Logical and Computational Structures for Linguistic Modeling
Part 1 – Introduction

Éric de la Clergerie
<Eric.De_La_Clergerie@inria.fr>

16 Septembre 2014
Part I

Introduction
Very large diversity with at least 6000 languages over the world including sign languages
Natural Language Processing

NLP

Computer Science

Cognitive Sciences

Linguistics
NLP triggered by machine translation (cold war context)

- first demo in 1954 by IBM (Russian -> English)
- negative impact of Y. Bar-Hillel’s report (1960), advocating for computer-assisted translation
- several generations of methods and software

- direct approach (word to word + reordering)
- by transfer (syntactic source tree to target tree)
- by interlingua (language independent semantic representation)

Pb of lexical ambiguities

- the spirit is willing but the flesh is weak.
- (en → ru → en) The vodka is good but the meat is rotten.
Nowadays, statistical machine translation (SMT): *Google Translate*

Principles:
- given \( f \), search \( e \) such that

\[
\hat{e} = \arg \max_{e \in E(f)} p(e|f) = \arg \max_{e \in E(f)} \frac{p(f|e)p(e)}{p(f)} = \arg \max_{e \in E(f)} p(f|e)p(e)
\]

- translation model probabilities \( p(f|e) = \prod_i p(f_i|e_i) \)
  estimated from aligned segments \( (f_i/e_i) \) from parallel texts
- language model probabilities \( p(e) \) estimated from huge corpora
  smoothed \( n \)-gram models
An old dream: creating artificial intelligence (AI)

**Turing Test** (1950): based on conversation between a judge, an human and a program

⇒ language proficiency

**ELIZA** (Weizenbaum 1966, ancestor of chatbots)

I am the psychotherapist. Please, describe your problems.

I'm not feeling well

Why do you say "I'm not feeling well"?

Well, I've no energy left

Is it because of your plans that you say "well you have no energy left"?

All my plans are total failures

Maybe your life has something to do with this.

Each year since 1991, the Loebner price awards the best chatbots

In 2014, the chatbot Eugene Goostman passed the Turing test (but dubious)
Importance of semantics through knowledge and implicits

→ in the 70s, development of several systems associated to micro-worlds

**SHRUDLU** (block-world) Winograd 1970

Knowledge representation and inferences

- notion of frames (Minksy) and scripts
  - SHOPPING script to understand: *I am going shopping / did you bring enough money?*
- Conceptual dependency theory (Schank)
  - states, primitives & (conceptual) dependencies

but,

- many such scripts/frames/scenarii
- scaling problems

Nevertheless, manual efforts for developing large resources about language and knowledge

**FRAMENET** (Baker & Fillmore, 1998), **WORDNET** (Miller), ontologies, ...

Nowadays, **knowledge acquisition** from large textual corpora
Progressive development of grammatical formalisms for describing syntax, inspired by Noam Chomsky

- Regular grammars: too simple!
- Augmented Transition Networks (ATN) and CFGs: not adequate for linguistic description, not expressive enough
- Transformational Grammars: too powerful
- HPSG (Pollard & Sag, 1994), LFG (Bresnan & Kaplan, 70s), TAGs (Joshi, 1975), CCG (Steedman, 1987), . . . adequate for description, reflecting linguistic theories, more or less tractable

Development of relatively efficient parsing techniques chart parsing, lexicalization, . . .

But,
- difficulty to develop and maintain large coverage grammars
- difficulty to select the correct analysis for a sentence (ambiguity)
First successes of statistical models in Speech processing

Hidden Markov Models (HMM)

Very successful for more and more NLP tasks, due to the conjunction of

1. large amount of available electronic spoken and written data
2. powerful computers for handling data (time and memory)
3. more and more sophisticated machine learning techniques

More specifically, 2 main approaches:

- preparation & distribution of annotated data
  (BROWN CORPUS, PENN TREEBANK 1993, . . .)
  \(\leadsto\) supervised learning

