY AND ITS 'FRIENDS'

LINGUISTIC RULES

- Simple Rules of word formation
- i.e. the plural cases of English
  - cat > cat**s**
  - bus > bus**es**
  - ferry > ferr**ies**
  - knife > kni**ves**
  - fish > fish
  - child > child**ren**
  - mouse > m**ic**e

LINGUISTIC RULES

- Complex rules of word formation
- i.e. the generation of a compound word

ασπρόμαυρος	(black-and-white)
σπιρτ**ό**κουτ**ο**	(matchbox)
κυμα**τ**ομορφή	(waveform)
χορ**ο**πηδώ	(jump-up-and-down)
ανα**ψ**οκοκκινίζω	(blush)
αιμα**τ**οβα**μ**μένος	(bloodstained)
κοκκινογένης	(red-beard)

AMBIGUITY

- A perhaps surprising fact about these categories of linguistic knowledge is that most tasks in speech and language processing can be viewed as resolving **ambiguity** at one of these levels.
- We will introduce the models and algorithms we discuss as ways to **resolve** or **disambiguate** these ambiguities. For example deciding whether *contact* is a verb or a noun can be solved by **part-of-speech tagging**.

- TASK IV:
  - *Find an ambiguous example for each linguistic level.*

AMBIGUITY

- Part-of-Speech Ambiguity
  - cut: VB VBP VBN N
  - ενημερώσεις: VBC NN
- Syntactic attachment ambiguity
  - John hits an old lady with a stick.
  - ἤξεις ἀφήξεις οὐκ ἐν πολέμῳ θνήξεις
- Word sense ambiguity:
  - interest: curiosity, concern | earnings | personal benefit
  - τέλος: end | fees, rates | death
- Semantic interpretation (ambiguities above the word level)

AMBIGUITY: A WORDNET EXAMPLE

- 1. table, tabular array -- (a set of data arranged in rows and columns; "see table 1")
- 2. table -- (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs; "it was a sturdy table")
- 3. table -- (a piece of furniture with tableware for a meal laid out on it; "I reserved a table at my favorite restaurant")
- 4. mesa, table -- (flat tableland with steep edges; "the tribe was relatively safe on the mesa but they had to descend into the valley for water")
- 5. table -- (a company of people assembled at a table for a meal or game; "he entertained the whole table with his witty remarks")
- 6. board, table -- (food or meals in general; "she sets a fine table"; "room and board")

