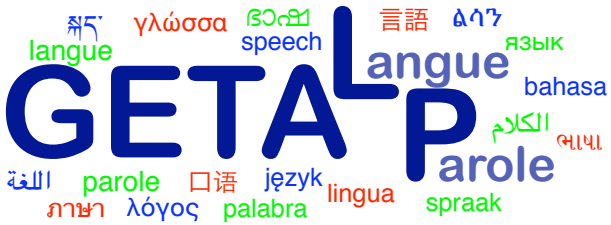
Natural Language Processing in a Nutshell

Hervé Blanchon
Laboratoire LIG
Équipe GETALP
herve.blanchon@univ-grenoble-alpes.fr



Outline

- Introduction
- Natural Language Processing Applications
- Levels of Treatment
- Ambiguity
- NLP Approaches in a Nutshell
- Tools
  - Words and Words in Context
  - Collocation
  - Stemming and Lemmatization
  - POS Tagging
  - Named-Entity Recognition
  - Parsing
  - Coreference Resolution
- NLP Frameworks and Packages
- Other Resources and Tools
- Credits

INTRODUCTION

What is this Session About?

- Broad overview of Natural Language Processing (NLP)
- Vocabulary
  - applications of NLP
  - levels of treatment
- Resources & Tools

Every Natural Language…

- …evolve over time
  - new vocabulary, changes of syntax
    - new artefacts, new concepts, new ideas,…
      - reading Shakespeare's writings from the sixteenth century?
      - reading Rabelais' writings from the fifteenth century?
      - reading Dante's writing from the thirteenth century?
- …evolve over space
  - French from France, Canada, Africa
  - British English, American English, global English
- …is ambiguous

Natural language is…

- Highly ambiguous at all levels
- Complex and subtle use of context to convey meaning
- Fuzzy, probabilistic
- Involves reasoning about the world
- A social system
  - a key part of people interacting with other people (persuading, insulting & amusing them)

Text Understanding is Very Hard

- Example
  - John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

Text Understanding is Very Hard

- Example
  - John stopped at the **donut** store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.
- *What's hard about "**donut**"?*
  - spare tire only intended for temporary use
  - stupid individual
  - deep-fried piece of dough with a hole in the center
  - anything in the shape of a torus
  - …

Text Understanding is Very Hard

- Example
  - John stopped at the **donut store** on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.
- *What's hard about "**donut store**"?*
  - where donuts shop?
  - run by donuts?
  - which looks like a big donut?
  - made of donut?
  - which has an emptiness at its core?
  - …

Text Understanding is Very Hard

- Example
  - John stopped at **the donut store on his way home from work**. He thought a coffee was good every few hours. But it turned out to be too expensive there.
- *What's hard about "**the donut store … work**"?*
  - describes where the store is?
  - describes when he stopped?

Text Understanding is Very Hard

- Example
  - John stopped at the donut store on his way home from work. **He thought** a coffee was good every few hours. But it turned out to be too expensive there.
- *What's hard about "**He thought**"?*
  - He -> need to determine that it refers to John
  - he thought at that moment?
  - he thought habitually?

Text Understanding is Very Hard

- Example
  - John stopped at the donut store on his way home from work. He thought a coffee was good **every few hours**. But it turned out to be too expensive there.

- *What's hard about "**every few hours**"?*
  - **he thought every few hours** that a coffee was good?
  - he thought **a coffee every few hours** was good?
  - he thought **a coffee stays good for every few hours**?
  - *Similarly*: "In America a woman has a baby every 15 minutes. Our job is to find that woman and stop her" *Groucho Marx*

Text Understanding is Very Hard

- Example
  - John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But **it** turned out to be too expensive there.
- *What's hard about "**it**"?*
  - stands for the coffee?
  - stands for the donut store?

Text Understanding is Very Hard

- Example
  - John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be **too expensive** there.
- *What's hard about "**too expensive**"?*
  - connect "it" to "expensive"
  - too expensive for what? what are we supposed to conclude about what John did?

Dialogue undestanding is very hard

- Example

  U: Where is **A bug's life** playing in **Mountain View**?

  S: A bug's life is playing at the **Century 16 theater**.

  U: When is **it** playing **there**?

  S: **It**'s playing at 2pm, 5pm, and 8pm.

  U: I'd like 1 **adult** and 2 **children** for **the first show**. How much would **that** cost?

- Knowledge sources:
  - Domain knowledge: a bug's life, Mountain view, Century 16 theater
  - Discourse knowledge: it, there, that
  - World knowledge: adult, children, the first show

Levels of Language

- Phonetics/phonology/morphology
  - what words (or subwords) are we dealing with?
- Syntax
  - What phrases are we dealing with?
  - Which words modify one another?
- Semantics
  - What's the literal meaning?
- Pragmatics
  - What should you conclude from the fact that I said something?
  - How should you react?

NLP APPLICATIONS

Applications

- Machine Translation
  - topic of the next session
- Summarization
  - reduce the size of a document retaining the most important information
- Natural language generation
  - produce natural language from a knowledge base, a database or a logical form
- Text classification
  - assign predefined categories
    - eg: automatic span detection (binary classifier)
    - eg: organize news stories by topics, …

Applications

- Information extraction
  - extract relevant information for future use
    - eg: detect events in emails and add them to the calendar
- Question answering
  - answer question posed by humans expressed in a natural language

Question answering

START
Natural Language Question Answering System

what is the birthdate of Victor Hugo?

Ask Question ›

===> what is the birthdate of Victor Hugo?

Victor Hugo (I)'s date of birth: February 26, 1802 in Besançon, Doubs, France

Source: The Internet Movie Database

Victor Hugo was born on February 26, 1802.

I know about one more person called "Victor Hugo": Victor Hugo (footballer)

Source: Wikipedia

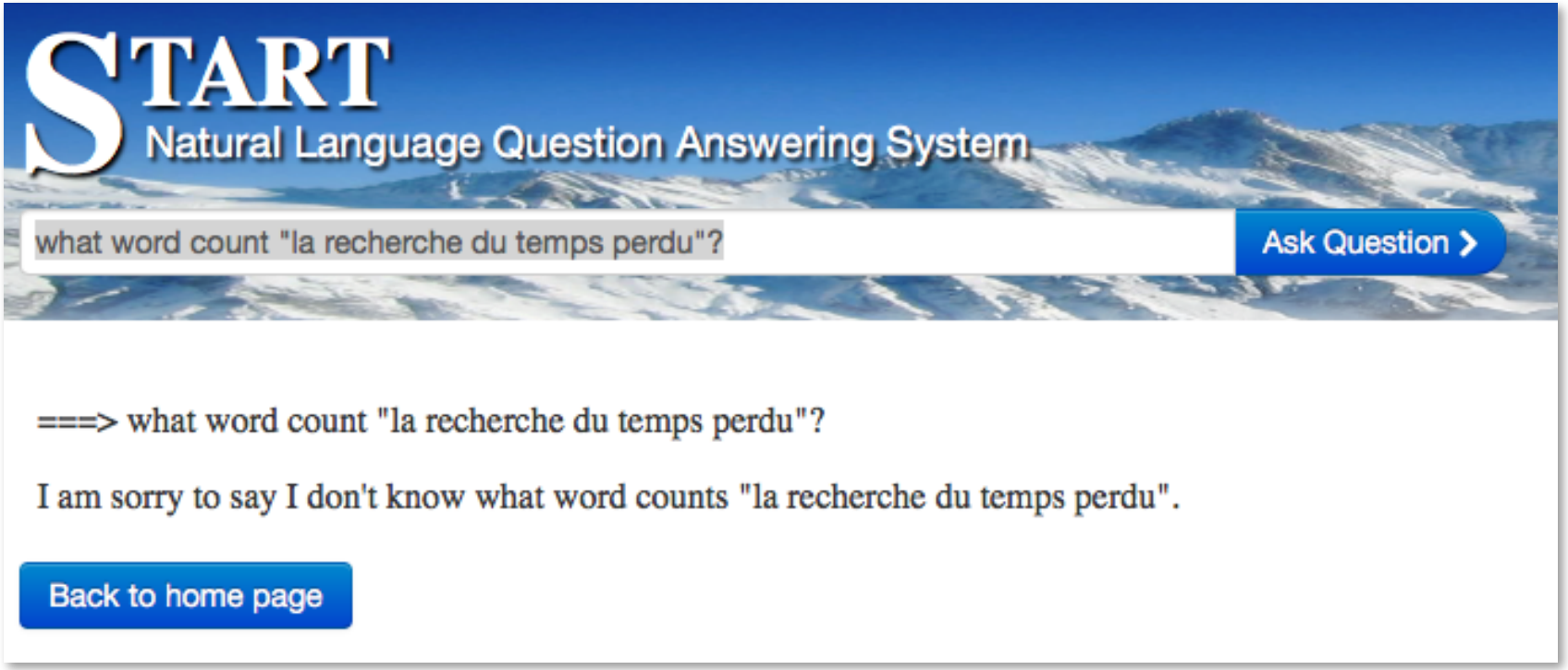
Hugo, Victor

Date of birth: 1802

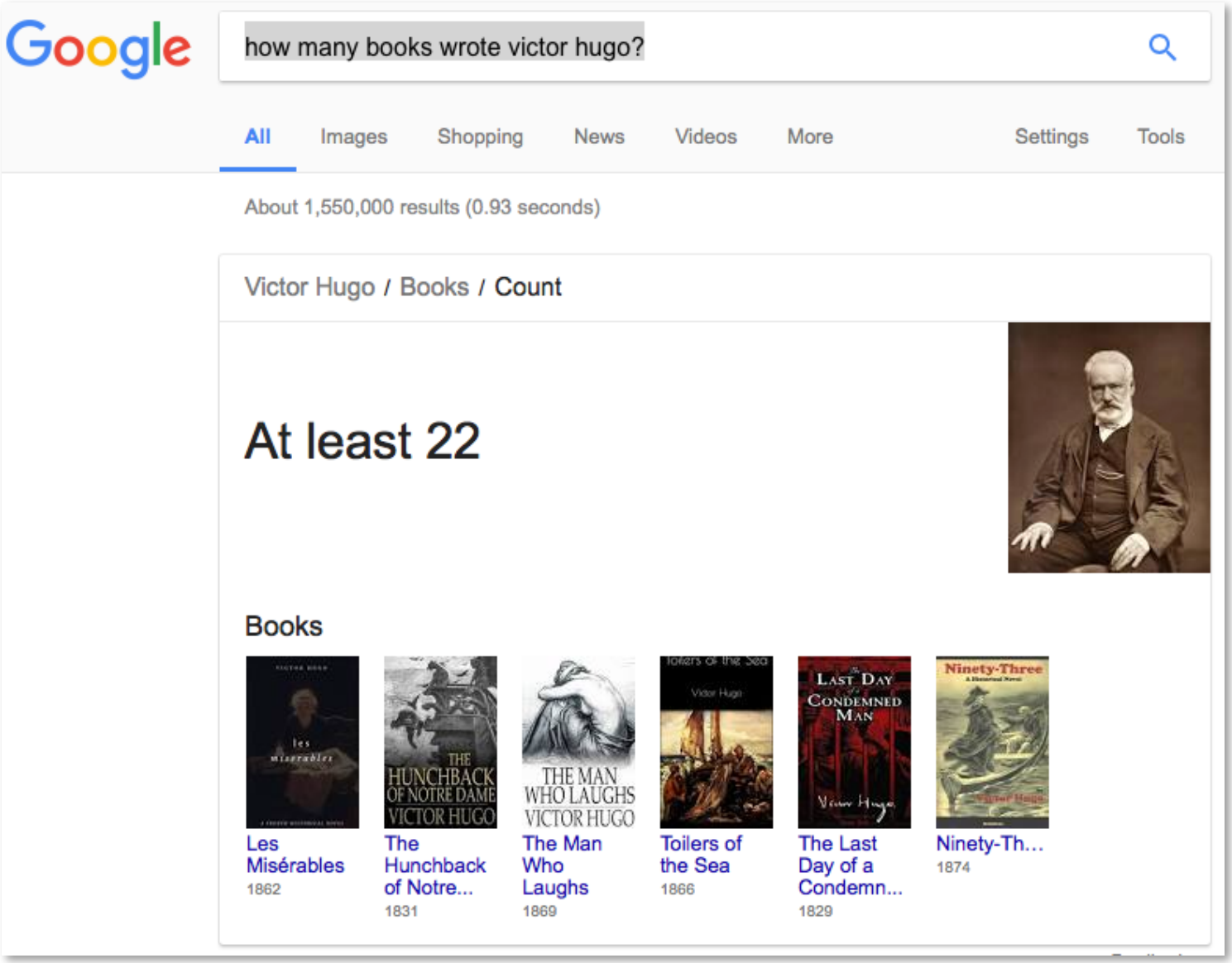
Source: The Gutenberg Project

Back to home page

Question answering



Question answering



Applications

- Information extraction
  - extract relevant information for future use
    - eg: detect events in emails and add them to the calendar
- Question answering
  - answer question posed by humans expressed in a natural language
- Information retrieval
  - retrieve documents relevant to a query among a collection of documents
    - search based on full-text or other content-based indexing
  - may be specialized: medical imagery
  - or generalized: searching on the Web