- huge amount of data, with web, video, . . .
  \(\leadsto\) unsupervised learning (more difficult !)
Siri, dois-je prendre mon parapluie ?

http://www.youtube.com/watch?v=xIBezLFLjiI

Apple’s vocal assistant SIRI doing its best to help you !
(but see also http://www.youtube.com/watch?v=WGxDaX1__yI)
And the answer is ... Elementary, my dear Watson!

http://www.youtube.com/watch?v=WFR3lOm_xhE

**Watson**, a software (and a supercomputer) developed by **IBM**, winner of TV game *Jeopardy*
Query in category **literary character**

*Wanted for general evil-ness; last seen at the tower of Barad-dur; it’s a giant eye, folks. Kinda hard to miss*

And the answer is: Sauron

Relation extraction based on “deep” patterns:

```
authorOf :: [Author] [WriteVerb] [Work]
```

- In 1936, **he wrote** his last play, *The Boy David*
- **Robert Louis Stevenson** fell in love with Fanny Osbourne, a married woman, and later **wrote** this tale for her son
- **Somnium**, an early work of science fiction, was **written** by **this German**
- **This French Connection actor** coauthored the 1999 novel *Wake of the Perdido Star*

*Deep parsing in Watson* *(McCord, Murdock, & Boguraev)*
NLP: which applications?

Many potential or existing applications:
- spelling/grammatical/stylistic correction (CORDIAL, WORD, . . .)
- information retrieval (IR)
- text mining, knowledge acquisition
- opinion/sentiment mining (e-reputation)
- information extraction (IE) & Question-Answering (QA) systems (WATSON),
- machine translation (GOOGLE TRANSLATE, SYSTRAN, MOSES, . . .) and computer-assisted translation
- automatic summarization
- generation
- Human-Machine Communication (SIRI), chatbot (ELIZA, ALICE)
- speech recognition, dictation (NUANCE)
- speech synthesis
- . . .
Part II

A “poor” view of language
A few simple experiments

**Objective:** to explore some properties of language with simple but nevertheless powerful methods

**Methods:**
- characters, char sequences (n-grams), words
- frequencies
- probabilities
- language models

Using documents available on Gutenberg Project
http://www.gutenberg.org
- for French: Jules Vernes, Proust, Maurice Leblanc, Gaston Leroux, Stendhal (∼ 1Mots)
- for English: Shakespeare (∼ 1Mmots)

A few simple Perl scripts (available on demand)
alternative languages: Python (numpy), R, Octave, ...

quantitative linguistics, data-driven linguistics, corpus linguistics
1. Do we get a message?

2. Language identification

3. Authorship attribution

4. Sequence prediction

5. Capturing word meaning
The necklace tree is being buttonholed to play cellos and the burgundian premeditation in the Vinogradoff, or Wonalancet am being provincialised to connect. Were difference viagra levitra cialis then the batsman’s dampish ridiculousnesses without Matamoraras did hear to liken, or existing and tuneful difference viagra levitra cialis devotes them.

Detecting Fake Content with Relative Entropy Scoring (Yvon and al)
If we should identify or design an (efficient) language, which expected properties/constraints? (some from C. Hocket)

- signal over a noisy channel $\implies$ robustness, redundancy
- **Semanticity:** primary function of language is *communication* inform, query, order about things, events, sentiments, . . .
- linearity $\implies$ ordering (syntax?)
- **Discreteness:** combinable elementary parts (possibly at various levels) phonemes /ˈlæŋwɪdʒ/, letters *l.a.n.g.u.a.g.e*, words *language*, . . .
- **Productivity:** ability to describe complex and new situations word creation, longer and longer messages
- **Arbitrariness:** no direct relationship between a word and its meaning
  Ferdinand de Saussure: *signifiant* / *signifié*
- cultural artifact $\implies$ learnability contingency, evolution, diversity
- **Efficiency,** fast real time $\implies$ fast emitting (speaker), short messages, fast decoding (listener)
  frequent short words, information delta (shared knowledge), ambiguity (but context) E. Gibson
An Expedient was therefore offered, that since Words are only Names for Things, it would be more convenient for all Men to carry about them, such Things as were necessary to express the particular Business they are to discourse on.