FROM AMBIGUITY TO GRAMMATICALITY

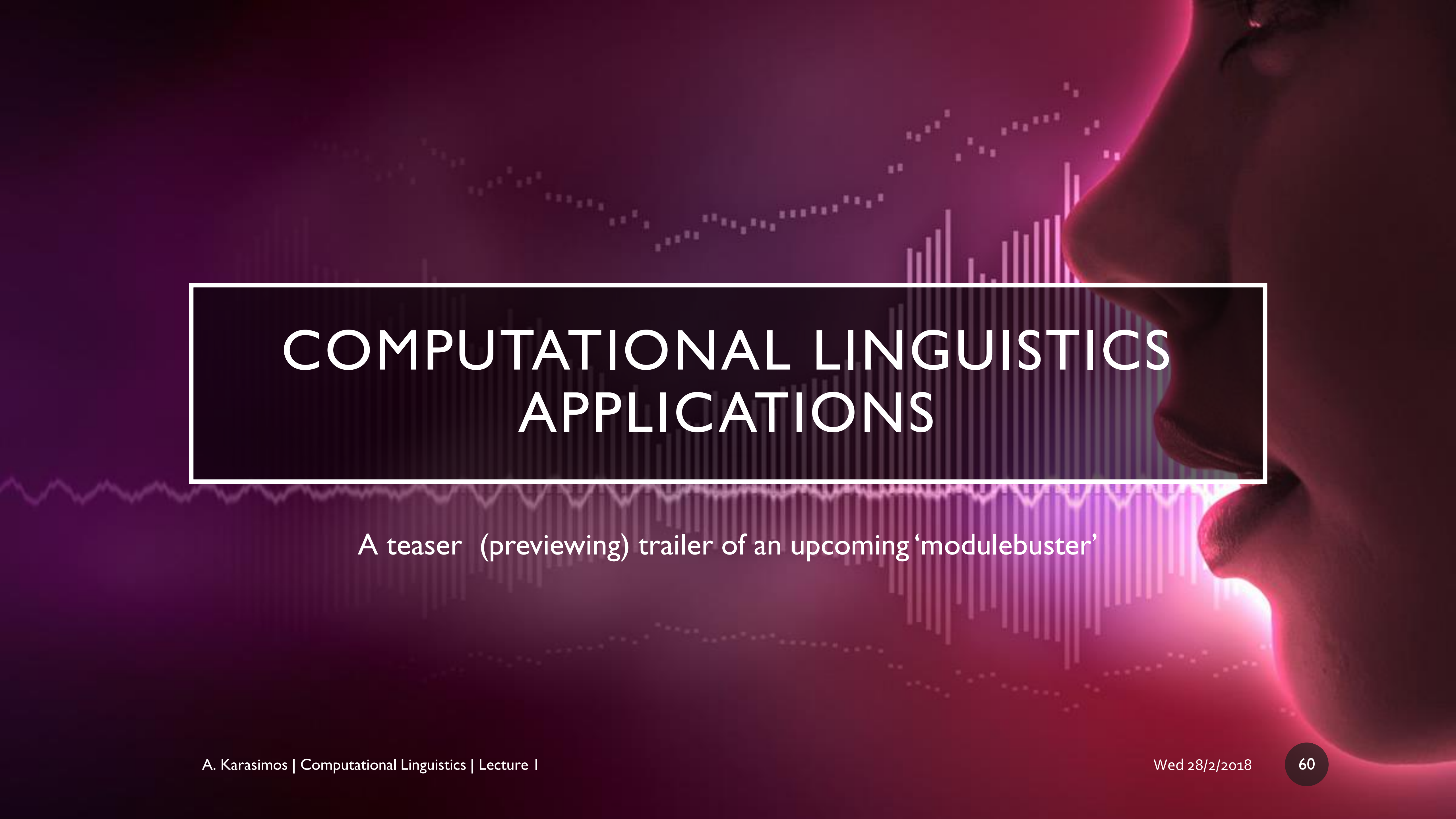
- **What is grammatical and what isn't?**
  - John I believe Sally said Bill believed Sue saw.
  - What did Sally whisper that she had secretly read?
  - John wants very much for himself to win.
  - Who did Jo think said John saw him?
  - The boys read Mary's stories about each other.
  - Mary, while John had had had had had had had had had was the correct answer.

FROM AMBIGUITY TO GRAMMATICALITY

- **What is grammatical and what isn't?**
  - John I believe Sally said Bill believed Sue saw.
  - What did Sally whisper that she had secretly read?
  - John wants very much for himself to win.
  - Who did Jo think said John saw him?
  - The boys read Mary's stories about each other.
  - Mary, while John had had "had" had had "had had;" "had had" was the correct answer.

LANGUAGE EVOLUTION: A DISGUISED AMBIGUITY?

- **Morphology**
  - We learn new words (neologisms) all the time:
    - bioterrorism, infotainment, fitspiration,
      cat lady, craptacular, haterade, lovefest
- **Part-of-speech**
  - Historically: "kind" and "sort" were always *nouns*:
    - "I knowe that sorte of men ryght well." [1560]
  - Now also used as *degree modifiers:*
    - "I'm sort of hungry." [Present]
    - "It sort o' stirs one up to hear about old times." [1833]
  - The case of «λίγο» and Greek Future tenses.



COMPUTATIONAL LINGUISTICS APPLICATIONS

A teaser (previewing) trailer of an upcoming 'modulebuster'

PROCESSING TEXT AND SPEECH: WHY?