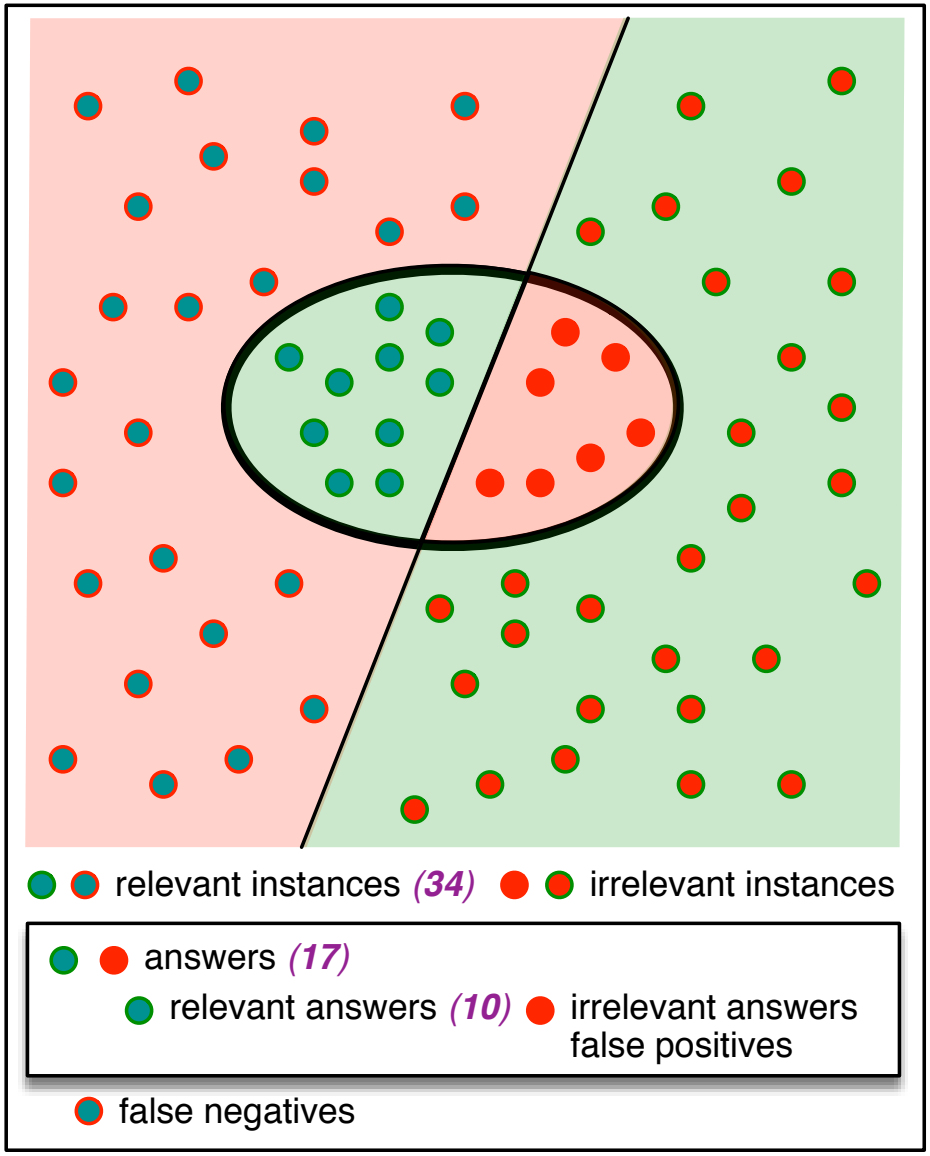
Applications

- Sentiment analysis
  - identify and categorize opinions expressed in a piece of text (attitude towards a topic -> positive, negative, neutral)
- Spelling correction & Grammar checking

- Speech recognition
  - produce a text from an acoustic signal
- Speech synthesis
  - produce an acoustic signal from a text

Evaluation Measures: Precision, Recall

- Precision
  - fraction of retrieved instances that are relevant
  - $P = \frac{\#\ of\ relevant\ answers}{\#\ of\ answers}$
- Recall
  - fraction of relevant instances that are retrieved
  - $R = \frac{\#\ of\ relevant\ answers}{\#\ of\ relevant\ instances}$
- For the example
  - $P = \frac{10}{17} = 0.59$
  - $R = \frac{10}{34} = 0.29$



LEVELS OF TREATMENT

Phonetic/Phonological

- Sounds to “characters” or “words”
  - problem with homophonous sequences
    - fr: ils étaient très amis vs ils étaient treize amis
    - en: the stuffy nose can lead to problems vs the stuff he knows can lead to problems

Words

- Basic units for most NLP tasks
  - languages with word separator: OK
  - languages with no word separator: segmentation needed
    - jp: 単語分割を行う (perform word segmentation)
      - jp: 単語/分割/を/行う
        (tango, word/bunkatsu, segmenting/particle o/okonau, to perform)
    - zh: 当原子结合成分子时
      - 当/原子/结合/成/分子时 **(OK)**
        (when/atoms/combine/molecule/the time)
      - 当/原子/结合/成分/子时 **(KO)**
        (when/atoms/combine/ingredient/midnight)

Words

- Construction from minimal sense units (lemma, dictionary entry)
  - study, studies, studying, studied
  - remarkable, unremarkable
  - short-sighted, fat-free, eyelashes, car park
- Lexical category (Part of Speech, POS)
  - common noun (cat, cats, snow), proper noun (IBM, Italy)
  - adjective (old, older, oldest), adverb (slowly)
  - verb (see, count)
  - …
  - number (122, 6589, one)
  - conjunction (and, or) , determiner (the, some), pronoun (he, its)
  - preposition (to, with), particle (off, up), interjection (Ow, Eh)
  - modal (can, had)
  - …

Open class (lexical) words

Closed class (functional)

Words

- Grammatical category
  - tense (present, past,...), number (singular, plural,...), gender (masculine, feminine, neutral), case
  - Case in German
    - nominative (subject, attribute):
      - **Der gute Mann** ist groß.
      - The good man is big.
    - accusative (complement of noun)
      - Das Hemd **des guten Mannes** ist schön.
      - The good man's shirt is beautiful.
    - dative (indirect object)
      - Ich gebe **dem guten Mann** ein Buch.
      - I give the good man a book.
    - genitive (direct object)
      - Ich höre **den guten Mann**.
      - I hear the good man.

Phrase

Syntagmatic group of words

- Phrasal category
  - adjective phrase (very hot)
  - adverbial phrase (too slowly)
  - adpositional phrase
    - prepositional phrase (around his desk, in the room)
    - postpositional phrase (jp: mise ni, ie kara, hashi de)
    - circumpositional phrase (from now on, de: Von mir aus)
  - noun phrase (the big man)
  - verb phrase (reads books)
  - …

Phrases

- Syntactic function
  - subject (**The big man** gave me a book yesterday.)
  - direct object (The big man gave me a **book** yesterday.)
  - indirect object (the big man gave **me** a book yesterday.)
  - adverbial phrase of…
    - …time (the big man gave me a book **yesterday**.)
    - …place (the big man gave me a book **at the library**.)
    - …
- Logico-semantic functions
  - semantic role [Fillmore 68] predicat/argument

Phrases

- Logico-semantic functions
  - semantic role [Fillmore 68]

Thematic Role	Definition
AGENT	The volitional causer of an event
EXPERIENCER	The experiencer of an event
FORCE	The non-volitional causer of the event
THEME	The participant most directly affected by an event
RESULT	The end product of an event
CONTENT	The proposition or content of a propositional event
INSTRUMENT	An instrument used in an event
BENEFICIARY	The beneficiary of an event
SOURCE	The origin of the object of a transfer event
GOAL	The destination of an object of a transfer event

Phrases

- Logico-semantic functions (of the predicate/verb)
  - semantic role [Fillmore 68]

Thematic Role	Example
AGENT	***The waiter*** spilled the soup.
EXPERIENCER	***John*** has a headache.
FORCE	***The wind*** blows debris from the mall into our yards.
THEME	Only after Benjamin Franklin broke ***the ice***...
RESULT	The city built ***a regulation-size baseball diamond***...
CONTENT	Mona asked ***"You met Mary Ann at a supermarket?"***
INSTRUMENT	He poached catfish, stunning them ***with a shocking device***...
BENEFICIARY	Whenever Ann Callahan makes hotel reservations ***for her boss***...
SOURCE	I flew in ***from Boston***.
GOAL	I drove ***to Portland***.

Phrases

- Logico-semantic functions (examples)
  - John (agent) broke the window (theme).
  - John (agent) broke the window (theme) with a rock (instr.).
  - The rock (instrument) broke the window (theme).
  - The window (theme) broke.
  - The window (theme) was broken by John (agent).

AMBIGUITY

Ambiguity

- A pervasive phenomenon in natural languages (NL)
  - a fundamental property of linguistic expressions
- A mean of flexibility & usability for NLs
  - it cannot be eliminated
- Most of the time humans do not see it
  - we share word knowledge, “common sense”
- On purpose, conscious uses
  - songs, poetry, humor, jokes, advertisements
- Accidental, unconscious uses
  - may lead to mistakes, errors, accidents when not detected
  - may lead to clarification sub-dialogues in conversations when detected

Ambiguities in English

Hirst, G. (1992) *Semantic interpretation and the resolution of ambiguity*

- A rough classification
  - Lexical ambiguities
    - Polysemy
    - Homophony
    - Categorical ambiguity
  - Structural ambiguities
    - Attachment problem
    - Gap finding & filling
    - Analytical ambiguity
  - any of those has implication at the semantic and pragmatic level

English: lexical ambiguity

- Polysemy
  - several related “meanings” associated to a string (sequence of letters [word, term], sequence of phonemes)
- Homonymy
  - several non-related “meanings” associated to a string (sequence of letters [word, term], sequence of phonemes)
    - written: homographs
    - spoken: homophones
- Categorical ambiguity
  - several syntactic categories associated to a string

Polysemy

Several **related** "meanings"

- The verb ***open***
  - unfolding,
  - expanding,
  - revealing,
  - moving to an open position,
  - making openings in

Homonymy

several non-related "meanings"

- Homographs (written)
  - ***row*** as a noun
    - a number of people or things in a more or less straight line
    - a noisy acrimonious quarrel
  - ***bark*** as a noun
    - the sharp explosive cry of certain animals
    - the tough, protective outer sheath of the trunk, branches, and twigs of a tree
    - a sailing ship …
- Homophones (spoken)
  - ***four*** as a noun
    - cardinal number
  - ***fore*** as an adjective
    - situated or placed in front

Polysemy + Homonymy

- the word ***right***
  - polysemy
    - senses concerning correctness & righteousness
  - homonymy
    - + senses concerning the right-hand side

- Linked also with metaphor

today's metaphor may be tomorrow's polysemy or homonymy

  - a person's *mouth* and the *mouth* of a river

Categorical ambiguity

several syntactic categories

- The string ***sink***
  - a noun
    - describing a plumbing fixture
  - a verb
    - meaning become submerged

- *It is mainly a problem of parsing*

Categorical ambiguity

several syntactic categories

- Orthogonal to the other types
  - the string ***respect***
    - categorical and polysemous
    - … noun and verb meanings are related
  - the string ***sink***
    - is categorical and homonymous
    - … noun and verb meanings are not related

Structural ambiguity

- Attachment
  - There is more than one node to which a particular syntactic constituent may be attached
- Gap finding and filling
  - A moved constituent has to be returned to its pre-transformational starting position, and there is more than one place it might go
- Analytical ambiguity
  - The nature of the constituent is itself in doubt, that is, when there is more than one possible analysis for it