Another great Advantage proposed by this Invention, was that it would serve as a Universal Language to be understood in all civilized Nations

*Gulliver’s Travels – J. Swift*

Close alternatives: iconic languages
Productivity

No bound on what can be produced
Noam Chomsky: embedding, recursion (e.g. relative clauses)
strong principle of an Universal Grammar

Maudit soit le père de l’épouse du forgeron qui forgea le fer de la cognée avec laquelle le bûcheron abattit le chêne dans lequel on sculpta le lit où fut engendré l’arrière-grand-père de l’homme qui conduisit la voiture dans laquelle ta mère rencontra ton père! (Desnos)

In most languages, many recursive constructions
relative clauses, subordinates, coordination, prepositional phrases (PPs), . . .

But recent controversy about recursion: Pirahã (D. Everett)
Message A
Les blaireaux viennent de gagner une bataille décisive au Royaume-Uni.

Message B
uyf pven-yexo anyccycb gy 3e3cy- xcy pebenvvy gs’nfnay ex UdlexqyiAcn.

Message C
éev -dfvonèné axeé3o’t -t èfjvmv ec3 galqjvfu bmlpspcb è3 UpcuèuAb3ix.

Message D
Aq’sRv AUxUplRv-URèlquyci q3dppgciyx-UxsIn AUmp lqplbbRv3fRv dlgUyx iAf-iqAqbbRvpl-U 3p3fApstjsstgU3p lqyx -lstgU’glq-Ufm3pyxx-dp.
Natural languages exhibit a typical mix of:

- redundancy
  function words (determiners, prepositions, conjunctions, \ldots) and other very frequent words
- diversity (richness of vocabulary and constructions)
- distribution over word length
  frequent words are generally short

$\Rightarrow$ impact on the *entropie* of messages

**Base:** *Prediction and Entropy of Printed English*

Shannon (1950)
**Starting point:** How well can we predict the next char $c_{n+1}$ extending a sequence $c_1 \cdots c_n$

- fully random *fdabRr pne-ba-RècU*
- fully predictable *abababababab*
- partly predictable *je me demande ce qu*

More formally, limit of conditional entropy (per-char entropy)

$$H = \lim_{n \to \infty} H_n$$

with

$$H_{n+1} = -\sum_{c_1 \cdots c_n c_{n+1}} p(c_1 \cdots c_n c_{n+1}) \log_2 p(c_{n+1} | c_1 \cdots c_n)$$

limit cases:

- $H_0 = \log_2 |\text{alphabet}|$ (equiprobable distribution)
- $H_1 = -\sum c p(c) \log_2 p(c)$
In practice

$H_n$ computed over large textual corpora, considering $n$-grams $c_1 \cdots c_n$, and

$$p(c_1 \cdots c_n) = \frac{\#(c_1 \cdots c_n)}{\#(\text{sequences of size } n)}$$

Problems:
- the number of n-grams grows exponentially with $n$ ($|V|^n$)
  $\implies$ cost in time for collecting and in place for storing
- never enough data (data sparseness) to observe enough occurrences of $c_1 \cdots c_n$ for $n$ large enough
  not observing $c_1 \cdots c_n$ in a corpus doesn’t mean the sequence is impossible! $\implies$ need for smoothing techniques

Google N-grams

Google distributes (word) n-grams ($n \leq 5$) computed over huge corpora (5M books) for several languages

https://books.google.com/ngrams
Some results

For English (27 chars), Shannon found $H_3 = 3.3$ and postulates $H$ between 1 and 2. Also based on the use of a deduction letter game.