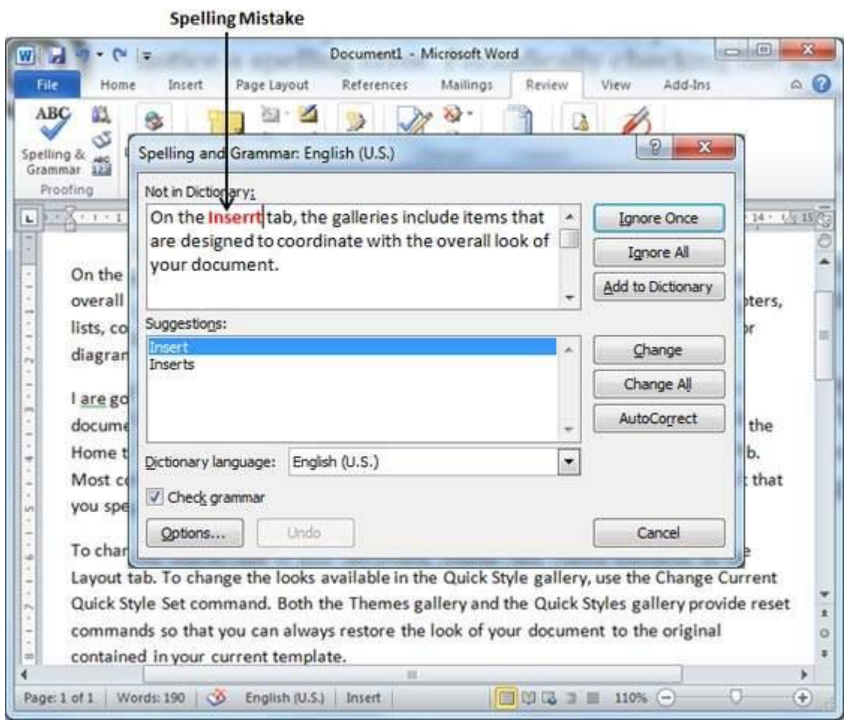
80% of information and texts are unstructured

- Full-information chains in organization texts
- Products and e-markets: presentations, advs, promotions
- Material from users: blogs, forums/ fora, wikis
- Customer opinions: social media, personal analysis

Enormous amount of data

- 161,000,000 GBs in digital content in 2006
- ~ 1000 EBs in digital content in 2010
- Sound and image need abstracts and labels
- Numerous text bodies without annotation and metadata

SPELL AND GRAMMAR CHECKING



Spelling Mistake

On the Inserrt tab, the galleries include items that are designed to coordinate with the overall look of your document.

Suggestions:
- Insert
- Inserts

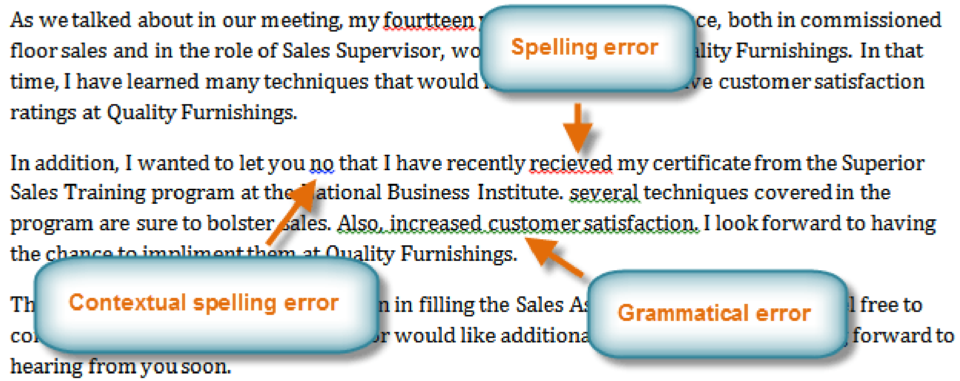
☑ Check grammar

On the
overall
lists, co
diagrar

I are go
docume
Home t
Most c
you spe

To char

Layout tab. To change the looks available in the Quick Style gallery, use the Change Current Quick Style Set command. Both the Themes gallery and the Quick Styles gallery provide reset commands so that you can always restore the look of your document to the original contained in your current template.



As we talked about in our meeting, my fourtteen … ce, both in commissioned floor sales and in the role of Sales Supervisor, wo … lity Furnishings. In that time, I have learned many techniques that would … ve customer satisfaction ratings at Quality Furnishings.

In addition, I wanted to let you no that I have recently recieved my certificate from the Superior Sales Training program at the ational Business Institute. several techniques covered in the program are sure to bolster sales. Also, increased customer satisfaction. I look forward to having the chance to impliment them at Quality Furnishings.

Th … n in filling the Sales As … l free to co … r would like additiona … forward to hearing from you soon.