Attachment

- Prepositional phrases may have more than one noun phrase available to attach it to (as well as possibly a verb)
- Example
  - the door near the stairs with the “member only” sign
    - the sign is one the door
    - the sign is on the stairs

Attachment

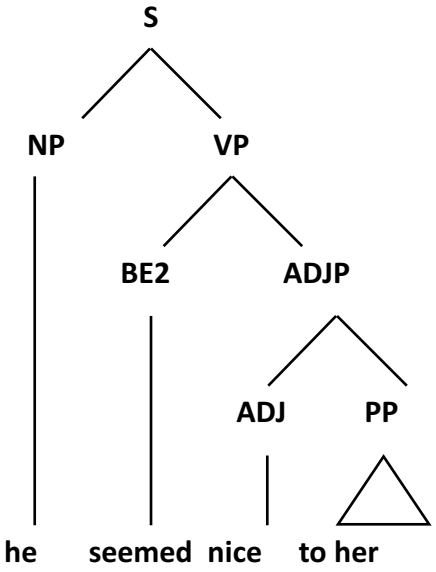
- A prepositional phrases may have more than one noun phrase available to attach it to …
- Example
  - I saw the man in the park
    - in the park, I saw the man
    - I saw the (man in the park)
  - I saw the man in the park with a telescope
    - in (the park with a telescope)
      - in (the park with a telescope), I saw the man
      - I saw the (man in (the park with a telescope))
    - in (the park) ; with (a telescope)
      - in (the park)$_{location}$, with (a telescope)$_{mean}$, I saw the man
      - with (a telescope)$_{mean}$, I saw the (man in (the park)$_{location}$)
      - I saw the (man in (the park with (a telescope)$_{attribute}$)$_{location}$)

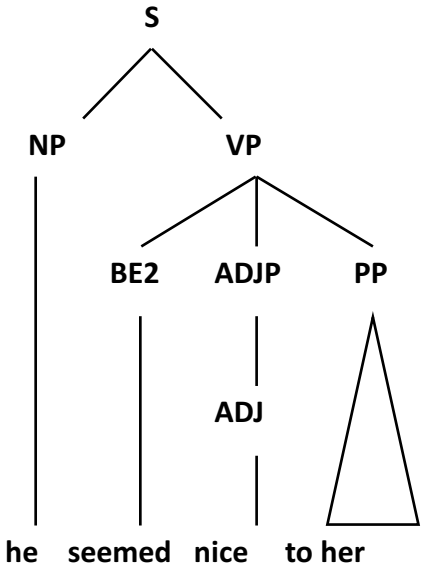
Attachment

- Relative clauses have similar attachment ambiguity
- Example
  - The door near the stairs that had the “Members Only” sign had tempted Nadia.
    - the sign is one the door
    - the sign is on the stairs

Attachment

- Prepositional phrases can also be attached to an adjective phrase
- Example
  - He seemed nice to her
    - He seemed to act nicely towards her
      - Attachment to the adjective phrase
    - He seemed to her to be nice
      - Attachment to the verbal phrase





Attachment

- A sentence contains a subsentence, both may contain place for the attachment of a prepositionnal phrase or adverb
- Example
  - Ross said that Nadia had taken the cleaning out *yesterday*
    - said yesterday
    - taken out yesterday

Attachment

- An adverbial may modify the sentence verb or the whole sentence
- Example: Happily, Nadia cleaned up the mess Ross had left
  - happily could be attached to the sentence
    - meaning that the event was a fortunate occurrence,
  - or it could be attached to the VP
    - meaning that Nadia was quite happy to clean up the mess

Attachment

- Adverbial placed between two clauses can be attached to the verb of either
- Exemples
  - The lady you met *now and then* came to visit us
    - We were visited by the lady you met now and then
    - We were visited now and then by the lady you met
  - The friends you praise *sometimes* deserve it
    - Sometimes the friend you praise deserve it
    - The friends you sometimes praise deserve it

Gap finding and filling

- A moved constituent has to be returned to its pre-transformational starting position, and there is more than one place it might go
- Example: Those are the boys that the police debated △ about fighting △.
  - The police debated with the boys on the topic of fighting
  - The police debated (among themselves) about fighting the boys

Analytical ambiguity

- The nature of the constituent is itself in doubt, that is, when there is more than one possible analysis for it
- Example
  - "You can have the music box that's in the closet or the one that's on the table" said Ross. "**I want the music box △ on the table**" said Nadia.
    - I want the music box **that is** on the table
  - "I put the music box on the mantelpiece. is that okay?" asked Ross. "No," said Nadia, "**I want the music box △ on the table**."
    - I want the music box **to be** on the table

Analytical ambiguity

- Present participle or adjective?
- Example
  - Ross and Nadia are singing madrigals
  - Pen and pencils are writing implements
- Ambiguity
  - They are cooking apples
    - What are they doing?
    - What are those apples?

Analytical ambiguity

- Present participle or noun?
  - Distinguishing between a present participle or a noun
- Example
  - We discussed running
    - We discussed the sport of running
    - We discussed the possibility of our running

Analytical ambiguity

- What is the subject of the supplementive?
  - Participles and adjectivals at the end of a clause. A subject and an object can be the subject of a supplementative.
- Example
  - We meet him leaving the room
    - we were leaving the room
    - he was leaving the room
  - I saw him going home
    - I was going home
    - he was going home

Analytical ambiguity

- Supplementive, restrictive relative or verb complement?
  - the participle, instead of being a supplementive, could be attached to the object NP either as a reduced restrictive relative clause or as a verb complement
- Example
  - The manager approached the boy smoking a cigar
    - the manager is smoking (supplementive)
    - the boy is smoking (relative clause)

Analytical ambiguity

- How is the predicate formed?
  - different structures that can underlies sentences of the form

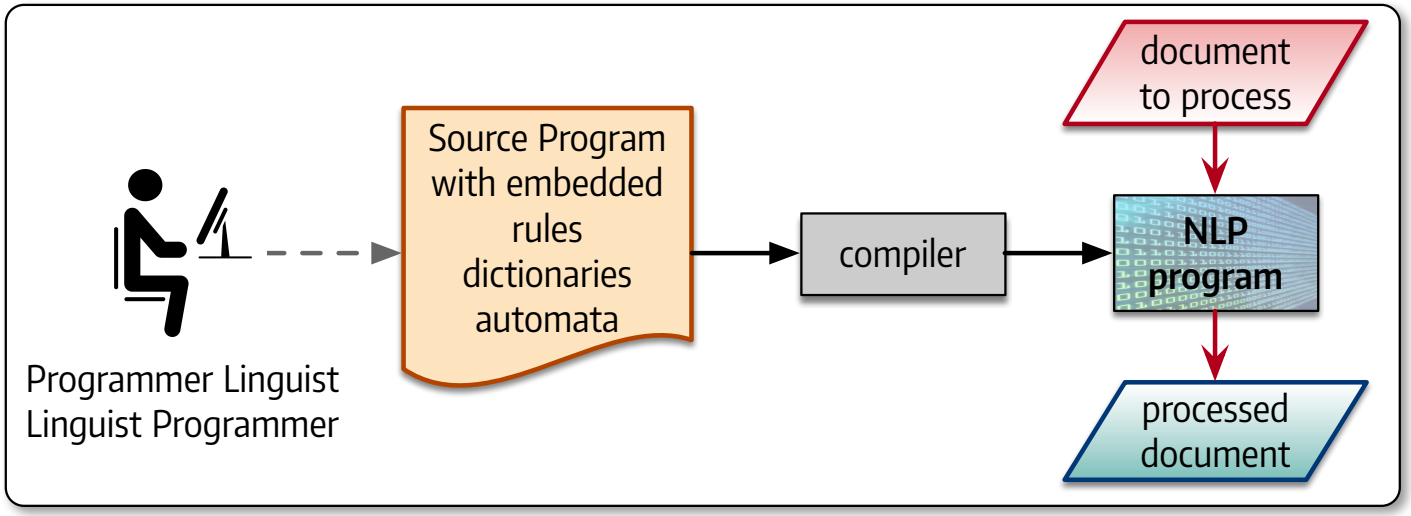
    NP be ADJ to V

- Examples
  - The chicken is ready to eat
    - the chicken will eat
    - the chicken will be eaten

NLP APPROACHES IN A NUTSHELL

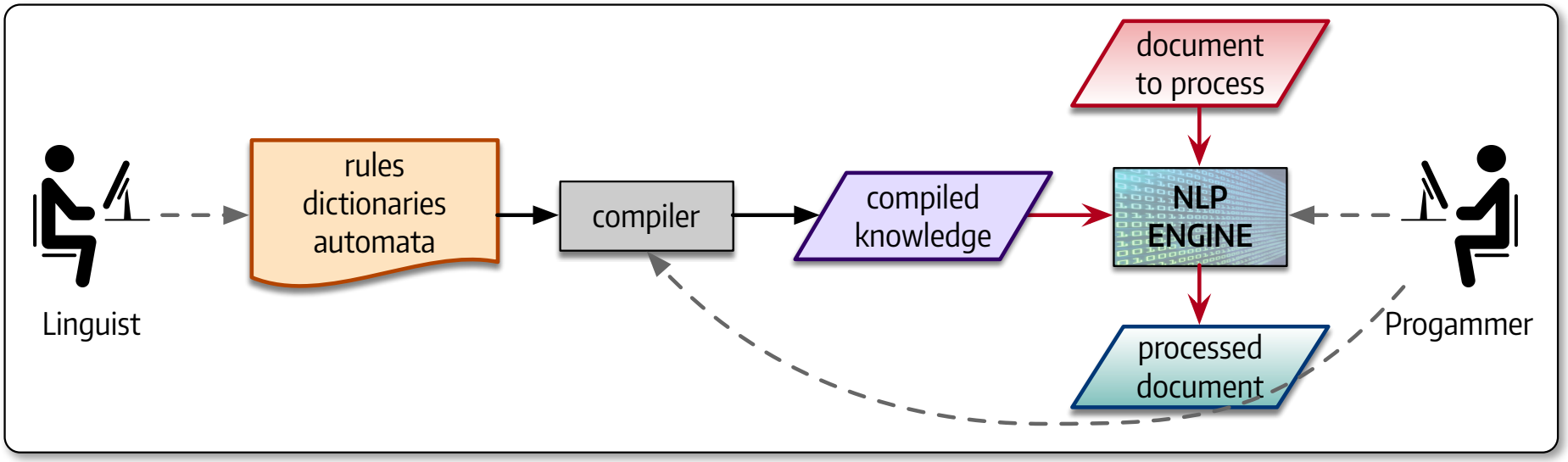
Early approach

- Linguistic hacking
  - rules, dictionaries (knowledge) are (is) encoded within the programs



Expert approach

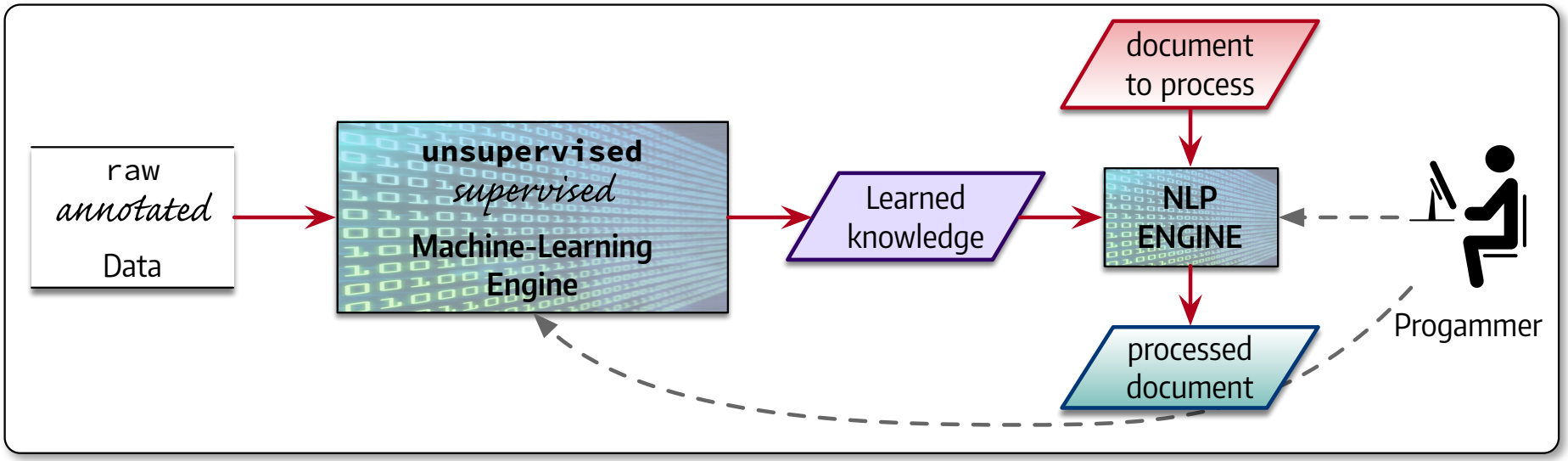
- Formal linguistics and compilation
  - rules, dictionaries (knowledge) are (is) encoded by specialists using grammars and automata
  - rules (knowledge) are (is) then compiled into an internal format
  - the internal format act as an input for an NLP-engine



  - analogy with Java: source code (rules, dicos) into bytecode (compiled rules, dicos) executed by a virtual machine (NLP-engine)

Empirical approach

- Machine learning
  - machine-learning approaches using statistical inference to learn automatically rules, knowledge, through the analysis of large, raw or annotated, corpora
  - machine-learning approaches using artificial neural networks



WORDS
WORDS IN CONTEXT

What can be done?