For $H_0 \Longrightarrow$ coding of chars on 7 or 8 bits. Less bits for longer sequences $\Longrightarrow$ compression.
Going further

Entropy is only a first step for determining the status of a message

Other hints

- word diversity (if easy notion of “word”)
- rate of emergence of new words
- relationship between frequency and word length
- distribution of words in potential word space
- ...
Zipf law (1949)

Power law strongly present in linguistic data, denoting an exponential decrease of frequency $f$ w.r.t. rank $r$:

$$f_r \propto \frac{1}{r^\alpha} \text{ with } \alpha = 1 + \epsilon$$

or better, Mandelbrot (1982) $f_r \propto \frac{1}{(r+\rho)^\alpha}$ with $\rho \gg 1$

- a few words/structures are frequently used;
- many many words are very rarely used (long tail)

- possible interpretation: language rewards reuse but is open to creativity
- maybe related to cognitive and/or evolution constraints (least effort)

but see also Lukasz Debowski Zipf’s Law: What and Why?

**Note:** similar relation on word lengths

$$l \approx 1 + \frac{a}{fb}$$

frequent words tend to be short (faster coding/decoding)
Distribution of words (lemmas) in a corpus of 500 millions words, avec 3,234,274 distinct lemmas, including 71,348 not proper nouns:

Most frequent French words: le, de, “,”, “.”, à, un, et, cln, “:”, en, être/v, …

80% occurrences covers with \( \sim 1500 \) lemmas and 90% with 6000 lemmas.
Distribution of FRMG constructions (trees) over 10,096 sentences from FRENCH TREEBANK (journalistic texts, Le Monde).

- only 223 over 344 possible trees are used
- 90% of occurrences covered with 25 trees; 99% with 100 trees
- note: coverage: 94.3%, accuracy 86.6%
A kind of probabilistic distribution over distributions close to Zipf law, popularized with a variant, the Chinese Restaurant Process

\[ n + 1^{\text{th}} \text{ customer sits, with probability } p \text{ (and } \alpha > 0, 0 < \mu < 1), \]

- at table \( k \) with \( n_k \) customers (old word)
  \[ p(x_{n+1} = k|x_{1:n}) = \frac{n_k - \mu}{n + \alpha} \]

- at a new table \( K + 1 \) (new word) with \( n = \sum_{k=1}^{K} n_k \)
  \[ p(x_{n+1} = K + 1|x_{1:n}) = \frac{\alpha + \mu.K}{n + \alpha} \]

In other words,

*The rich get richer (but some hope remains!)*

Also related to: Pòlya’s Urn, stick-breaking construction, Pitman-Yor process,
Occurrences of new words

![Graph showing occurrences of new words with vocabulary size on the y-axis and corpus size on the x-axis. Three lines represent different corpus sizes: French corpus, English corpus, and CRP distributions with parameters α = 900, µ = 0.44, and α = 500, µ = 0.46.](image-url)
234 pages book written between 1450 and 1520, with illustrations, but unknown author and content. But satisfy most criteria for an human language
http://fr.wikipedia.org/wiki/Manuscrit_de_Voynich
1. Do we get a message?

2. Language identification

3. Authorship attribution

4. Sequence prediction

5. Capturing word meaning

INRIA
É. de la Clergerie
TAL
16/09/2014 35 / 86
An easy task

Software:

- online: http://whatlanguageisthis.com/
- free: MGUESSER http://www.mnogosearch.org/guesser/

```bash
> echo "Beware_the_Jubjub_bird_and_shun_The_frumious_Bandersnatch" | ./mguesser -d maps/ -n3
0.6202442646 en iso-8859-1
0.6046028733 de latin1
0.5912522078 fr utf8
```

```bash
> echo "Il_était_grilheure;_les_slichtueux_toves_Gyraient_sur_l’alloinde_et_vriblaient" | ./mguesser -d maps/ -n3 -l l1
0.6878187060 fr utf8
0.6851934791 fr latin1
0.6823609471 fr iso-8859-1
```

```bash
> echo "Nakita_kitá_sa_tindahan_kahapon" | ./mguesser -d maps -n3
0.5999047756 tl ascii
0.5547670126 tl ascii
0.5282356739 fi latin1
```
### Simple language models