Spelling error

Contextual spelling error

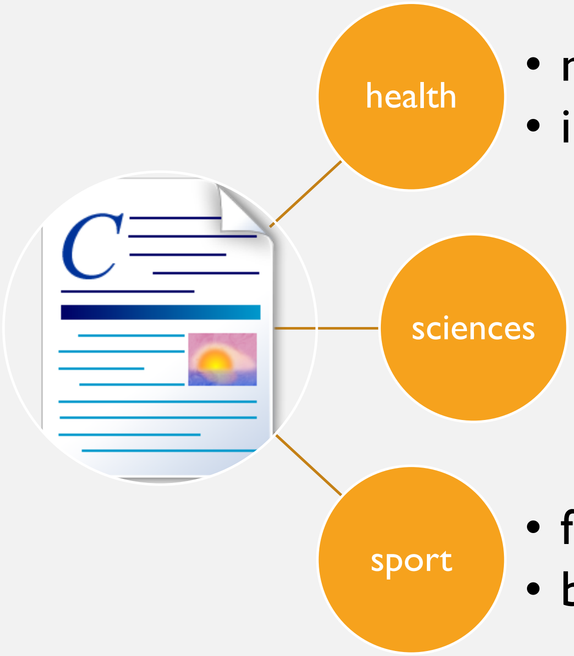
Grammatical error

INFORMATION RETRIEVAL

An earthquake of 3.7 magnitude (according to the EMSC automated solution and the Geodynamic Institute of Athens) took place on 14 March 2015. The precise focus of the vibration is 93km Southeast of Agios Nikolaos Crete and 128km Southwest of Karpathos. The focal depth of the earthquake is estimated at 5km. The earthquake, as recorded by the seismograph of the Seismological Network of the Geodynamic Institute of the National Observatory of Athens, which is located in Zakros, in the prefecture of Lassithi. Only a few disasters were recorded in old houses in villages in the prefecture of Lasithi.

```
Destruction type: earthquake
    location: Crete
    date: 14/3/2015
    size: 3.7
    Focus: 93km Southeast of Agios Nikolaos Crete and
    128km Southwest of Karpathos.
    Source: EMSC and Geodynamic Institute of Athens
damage:
    human:
        victim: -
        number: -
        result: -
    materials:
        object: villages of Lassithi prefecture
        result: losses
```