- Stemming/Morphological segmentation
- Lemmatization
- Counting individual:
  - forms
  - lemmas
- Clouding over a document/set of documents

Word Cloud for "Pride & Pejudice" J. Austen

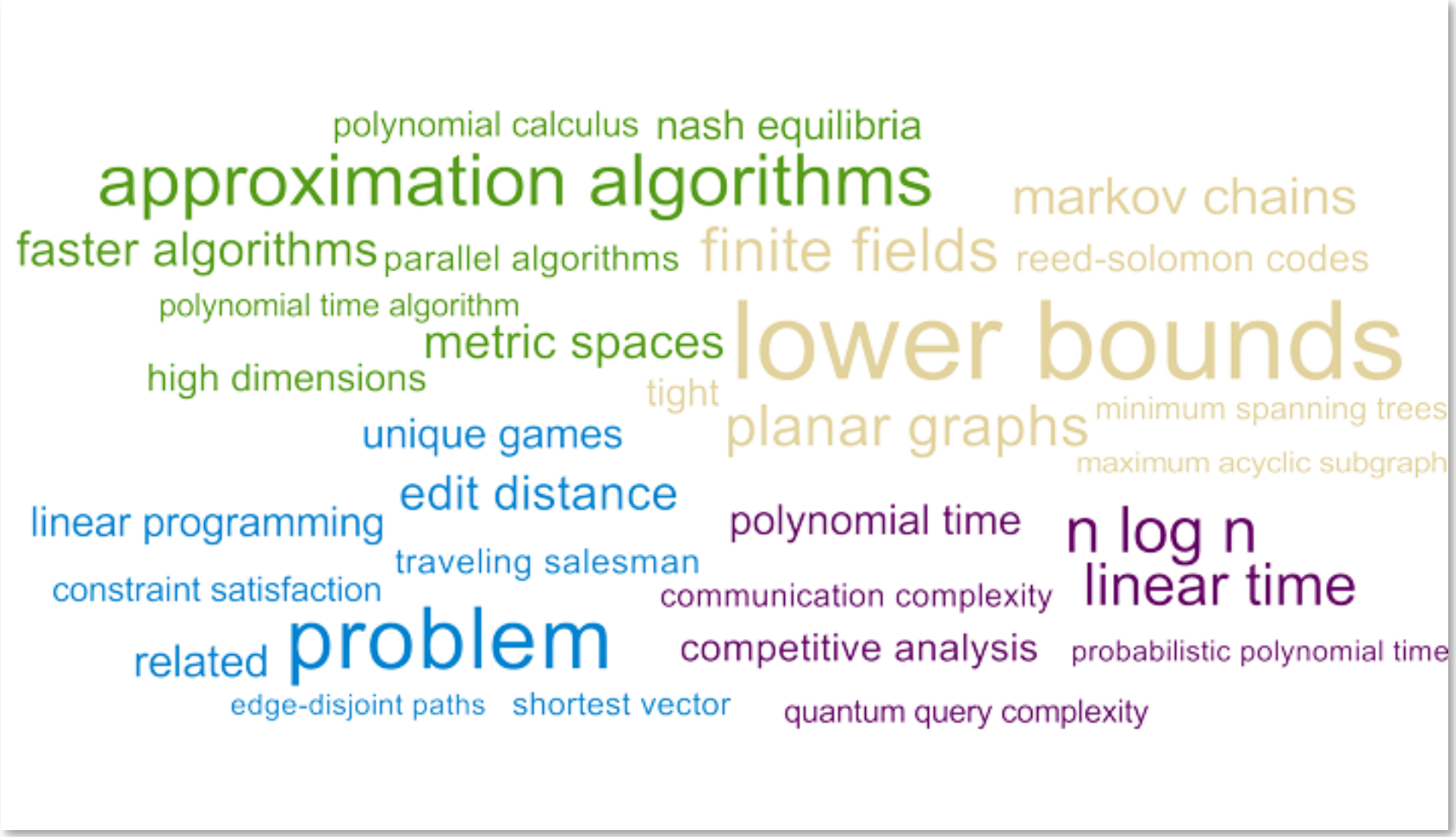
Wordle (wordle.net)

- placement and color are random

- Random
  - wordle.net

Semantic

- wordcloud.cs.arizona.edu



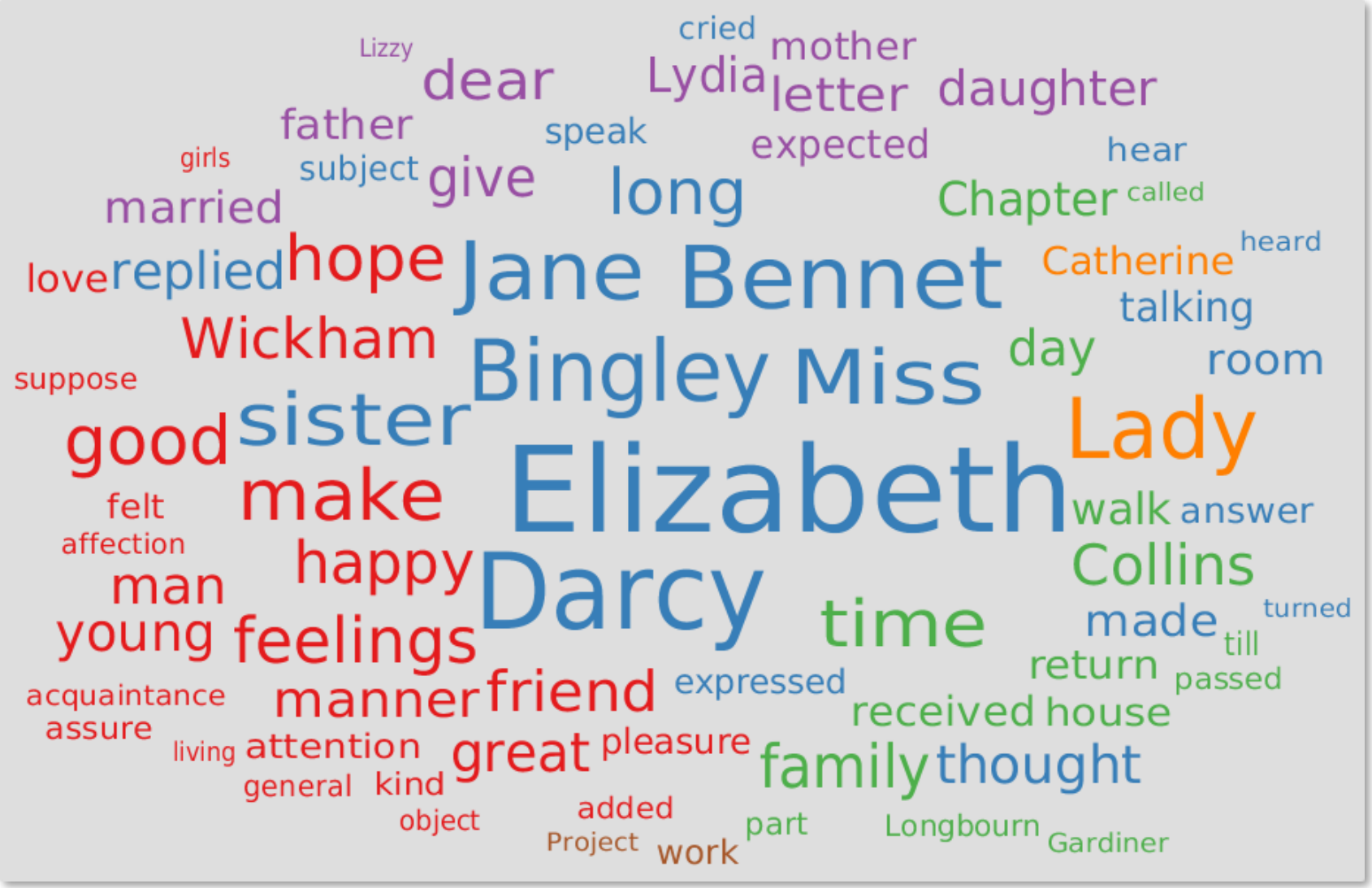
Word Cloud for "Pride & Pejudice" J. Austen

- with WordCloud (wordcloud.cs.arizona.edu)
  - sorted by rank

Word Cloud for "Pride & Pejudice" J. Austen

with WordCloud (wordcloud.cs.arizona.edu)

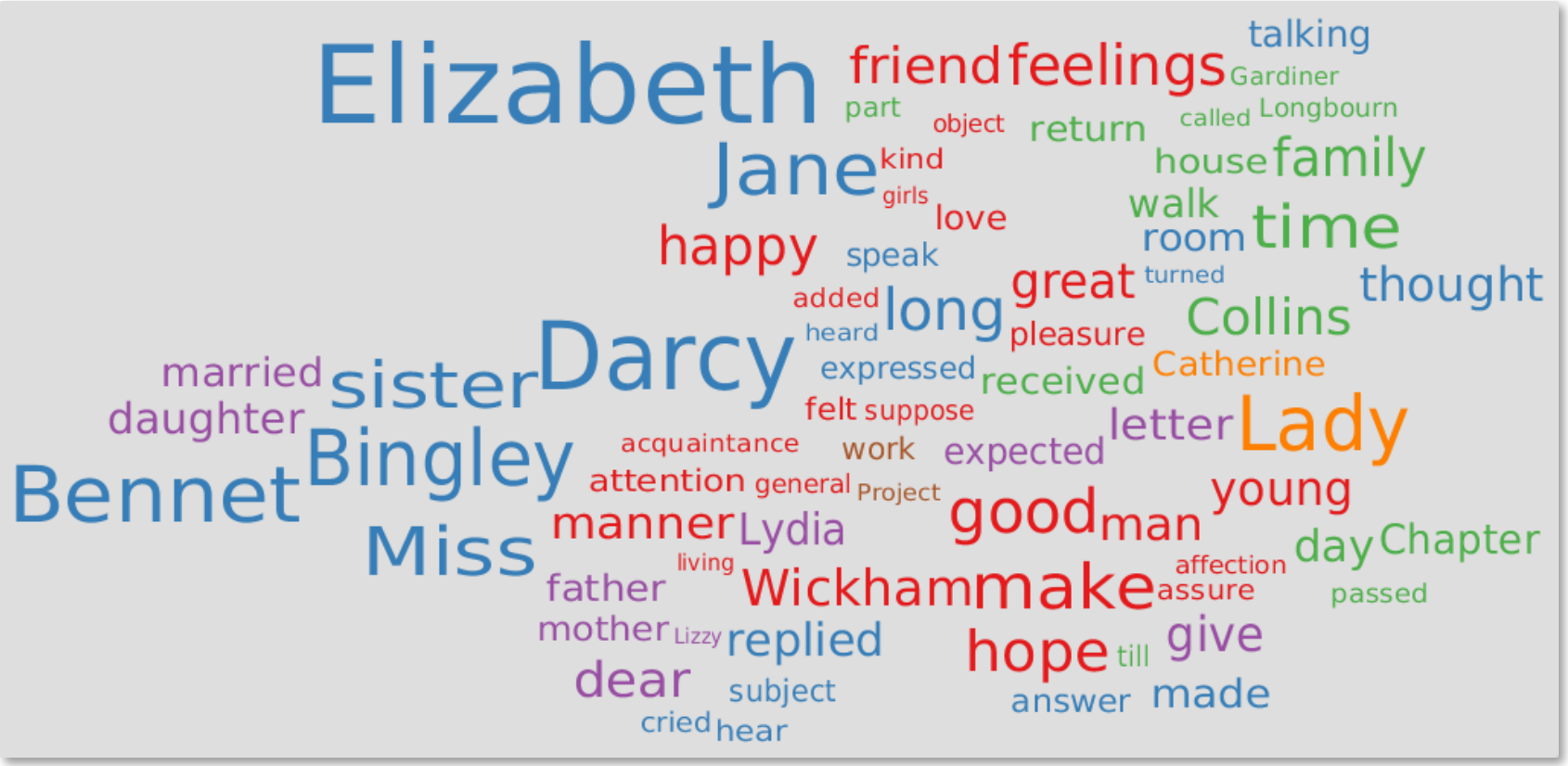
- Seam Carving



Word Cloud for "Pride & Pejudice" J. Austen

with WordCloud (wordcloud.cs.arizona.edu)