**language model files for MGUESSER**

<table>
<thead>
<tr>
<th>seq</th>
<th>freq</th>
<th>mot</th>
<th>freq</th>
<th>mot</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>_</td>
<td>4,762,268</td>
<td>_</td>
<td>8,097,193</td>
<td>_</td>
<td>7,119,158</td>
</tr>
<tr>
<td>e</td>
<td>3,227,901</td>
<td>e</td>
<td>4,757,841</td>
<td>e</td>
<td>6,188,609</td>
</tr>
<tr>
<td>s</td>
<td>1,736,708</td>
<td>t</td>
<td>3,450,856</td>
<td>n</td>
<td>3,781,083</td>
</tr>
<tr>
<td>a</td>
<td>1,722,683</td>
<td>o</td>
<td>3,181,965</td>
<td>i</td>
<td>2,867,838</td>
</tr>
<tr>
<td>t</td>
<td>1,573,003</td>
<td>a</td>
<td>2,910,346</td>
<td>r</td>
<td>2,540,532</td>
</tr>
<tr>
<td>i</td>
<td>1,544,233</td>
<td>n</td>
<td>2,617,886</td>
<td>s</td>
<td>2,085,127</td>
</tr>
<tr>
<td>n</td>
<td>1,451,396</td>
<td>i</td>
<td>2,601,399</td>
<td>t</td>
<td>2,047,798</td>
</tr>
<tr>
<td>r</td>
<td>1,395,479</td>
<td>s</td>
<td>2,330,971</td>
<td>h</td>
<td>1,939,960</td>
</tr>
<tr>
<td>u</td>
<td>1,343,622</td>
<td>r</td>
<td>2,232,821</td>
<td>a</td>
<td>1,932,605</td>
</tr>
<tr>
<td>o</td>
<td>1,262,006</td>
<td>h</td>
<td>2,157,803</td>
<td>d</td>
<td>1,796,659</td>
</tr>
<tr>
<td>l</td>
<td>1,167,742</td>
<td>l</td>
<td>1,423,346</td>
<td>en</td>
<td>1,488,315</td>
</tr>
<tr>
<td>e_</td>
<td>1,105,484</td>
<td>d</td>
<td>1,405,996</td>
<td>u</td>
<td>1,388,799</td>
</tr>
<tr>
<td>d</td>
<td>732,432</td>
<td>e_</td>
<td>1,340,805</td>
<td>l</td>
<td>1,319,841</td>
</tr>
<tr>
<td>s_</td>
<td>709,985</td>
<td>_t</td>
<td>1,120,482</td>
<td>n_</td>
<td>1,299,079</td>
</tr>
<tr>
<td>t_</td>
<td>662,637</td>
<td>th</td>
<td>1,051,445</td>
<td>er</td>
<td>1,266,324</td>
</tr>
<tr>
<td>m</td>
<td>591,466</td>
<td>u</td>
<td>988,874</td>
<td>c</td>
<td>1,241,121</td>
</tr>
</tbody>
</table>

---

**INRIA**

É. de la Clergerie

TAL

16/09/2014 38 / 86
Comparing the distributions

\[ d(a, b) = \sum_s |r_a(s) - r_b(s)| \]
Il était grilheure; les slictueux toves Gyraient sur l’alloinde et vriblaient

<table>
<thead>
<tr>
<th>seq</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>_</td>
<td>10</td>
</tr>
<tr>
<td>e</td>
<td>9</td>
</tr>
<tr>
<td>i</td>
<td>8</td>
</tr>
<tr>
<td>l</td>
<td>8</td>
</tr>
<tr>
<td>t</td>
<td>7</td>
</tr>
<tr>
<td>r</td>
<td>5</td>
</tr>
<tr>
<td>a</td>
<td>4</td>
</tr>
<tr>
<td>u</td>
<td>4</td>
</tr>
<tr>
<td>s</td>
<td>4</td>
</tr>
<tr>
<td>ai</td>
<td>3</td>
</tr>
<tr>
<td>n</td>
<td>3</td>
</tr>
<tr>
<td>t_</td>
<td>3</td>
</tr>
<tr>
<td>ient</td>
<td>2</td>
</tr>
<tr>
<td>ent</td>
<td>2</td>
</tr>
<tr>
<td>ien</td>
<td>2</td>
</tr>
<tr>
<td>ri</td>
<td>2</td>
</tr>
</tbody>
</table>