TEXT CATEGORIZATION



health

- nutrition
- illnesses

sciences

- Humanities
- Social Sciences

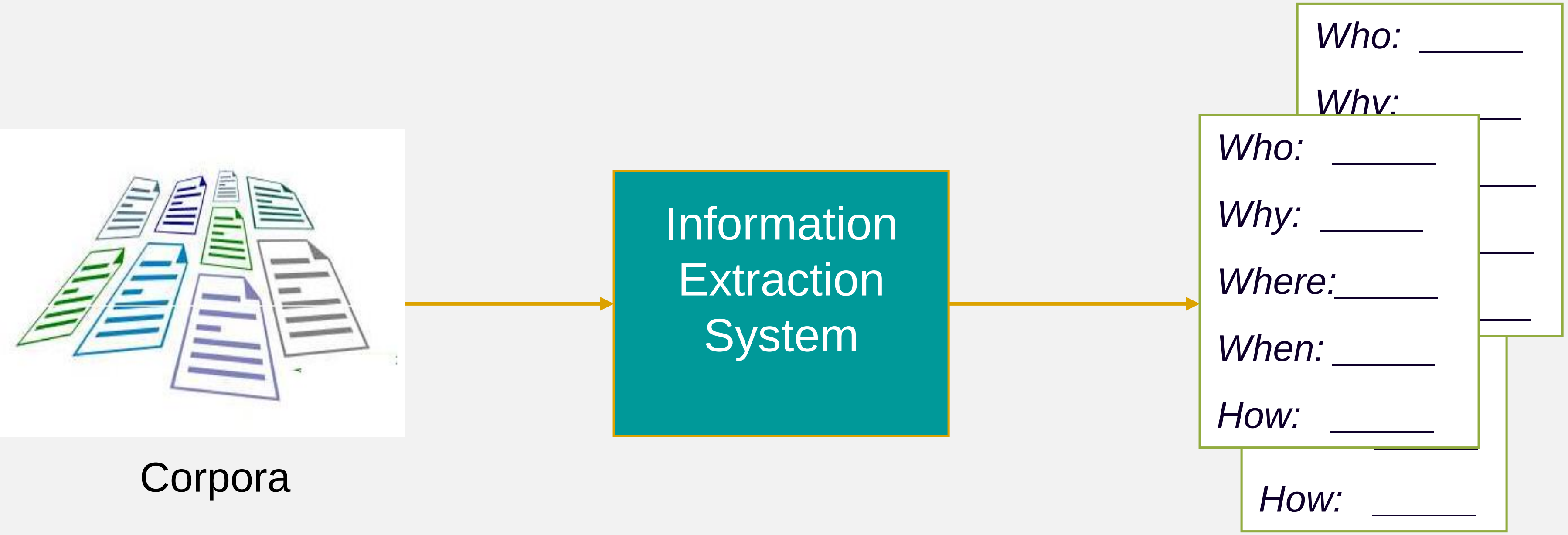
sport

- football
- basketball

TEXT CATEGORIZATION

- RSS Feed and News feeds
  - Categorize incoming news, news and stories
- Queries on search engines
  - Google: search "author of Metaphysics"
- Detect spam emails
  - http://www.paulgraham.com/spam.html
- Filtering emails to appropriate individuals and groups