- cycle cover
  - an approximation algo. to compute semantic word clouds



Word Cloud for "Pride & Pejudice" J. Austen

- with WordCloud (wordcloud.cs.arizona.edu)
  - force-directed
    - multidimensional scaling

Context

COLLOCATIONS

Definition

- Word, words, appearing in the neighborhood of a “target” word or sequence of words
- Collocations can be in a
  - syntactic relation (such as verb–object: 'make' and 'decision')
  - lexical relation (such as antonymy)
  - or they can be in no linguistically defined relation
- Processing of collocations involves a number of parameters, the most important of which is the measure of association, which evaluates whether the co-occurrence is purely by chance or statistically significant using measures of association

With a simple text editor

- A little concordancer good for close reading
  - jEdit (jedit.org)
  - SublimText (sublimetext.com)
  - TextWangler (barebones.com)

Example with TextWangler

TextWrangler File Edit Text View Search Go Window #! Help

Pride and Prejudice.txt

~/Documents/ Recherche/Articles/Articles-2017/EcoleEte Humanités Numériques/Pride and Prejudice.txt

Chapter 2

Mr. Bennet was among the earliest of those who waited on Mr. Bingley. He had always intended to visit him, though to the last always assuring his wife that he should not go; and till the evening after the visit was paid she had no knowledge of it. It was then disclosed in the following manner. Observing his second daughter employed in trimming a hat, he suddenly addressed her with:
"I hope Mr. Bingley will like it, Lizzy."
"We are not in a way to know what Mr. Bingley likes," said her mother resentfully, "since we are not to visit."
"But you forget, mamma," said Elizabeth, "that we shall meet him at the assemblies, and that Mrs. Long promised to introduce him."
"I do not believe Mrs. Long will do any such thing. She has two nieces of her own. She is a selfish, hypocritical woman, and I have no opinion of her."
"No more have I," said Mr. Bennet; "and I am glad to find that you do not depend on her serving you."
Mrs. Bennet deigned not to make any reply, but, unable to contain herself, began scolding one of her daughters.
"Don't keep coughing so, Kitty, for Heaven's sake! Have a little compassion on my nerves. You tear them to pieces."
"Kitty has no discretion in her coughs," said her father; "she times them ill."
"I do not cough for my own amusement," replied Kitty fretfully. "When is your next ball to be, Lizzy?"
"To-morrow fortnight."
"Aye, so it is," cried her mother, "and Mrs. Long does not come back till the day before; so it will be impossible for her to introduce him, for she will not know him herself."
"Then, my dear, you may have the advantage of your friend, and introduce Mr. Bingley to her."
"Impossible, Mr. Bennet, impossible, when I am not acquainted with him myself; how can you be so teasing?"
"I honour your circumspection. A fortnight's acquaintance is certainly very little. One cannot know what a man really is by the end of a fortnight. But if we do not venture somebody else will; and after all, Mrs. Long and her nieces must stand their chance; and, therefore, as she will think it an act of kindness, if you decline the office, I will take it on myself."
The girls stared at their father. Mrs. Bennet said only, "Nonsense, nonsense!"
"What can be the meaning of that emphatic exclamation?" cried he. "Do you consider the forms of introduction, and the stress that is laid on them, as nonsense? I cannot quite agree with you there. What say you, Mary? For you are a young lady of deep reflection, I know, and read great books and make extracts."
Mary wished to say something sensible, but knew not how.
"While Mary is adjusting her ideas," he continued, "let us return to Mr. Bingley."
"I am sick of Mr. Bingley," cried his wife.
"I am sorry to hear that; but why did not you tell me that before? If I had known as much this morning I certainly would not have called on him. It is very unlucky; but as I have actually paid the visit, we cannot escape the acquaintance now."
The astonishment of the ladies was just what he wished; that of Mrs. Bennet

L: 53 C: 31 Text File Unicode (UTF-8) Unix (LF) Saved: 20/06/2017 15:55:53 9 / 1 / 0 100%

Find

Find: Elizabeth

Replace:

Next Previous Find All Replace Replace All Replace & Find

Matching: Case sensitive Entire word Grep

Search in: Selected text only Wrap around

Pride and Prejudice.txt — Search Results (Elizabeth)

0 Errors 0 Warnings 635 Notes Open

635 occurrences of "Elizabeth" found in 1 files.

Pride and Prejudice.txt 635 occurrences found

File Pride and Prejudice.txt; Line 53: "But you forget, mamma," said Elizabeth, "that we shall meet him at the assemblies, and that Mrs. Long prom
File Pride and Prejudice.txt; Line 89: Elizabeth Bennet had been obliged, by the scarcity of gentlemen, to sit down for two dances; and during part
File Pride and Prejudice.txt; Line 95: "Which do you mean?" and turning round he looked for a moment at Elizabeth, till catching her eye, he withd
File Pride and Prejudice.txt; Line 96: Mr. Bingley followed his advice. Mr. Darcy walked off; and Elizabeth remained with no very cordial feelings to
File Pride and Prejudice.txt; Line 97: The evening altogether passed off pleasantly to the whole family. Mrs. Bennet had seen her eldest daughter
File Pride and Prejudice.txt; Line 109: When Jane and Elizabeth were alone, the former, who had been cautious in her praise of Mr. Bingley before,
File Pride and Prejudice.txt; Line 111: "He is also handsome," replied Elizabeth, "which a young man ought likewise to be, if he possibly can. His ch
File Pride and Prejudice.txt; Line 119: Elizabeth listened in silence, but was not convinced; their behaviour at the assembly had not been calculate
File Pride and Prejudice.txt; Line 132: Lady Lucas was a very good kind of woman, not too clever to be a valuable neighbour to Mrs. Bennet. They
File Pride and Prejudice.txt; Line 149: "That is very true," replied Elizabeth, "and I could easily forgive his pride, if he had not mortified mine."

~/Documents/ Recherche/Articles/Articles-2017/EcoleEte Humanités Numériques/Pride and Prejudice.txt

he should not go; and till the evening after the visit was paid she had no knowledge of it. It was then disclosed in the following manner. Observing his second daughter employed in trimming a hat, he suddenly addressed her with:
"I hope Mr. Bingley will like it, Lizzy."
"We are not in a way to know what Mr. Bingley likes," said her mother resentfully, "since we are not to visit."
"But you forget, mamma," said Elizabeth, "that we shall meet him at the assemblies, and that Mrs. Long promised to introduce him."
"I do not believe Mrs. Long will do any such thing. She has two nieces of her own. She is a selfish, hypocritical woman, and I have no opinion of her."
"No more have I," said Mr. Bennet; "and I am glad to find that you do not depend on her serving you."
Mrs. Bennet deigned not to make any reply, but, unable to contain herself, began scolding one of her daughters.
"Don't keep coughing so, Kitty, for Heaven's sake! Have a little compassion on my nerves. You tear them to pieces."
"Kitty has no discretion in her coughs," said her father; "she times them ill."
"I do not cough for my own amusement," replied Kitty fretfully. "When is your next ball to be, Lizzy?"
"To-morrow fortnight."
"Aye, so it is," cried her mother, "and Mrs. Long does not come back till the day before; so it will be impossible for her to introduce him, for she will not know him herself."
"Then, my dear, you may have the advantage of your friend, and introduce Mr.

L: 53 C: 31 Text File Unicode (UTF-8) Unix (LF) Saved: 20/06/2017 15:55:53 9 / 1 / 0 100%

With a real concordancer

- AntConc (laurenceanthony.net)

- WordSmith Tools (lexically.net)
  - commercial; Windows

- CasualConc (GoogleSite)
  - free; Mac OS X
  - a lot of functionalities

Example with CasualConc

CasualConc

File | Concord | Word Count | Collocation | Cluster | File Info

Simple | File | Text

Elizabeth | Search | Span 40 40 | Sort Choice | Sort | Context

Context Word | L1-Key-R1

Kwic - 635 found in 1 files

		File Name
370	er companions all the way to Longbourn. Elizabeth listened as little as she could, but th	Pride and Prejudice.txt
371	over, they should proceed to Longbourn. Elizabeth was surprised, however, that Wickham sh	Pride and Prejudice.txt
372	r chair, not knowing which way to look. Elizabeth found herself quite equal to the scene,	Pride and Prejudice.txt
373	ve been but for you, dearest, loveliest Elizabeth! What do I not owe you! You taught me a	Pride and Prejudice.txt
374	ons; and in spite of his being a lover, Elizabeth really believed all his expectations of	Pride and Prejudice.txt
375	on with you," said Catherine and Lydia. Elizabeth accepted their company, and the three y	Pride and Prejudice.txt
376	This information made Elizabeth smile, as she thought of poor Miss Bing	Pride and Prejudice.txt
377	ing run away with by his feelings, made Elizabeth so near laughing, that she could not us	Pride and Prejudice.txt
378	and the general pause which ensued made Elizabeth tremble lest her mother should be expos	Pride and Prejudice.txt
379	event of such happy promise as to make Elizabeth hope that by the following Christmas sh	Pride and Prejudice.txt
380	llowed them to see him before they met. Elizabeth, however astonished, was at least more	Pride and Prejudice.txt
381	"Miss Elizabeth Bennet!" repeated Miss Bingley. "I am a	Pride and Prejudice.txt
382	"Miss Elizabeth Bennet."	Pride and Prejudice.txt
383	is was all lost upon me. I thought Miss Elizabeth Bennet looked remarkably well when she	Pride and Prejudice.txt
384	eously married, but that you, that Miss Elizabeth Bennet, would, in all likelihood, be so	Pride and Prejudice.txt
385	I think she will. She is now about Miss Elizabeth Bennet's height, or rather taller."	Pride and Prejudice.txt
386	s opportunity of soliciting yours, Miss Elizabeth, for the two first dances especially, a	Pride and Prejudice.txt
387	"My dear Miss Elizabeth, I have the highest opinion in the worl	Pride and Prejudice.txt
388	"I know not, Miss Elizabeth," said he, "whether Mrs. Collins has ye	Pride and Prejudice.txt
389	"Believe me, my dear Miss Elizabeth, that your modesty, so far from doing y	Pride and Prejudice.txt
390	t. Only let me assure you, my dear Miss Elizabeth, that I can from my heart most cordiall	Pride and Prejudice.txt
391	Mrs. Bennet rang the bell, and Miss Elizabeth was summoned to the library.	Pride and Prejudice.txt
392	"Yes, Miss Elizabeth, you will have the honour of seeing Lad	Pride and Prejudice.txt

"She has nothing, in short, to recommend her, but being an excellent walker. I shall never forget her appearance this morning. She really looked almost wild."

"She did, indeed, Louisa. I could hardly keep my countenance. Very nonsensical to come at all! Why must she be scampering about the country, because her sister had a cold? Her hair, so untidy, so blowsy!"