INFORMATION EXTRACTION



Corpora

Information Extraction System

Who: ______

Why: ______

Who: ______

Why: ______

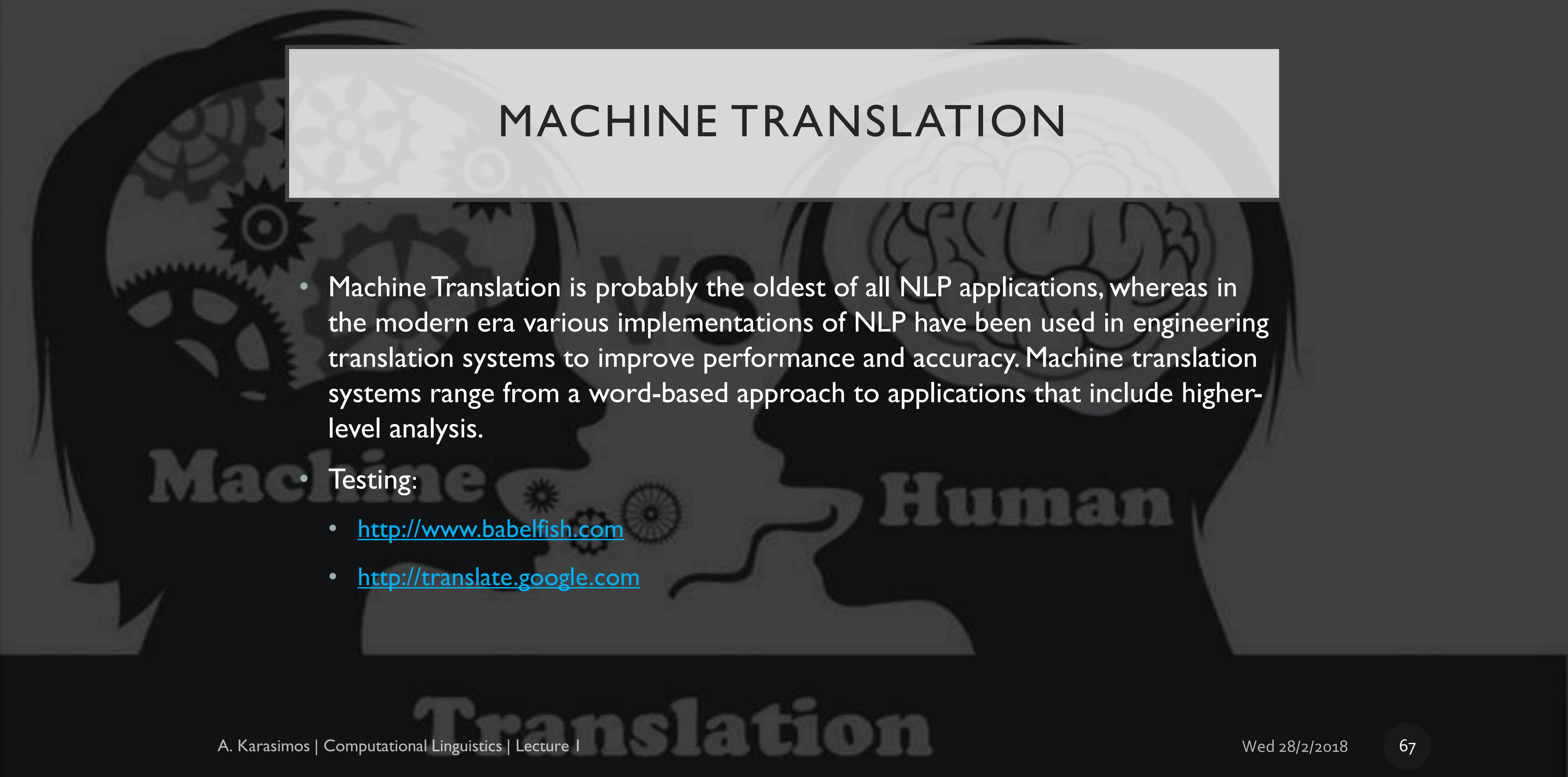
Where:______

When: ______

How: ______

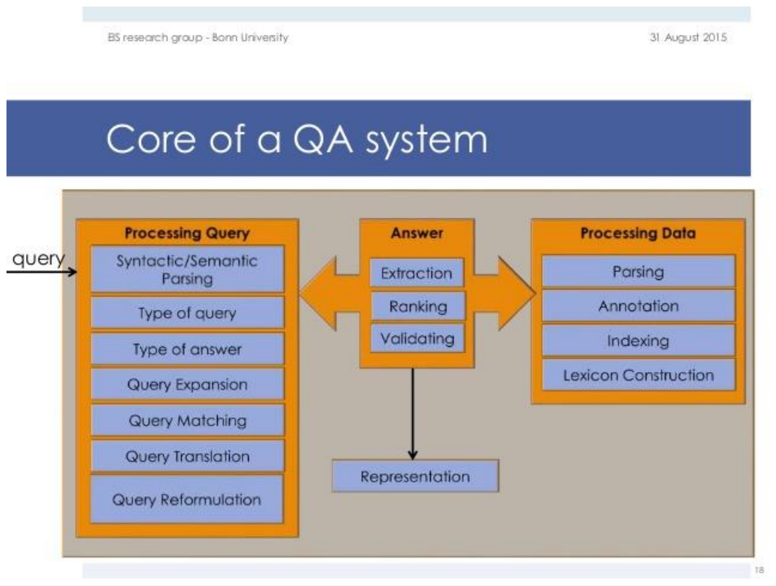
How: ______

MACHINE TRANSLATION



- Machine Translation is probably the oldest of all NLP applications, whereas in the modern era various implementations of NLP have been used in engineering translation systems to improve performance and accuracy. Machine translation systems range from a word-based approach to applications that include higher-level analysis.
- Testing:
  - http://www.babelfish.com
  - http://translate.google.com

QUESTION-ANSWERING SYSTEMS



EIS research group - Bonn University

31 August 2015

Core of a QA system

- Unlike information retrieval, which provides a list of related documents in response to a user query
- QAS provides the user with either the text of the answer itself or different alternative answers.
- Testing:
  - http://start.csail.mit.edu/index.php
  - http://www.wikiqa.de/index.php

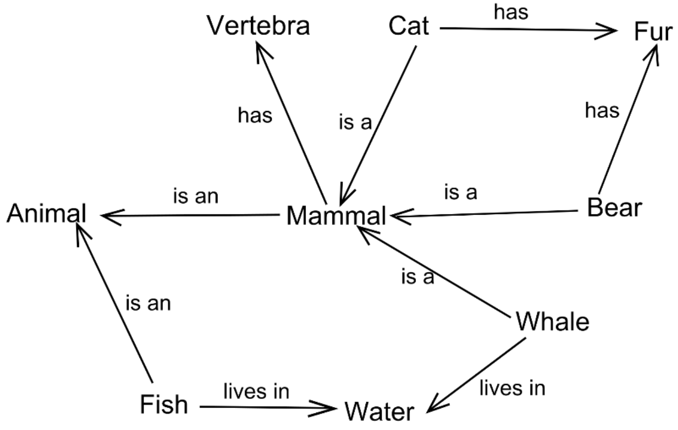
SEMANTIC NETWORKS

Computer-readable/ writeable data for linking and creating a hypertext.