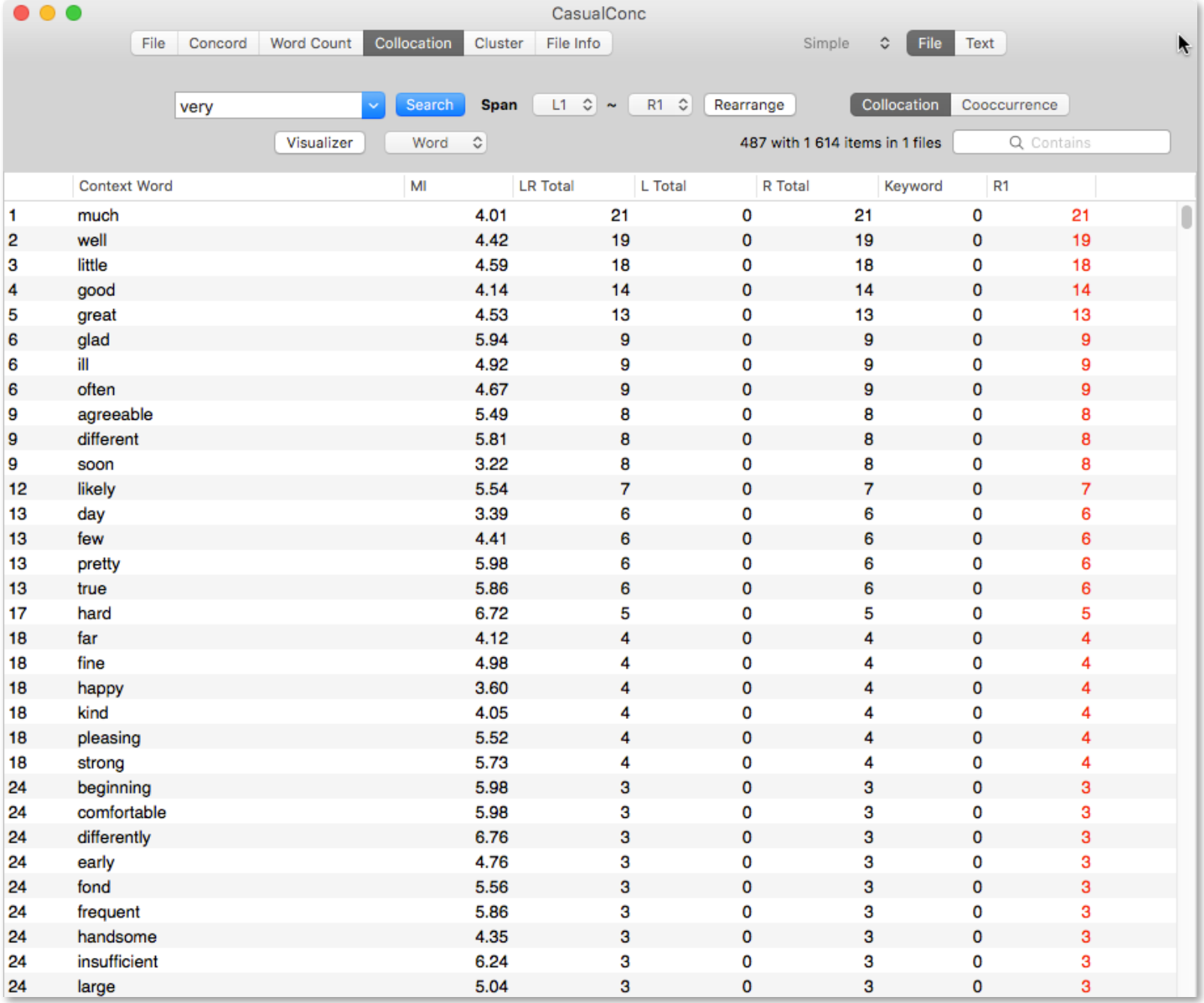
"Yes, and her petticoat; I hope you saw her petticoat, six inches deep in mud, I am absolutely certain; and the gown which had been let down to hide it not doing its office."

"Your picture may be very exact, Louisa," said Bingley; "but this was all lost upon me. I thought Miss **Elizabeth** Bennet looked remarkably well when she came into the room this morning. Her dirty petticoat quite escaped my notice."

"You observed it, Mr. Darcy, I am sure," said Miss Bingley; "and I am inclined to think that you would not wish to see your sister make such an exhibition."

"Certainly not."

Collocation with CasualConc: "very" L1



	Context Word	MI	LR Total	L Total	R Total	Keyword	R1
1	much	4.01	21	0	21	0	21
2	well	4.42	19	0	19	0	19
3	little	4.59	18	0	18	0	18
4	good	4.14	14	0	14	0	14
5	great	4.53	13	0	13	0	13
6	glad	5.94	9	0	9	0	9
6	ill	4.92	9	0	9	0	9
6	often	4.67	9	0	9	0	9
9	agreeable	5.49	8	0	8	0	8
9	different	5.81	8	0	8	0	8
9	soon	3.22	8	0	8	0	8
12	likely	5.54	7	0	7	0	7
13	day	3.39	6	0	6	0	6
13	few	4.41	6	0	6	0	6
13	pretty	5.98	6	0	6	0	6
13	true	5.86	6	0	6	0	6
17	hard	6.72	5	0	5	0	5
18	far	4.12	4	0	4	0	4
18	fine	4.98	4	0	4	0	4
18	happy	3.60	4	0	4	0	4
18	kind	4.05	4	0	4	0	4
18	pleasing	5.52	4	0	4	0	4
18	strong	5.73	4	0	4	0	4
24	beginning	5.98	3	0	3	0	3
24	comfortable	5.98	3	0	3	0	3
24	differently	6.76	3	0	3	0	3
24	early	4.76	3	0	3	0	3
24	fond	5.56	3	0	3	0	3
24	frequent	5.86	3	0	3	0	3
24	handsome	4.35	3	0	3	0	3
24	insufficient	6.24	3	0	3	0	3
24	large	5.04	3	0	3	0	3

Collocation with CasualConc: “very” L1

- Word cloud

Credits for this section

Introduction to Information Retrieval

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze

https://nlp.stanford.edu/IR-book/

STEMMING AND LEMMATIZATION

Goal

- Reduce inflectional forms and sometimes derivationally related form of a word (an occurrence in a text)…
  - am, are, is
- …to a common base form
  - be

Stemming

- Crude heuristic process, relying on rules,…
  - chop of the ends of words
  - removal of derivational affixes
- …empirically very effective

Rule	Example
SSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S → Ø	cats → cat
(m>1) EMENT → Ø	replacement → replac (but not cement → c)

Stemming

- Comparison of three stemming algorithms

Sample text	Such an **analysis** can reveal **features** that are not **easily** **visible** from the variations in the **individual** genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation
Lovins Stemmer	Such an **analys** can reve **featur** that ar not **eas** **vis** from th vari in th **individu** gen and can lead to a pictur of expres that is more biolog transpar and access to interpret
Porter Stemmer	Such an **analys** can reveal **featur** that are not **easi** **visibl** from the variat in the **individu** gens and can lead to a pictur of expres that is more biolog transpar and access to interpret
Paice Stemmer	Such an **analys** can rev **feat** that are not **easy** **vis** from the vary in the **individ** gen and can lead to a pict of express that is mor biolog transp and access to interpret

Lemmatizer

- Perform a full morphological analysis to accurately identify the lemma for each word
  - lemma
  - morphological properties

- *next slide morphological analysis from Xerox at https://open.xerox.com/Services/fst-nlp-tools*

analysis	<analyze>is +Noun+Sg
can	<can> +Aux
	<can> +Noun+Sg
	<can> +Verb+Pres+Non3sg
reveal	<reveal> +Verb+Pres+Non3sg
features	<feature> +Noun+Pl
	<feature> +Verb+Pres+3sg
that	<that> +Conj+Sub
	<that> +Det+Sg
	<that> +Pron+NomObl+3P+Sg
	<that> +Pron+Rel+NomObl+3P+SP
	<that> +Adv
are	<be> +Verb+Pres+Pl
not	<not> +Adv
easily	<easy>ly} +Adv
visible	<visible> +Adj
from	<from> +Prep
the	<the> +Det+Def+SP
variations	<variation> +Noun+Pl

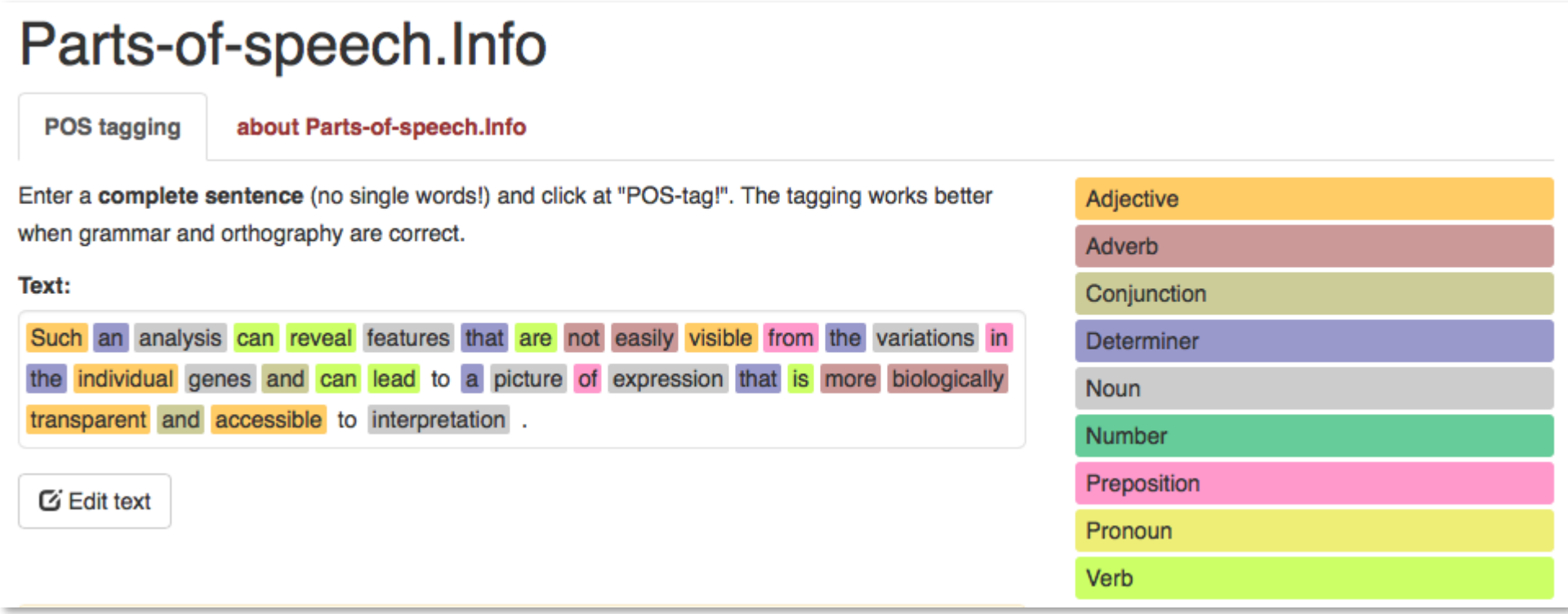
POS TAGGING

Implementation

- Normally done by a *sequence model* (HMM, CRM, MEMM/CMM)
  - A POS tag is to be assigned to each word
  - The model considers a local context of possible previous and following POS tags, the current word, neighboring words, and features of them (capitalized?, ends in -*ing*?)
  - Each such *feature* has a *weight*, and the evidence is combined, and the most likely sequence of tags (according to the model) is chosen
- Possibly several acceptable sequences of POS are produced to deal with ambiguity
- Different POS Taggers may use different POS sets
  - Penn Treebank, GRACE, Amalgam,…

Examples

- The Stanford POS Tagger



Examples

- Xerox POS Tagger

Such*+PREDET* an*+DET* analysis*+NOUN* can*+VAUX* reveal*+VI* features*+NOUN* that*+PRONREL* are*+VBPRES* not*+NOT* easily*+ADV* visible*+ADJ* from*+PREP* the*+DET* variations*+NOUN* in*+PREP* the*+DET* individual*+ADJ* genes*+NOUN* and*+COORD* can*+VAUX* lead*+VI* to*+PREP* a*+DET* picture*+NOUN* of*+PREP* expression*+NOUN* that*+PRONREL* is*+VBPRES* more*+QUANTCMP* biologically*+ADV* transparent*+ADJ* and*+COORD* accessible*+ADJ* to*+PREP* interpretation*+NOUN* .*+SENT*

NAMED-ENTITY RECOGNITION

What is a named-entity?

- Name
  - proper name, company name, museum, rivers, …
- Location
  - country, state, city,…
- Organization
  - NATO, DARPA, UN, …
- Products
- Medicine, Biology, Epidemiology, chemistry
  - disease, drugs, protein, DNA, RNA, cell line, cell type, chemical name, species
- Other types
  - film, research area,…

Examples

- Germany's representative to the European Union's veterinary committee Werner Zwingman said on Wednesday consumers should …
- IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.

Further reading

- A survey of named entity recognition and classification
  David Nadeau, Satoshi Sekine
  Journal of Linguisticae Investigationes 30:1 ; 2007
- *available on the Web*

PARSING

Definitions

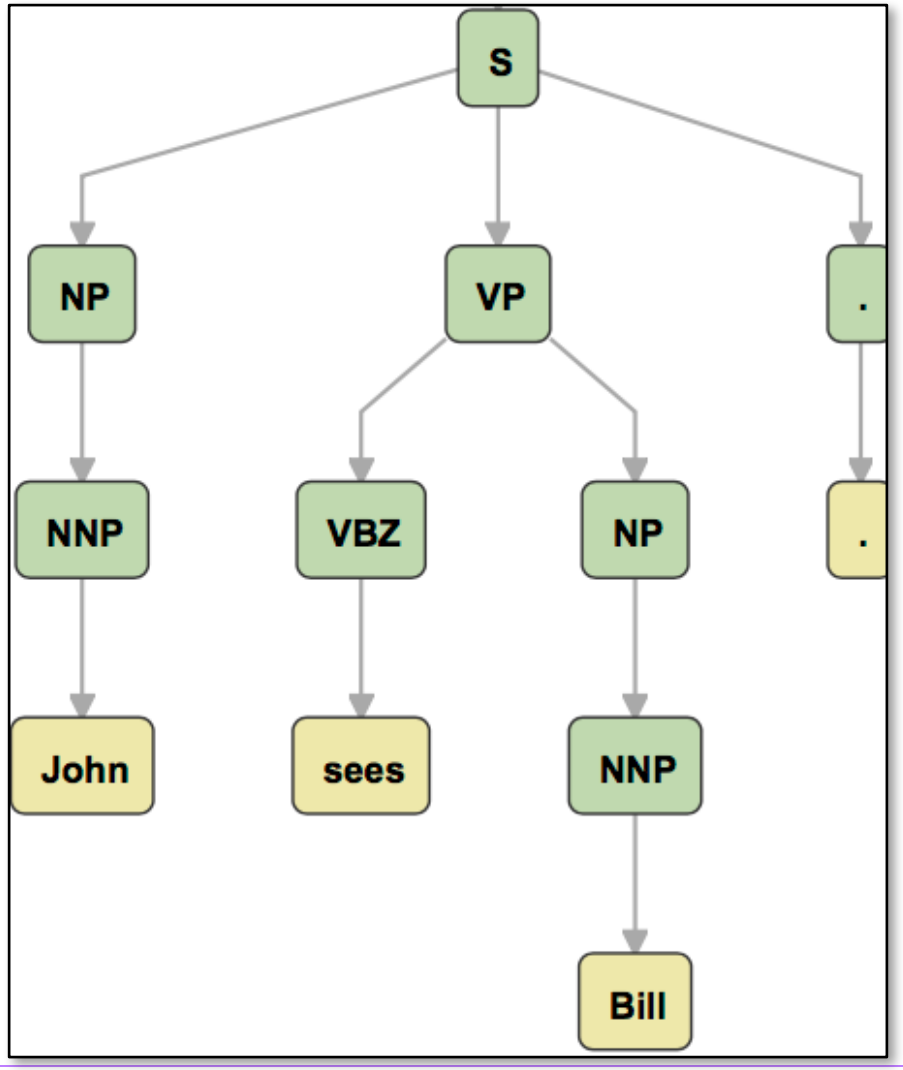
- Within computational linguistics “parsing” is used to refer to the formal analysis by a computer of a sentence or other string of words into its constituents, resulting in a parse tree showing their syntactic relation to each other, which may also contain semantic and other information
- Two main approaches
  - constituency parsing
  - dependency parsing

Constituency/Phrase Structure Parsing

- Principles
  - Text is broken into sub-phases (constituents)
  - Non-terminals in the tree are types of phrases, possibly with some additional labels
- Structure
  - Terminals are the words, with some additional labels
  - Edges are not labeled

Constituency Parsing

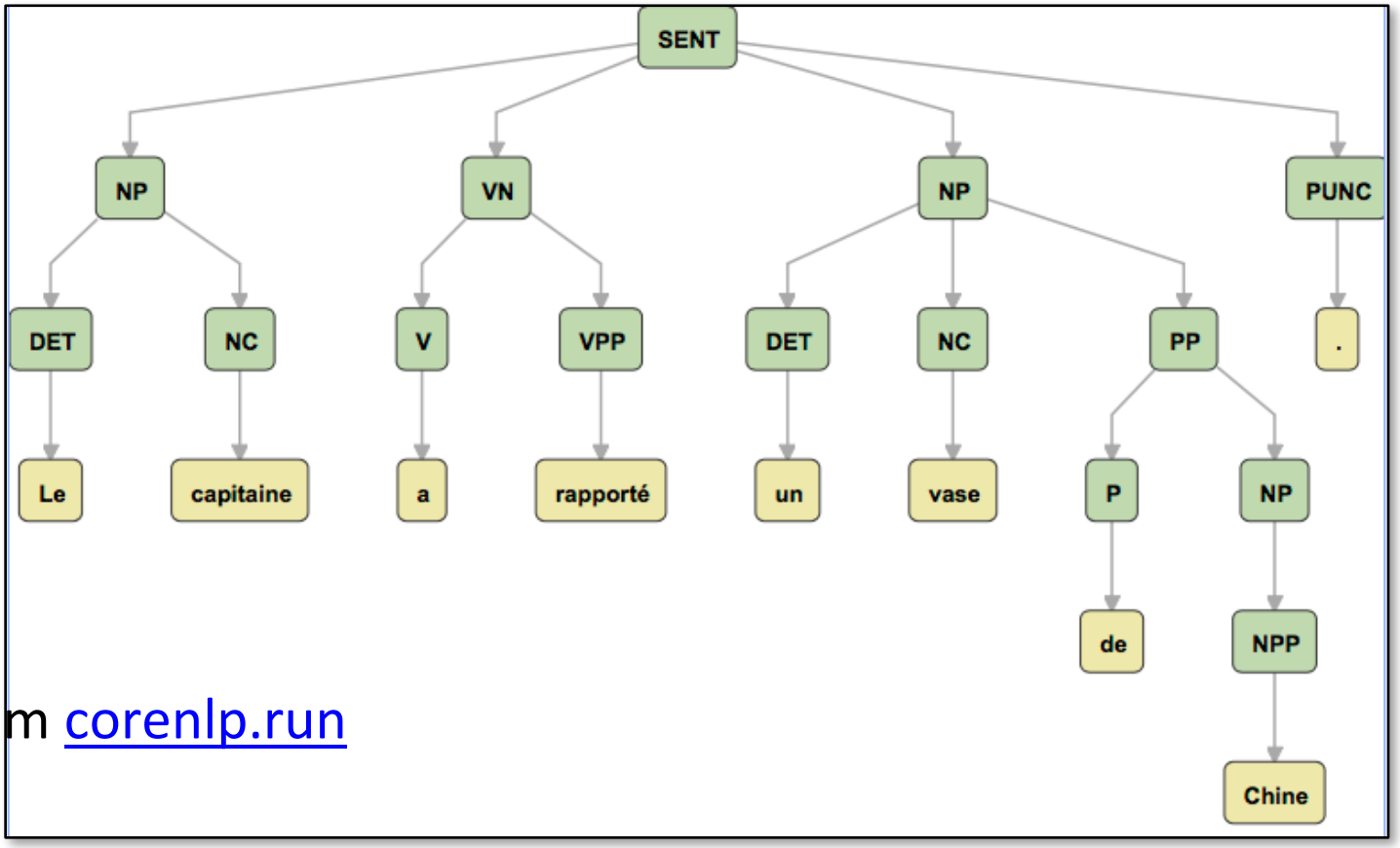
- Example of a "simple" parse tree
  - "John sees Bill."



  - from corenlp.run

Constituency Parsing

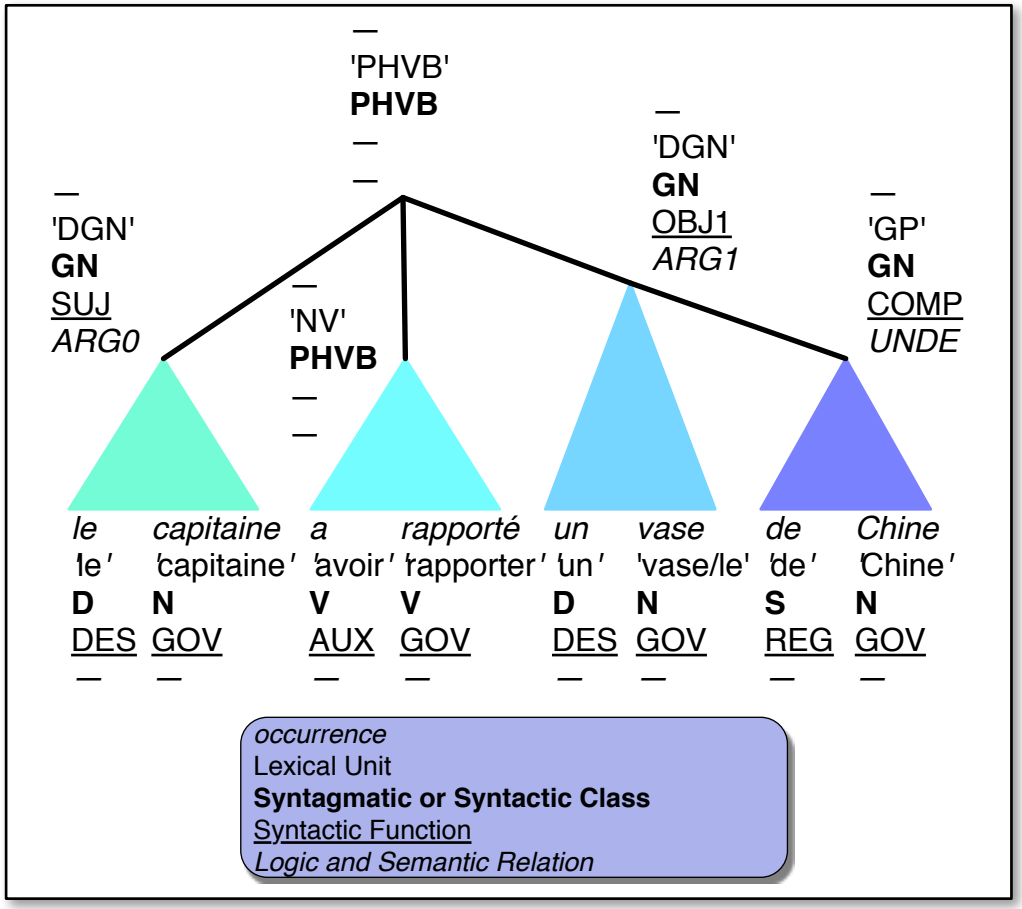
- Example of a “simple” parse tree
  - « Le capitaine a rapporté un vase de chine. »
    - the captain brought back a Chinese vase



    - from corenlp.run

Constituency Parsing

- Example of a multilevel structure
  - « Le capitaine a rapporté un vase de chine »
    - the captain brought back a Chinese vase



- with ARIANE-G5 (GETALP)

A phrase structure grammar

$S \rightarrow NP\ \ VP\ .$

$VP \rightarrow V\ \ NP$

$VP \rightarrow V\ \ NP\ \ PP$

$NP \rightarrow NP\ \ PP$

$NP \rightarrow N$

$NP \rightarrow NNP$

$NP \rightarrow DET\ NC$

$NP \rightarrow DET\ NC\ PP$

$PP \rightarrow P\ \ NP$

Grammar rules

$NC \rightarrow$ cats

$NC \rightarrow$ captain

$NC \rightarrow$ vase

$NC \rightarrow$ scratch

$V \rightarrow$ see

$NNP \rightarrow$ John

$P \rightarrow$ with

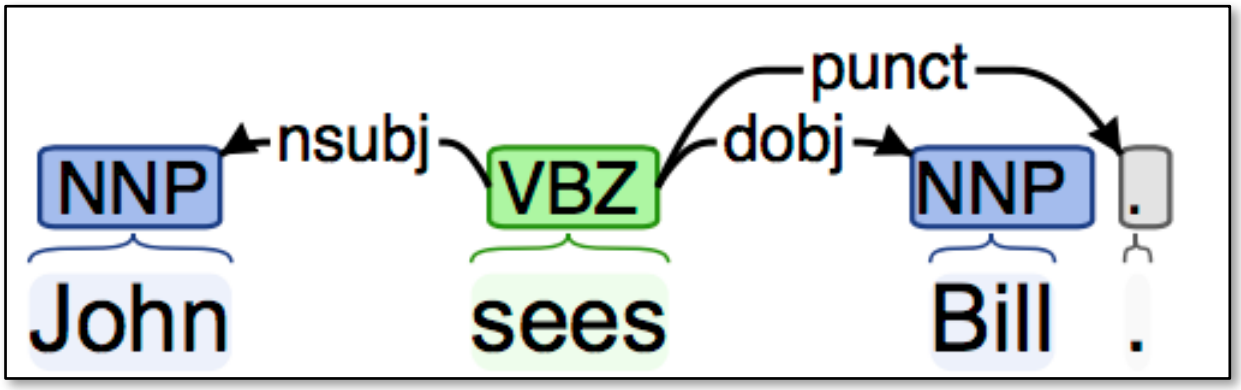
Dictionary

Dependency Parsing

- Key notions
  - governors with dependents
- Principles
  - Text is broken into words
  - establish relationship between “head” words and words which modify these head with a directed binary grammatical relation
- Structure
  - Terminals are the words, with some additional labels
  - Edges are labeled with a grammatical relation