The presence of metadata in documents is essential

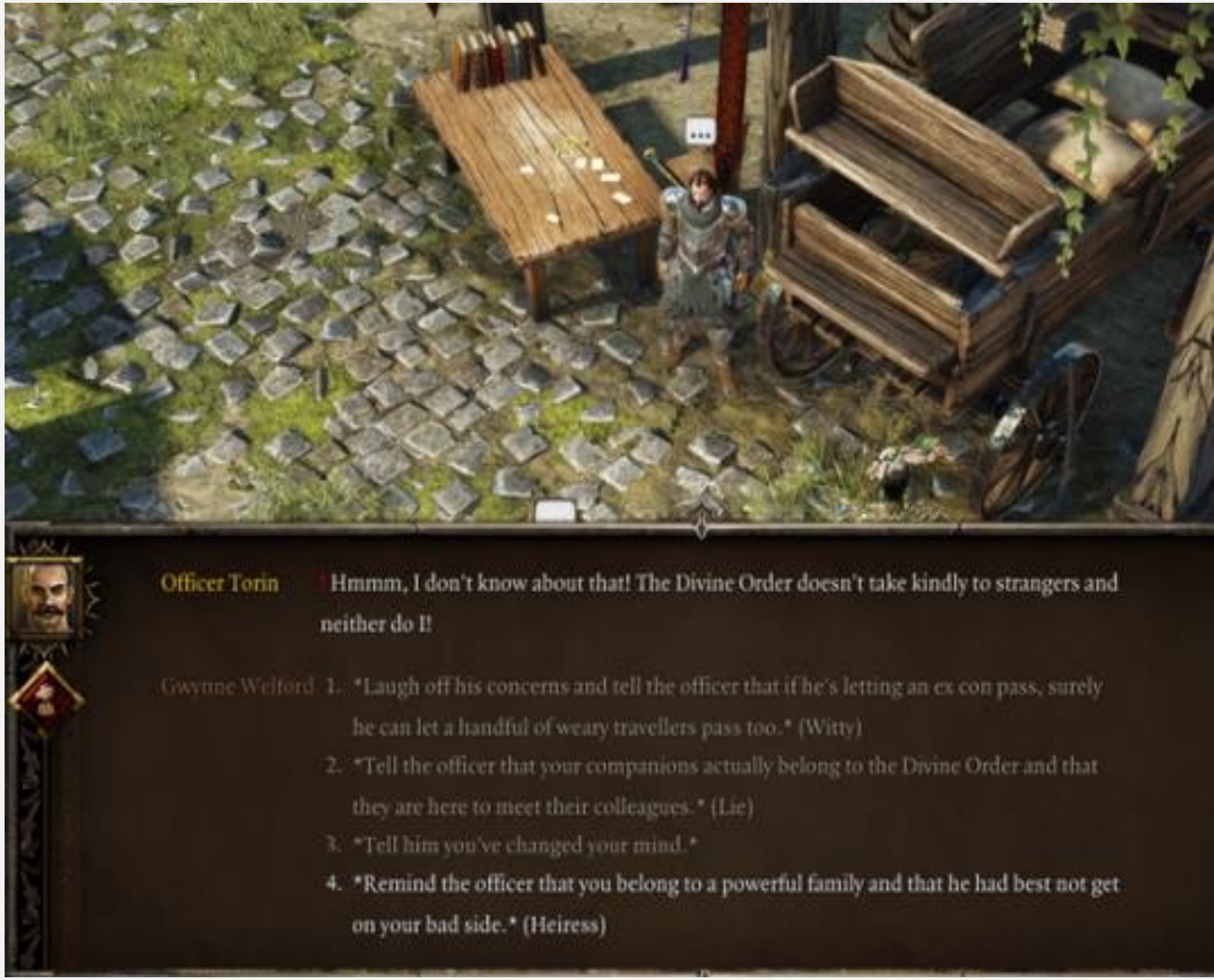
Direct: title, author, date of creation

Indirect: Related information such as entity names and their relationship



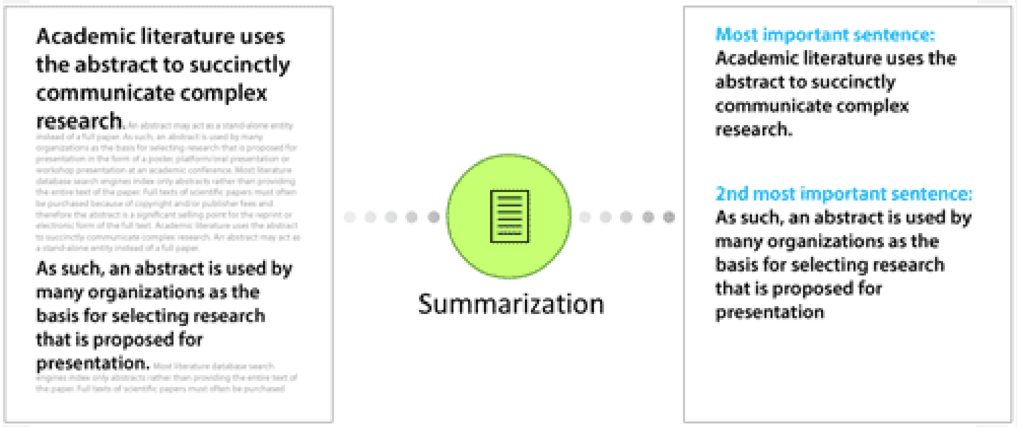
DIALOGUE SYSTEMS

- Perhaps the omnipresent application of the future, as the system that all major application providers/ creators/ producers envision it as an end use.
- Dialogue systems tend to focus on a limited application (e.g. from refrigerators, smart TVs, next generation mobile phones, to online digital assistants in companies or call centers),
- They are currently using the vocal or dictionary level of language. It is believed that the use of all levels of language processing at the levels mentioned above will enable real interactive systems.



Officer Torin Hmmm, I don't know about that! The Divine Order doesn't take kindly to strangers and neither do I!

Gwynne Welford
1. *Laugh off his concerns and tell the officer that if he's letting an ex con pass, surely he can let a handful of weary travellers pass too.* (Witty)
2. *Tell the officer that your companions actually belong to the Divine Order and that they are here to meet their colleagues.* (Lie)
3. *Tell him you've changed your mind.*
4. *Remind the officer that you belong to a powerful family and that he had best not get on your bad side.* (Heiress)



Academic literature uses the abstract to succinctly communicate complex research. An abstract may act as a stand-alone entity instead of a full paper. As such, an abstract is used by many organizations as the basis for selecting research that is proposed for presentation in the form of a poster, platform/oral presentation or workshop presentation at an academic conference. Most literature database search engines index only abstracts rather than providing the entire text of the paper. Full texts of scientific papers must often be purchased because of copyright and/or publisher fees and therefore the abstract is a significant selling point for the reprint or electronic form of the full text. Academic literature uses the abstract to succinctly communicate complex research. An abstract may act as a stand-alone entity instead of a full paper.

As such, an abstract is used by many organizations as the basis for selecting research that is proposed for presentation. Most literature database search engines index only abstracts rather than providing the entire text of the paper. Full texts of scientific papers must often be purchased