Dependency Parsing

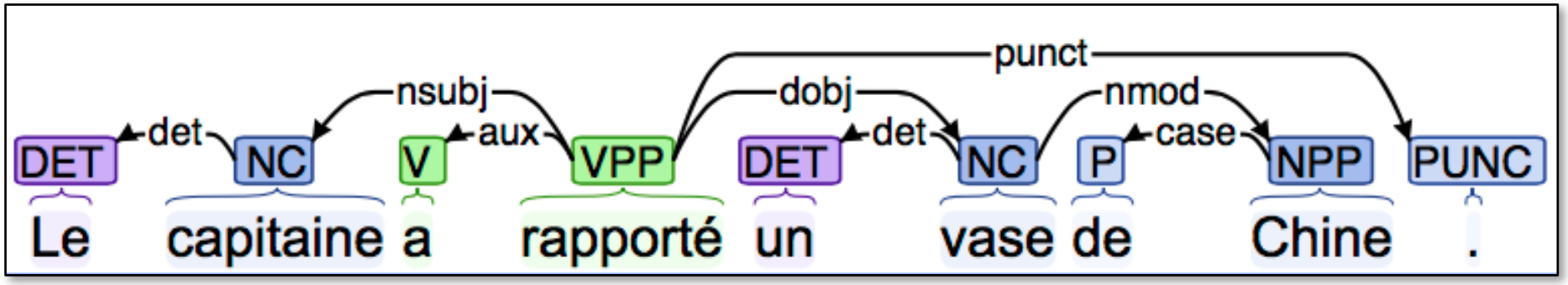
- Example
  - "John sees Bill."



  - from corenlp.run

Dependency Parsing

- Example
  - « Le capitaine a rapporté un vase de chine. »
    - the captain brought back a Chinese vase



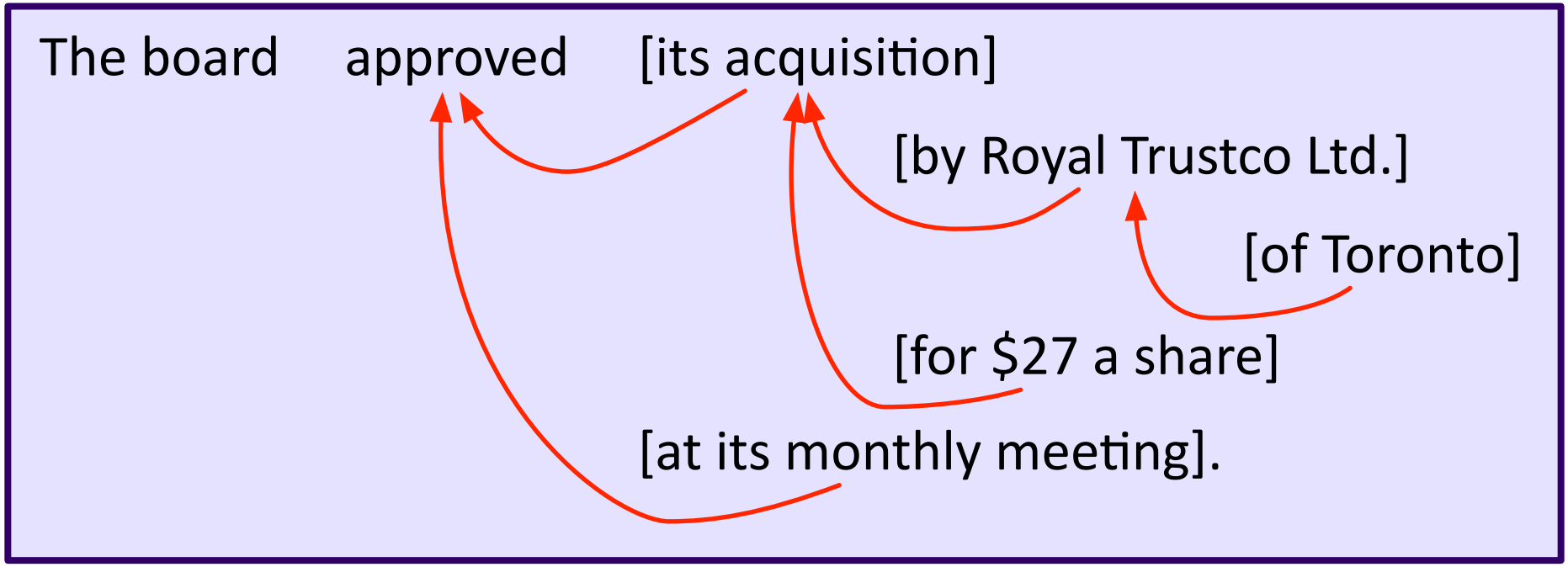
- from corenlp.run

Problem: Ambiguities

- I saw that gasoline can explode
- I saw the man in the park with a telescope.

Problem: Ambiguities

- In the ”real word”
  - The board approved its acquisition by Royal Trustco Ltd. of Toronto for $27 a share at its monthly meeting.



Approaches

- Expert
  - expert parser use handcrafted
    - rules
      &
    - dictionaries
- Empirical (statistical)
  - statistical parsers
    - are trained from a set of hand-parsed sentences (treebanks; eg: Penn TreeBank, French TreeBank, …)
    - know statistics about phrase structure and word relationships, and use them to assign the most likely structure to a new sentence

Tools available

- a lot…
  - Link Grammar
  - CoreNLP
  - SyntaxNet
  - TurboParser
  - spaCy
  - NLTK
  - Bllip-Parser
  - Berkeleyparser
  - MaltParser
  - MSTParser
  - …

9. COREFERENCE RESOLUTION

Coreference resolution

- Find out which (noun) phrases refer to the same entities in the world
  - Sarah asked her father to look at her. He appreciated that his eldest daughter wanted to speak frankly.
- ≈ anaphora resolution
- ≈ pronoun resolution
- ≈ entity resolution

A whole new problem

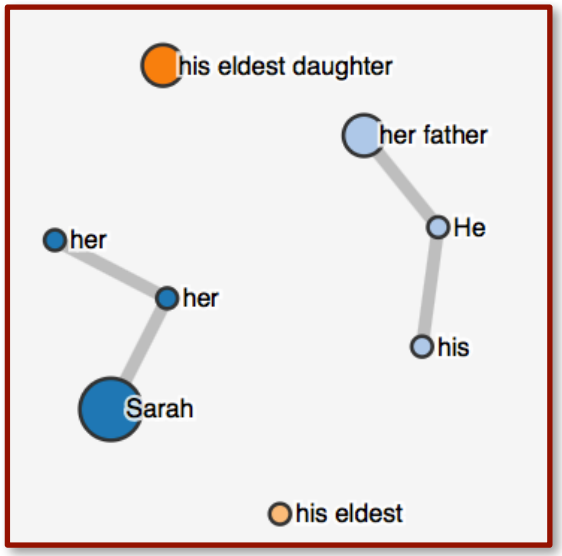
- Tools seen so far process one sentence at a time (or use the whole document but ignore all structure and just count)
- Coreference uses the whole document
- The resources used will grow with the document size – you might want to try a chapter not a novel
- Coreference systems normally require parsers, NER, etc. as preprocessors, and use of lexicons

Availability?

- English-only for the moment….
- While there are some papers on coreference resolution in other languages, I am aware of no downloadable coreference systems for any language other than English
- For English, there are a good number of downloadable systems, but their performance remains modest. It's just not like POS tagging, NER or parsing

OK example

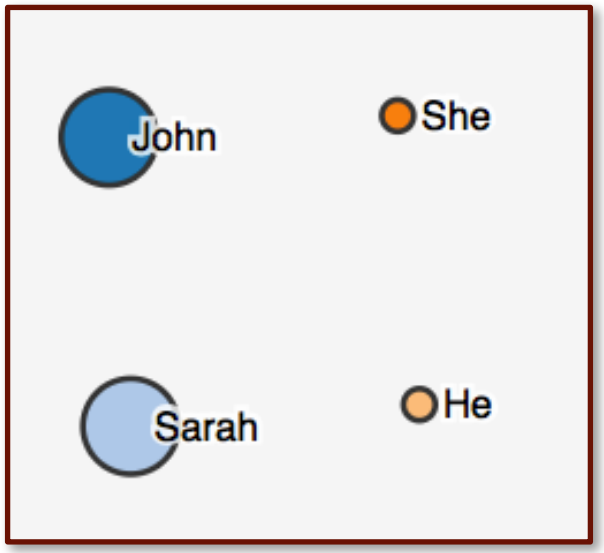
- Example human tagged
  - Sarah asked her father to look at her. He appreciated that his eldest daughter wanted to speak frankly.
- Tagged output*
  - [Sarah] asked [[her] father] to look at [her] . [He] appreciated that [[[his] eldest] daughter] wanted to speak frankly .



* the tool used is from
Cognitive Computation Group
Department of Computer Science
University of Illinois at Urbana-Champaign
link

KO example

- Example human tagged
  - John gave flowers to Sarah. She blushed. He was very happy.
- Tagged output*
  - [John] gave flowers to [Sarah] . [She] blushed . [He] was very happy .



* the tool used is from
Cognitive Computation Group
Department of Computer Science
University of Illinois at Urbana-Champaign
link

NLP FRAMEWORKS & PACKAGES

The Big 3 NLP Frameworks

- GATE – General Architecture for Text Engineering
  - U. Sheffield
  - http://gate.ac.uk/
  - Java, quite well maintained
  - Includes tons of components
- UIMA – Unstructured Information Management Architecture
  - Originally IBM; now Apache project
  - http://uima.apache.org/
  - Professional, scalable, etc.
  - Non-starter unless skills with Xml, Eclipse, Java or C++, etc.
- NLTK – Natural Language Toolkit
  - started by Steven Bird
  - http://www.nltk.org/
  - Big community; large Python package; corpora and *books* about it
  - But it's code modules and API, no GUI or command-line tools
  - Like R for NLP. But, R's becoming very successful….

The main NLP Packages

see also: other resources for NLP

- NLTK Python
  - http://www.nltk.org/
- OpenNLP
  - http://incubator.apache.org/opennlp/
- Stanford NLP
  - http://nlp.stanford.edu/software/
- LingPipe
  - http://alias-i.com/lingpipe/
- Statistical Natural Language Processing
  - http://nlp.stanford.edu/links/statnlp.html

The major resources

OTHER RESOURCES AND TOOLS FOR NLP

- A curated list of beginner resources in Natural Language Processing
  - Dibya Chakravorty
  - https://github.com/gutfeeling/beginner_nlp
- Awesome-NLP
  - Keon Kim & Martin Park
  - https://github.com/keon/awesome-nlp
- ACL Wiki (Association for Computational Linguistics)
  - https://www.aclweb.org/aclwiki/
- Corpus…the open parallel corpus
  - http://opus.lingfil.uu.se
- Voyant
  - https://voyant-tools.org
- Associations
  - ACL: https://www.aclweb.org/
  - EACL: http://www.eacl.org
  - AMTA: https://amtaweb.org
  - EAMT: http://www.eamt.org

CREDITS

Sources/Resources Used

- Bob Berwick (2003) *Artificial Intelligence*, Fall 2003: Slides on Language (link)
- Christopher Manning (2011) *Natural Language Processing Tools for the Digital Humanities* (link)
- Christopher Manning (2007) *Seven Lectures on Statistical Parsing* (link)
- Wikipedia for “simple” definitions of the terms