Summarization

Most important sentence:
Academic literature uses the abstract to succinctly communicate complex research.

2nd most important sentence:
As such, an abstract is used by many organizations as the basis for selecting research that is proposed for presentation

SUMMARIZATION

Summarization reduces a larger text by converting it to a smaller one, but retaining its richly hierarchical structural representation of the original document with key information.

Testing:

- http://apidemo.pingar.com/Summarize.aspx#wrapper
- http://textsummarization.net/text-summarizer

SUMMING UP

SUMMARY AND COMING SOON…

- Computational Linguistics is an applied discipline with an increasingly large inventory of applications.
- A wide variety of levels of analysis are used to implement these applications.
  - Many, but not all of these levels, are derived from or inspired by theoretical linguistics
- Corpus Linguistics: In the next lecture we will introduce some models of the patterns these produce and programs for identifying these patterns.
- In the next lecture we will introduce some models of the patterns these produce and programs for identifying these patterns (a.k.a. Regular Expressions).

SUMMARY

- Space Odyssey 2001: A good way to understand the concerns of speech and language processing.
- Speech and Language technology relies on formal models, or representations, of knowledge of language at the levels of phonology and phonetics, morphology, syntax, semantics, pragmatics and discourse.
- The foundations of speech and language technology lie in computer science, linguistics, mathematics, electrical engineering and psychology.
- The critical connection between language and thought has placed speech and language processing technology at the center of debate over intelligent machines.
- Revolutionary applications of speech and language processing are currently in use around the world

WEEK READINGS

- Karen Sparck Jones (1994). Natural Language Processing: A Historical Review. In A. Zampolli, N. Calzolari & M. Palmer *Current issues in computational linguistics:* in honour of Don Walker. *Linguistica Computazionale*, vol. 9-10, pp. 3-16. Pisa, Dordrecht.
- Jurafsky D. & J. Martin (2008). SPEECH and LANGUAGE PROCESSING An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (2nd Edition + 3rd Edition). CHAPTER 1.
- Benett, G. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. CHAPTER 1. University of Michigan Press.