paste fr . latin1 . mdl msg.mdl | perl ./ngram_diff.pl

<table>
<thead>
<tr>
<th>langue</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr</td>
<td>26,832</td>
</tr>
<tr>
<td>br</td>
<td>29,262</td>
</tr>
<tr>
<td>af</td>
<td>29,506</td>
</tr>
<tr>
<td>ca</td>
<td>29,576</td>
</tr>
<tr>
<td>es</td>
<td>29,624</td>
</tr>
<tr>
<td>no</td>
<td>29,656</td>
</tr>
<tr>
<td>ca</td>
<td>29,874</td>
</tr>
<tr>
<td>nl</td>
<td>30,030</td>
</tr>
<tr>
<td>la</td>
<td>30,036</td>
</tr>
<tr>
<td>da</td>
<td>30,152</td>
</tr>
<tr>
<td>ro</td>
<td>30,452</td>
</tr>
<tr>
<td>de</td>
<td>30,458</td>
</tr>
<tr>
<td>is</td>
<td>30,530</td>
</tr>
<tr>
<td>af</td>
<td>30,560</td>
</tr>
<tr>
<td>it</td>
<td>30,648</td>
</tr>
<tr>
<td>en</td>
<td>30,694</td>
</tr>
</tbody>
</table>
In 2011, Kevin Knight and colleagues break the Copiale cypher, used in 105 page manuscript (~ 75Kchar), dated between 1760-1780.

http://stp.lingfil.uu.se/~bea/copiale/
Comparison with the distribution of various languages:

- not a substitution cypher
- slight proximity with German (coherent with other hints)

Hypothesis of an homophonic cypher

- a char $c$ with strong frequency $f$ may be substituted by any char $x$ selected in set $\{x_1, \ldots, x_n\}$, with $n$ proportional to $f$
- used for D messages (entropy computation)

This kind of cyphers:

- hides the distribution over chars (unigram distribution)
- but is imperfect over char sequences, in particular for sequences involving rare chars
  example: qu in French
Copiale cypher = homophonic code for German
Initiation manuscript for a secret society

<table>
<thead>
<tr>
<th>Plain</th>
<th>Cipher</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ã</td>
</tr>
<tr>
<td>Ä</td>
<td>ø</td>
</tr>
<tr>
<td>B</td>
<td>ß</td>
</tr>
<tr>
<td>C</td>
<td>ç</td>
</tr>
<tr>
<td>D</td>
<td>ð</td>
</tr>
<tr>
<td>E</td>
<td>ë</td>
</tr>
<tr>
<td>F</td>
<td>ñ</td>
</tr>
<tr>
<td>G</td>
<td>ß</td>
</tr>
<tr>
<td>H</td>
<td>û</td>
</tr>
<tr>
<td>I</td>
<td>ï</td>
</tr>
<tr>
<td>J</td>
<td>ÿ</td>
</tr>
<tr>
<td>K</td>
<td>ì</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Plain</th>
<th>Cipher</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ã</td>
</tr>
<tr>
<td>Ä</td>
<td>ø</td>
</tr>
<tr>
<td>B</td>
<td>ß</td>
</tr>
<tr>
<td>C</td>
<td>ç</td>
</tr>
<tr>
<td>D</td>
<td>ð</td>
</tr>
<tr>
<td>E</td>
<td>ë</td>
</tr>
<tr>
<td>F</td>
<td>ñ</td>
</tr>
<tr>
<td>G</td>
<td>ß</td>
</tr>
<tr>
<td>H</td>
<td>û</td>
</tr>
<tr>
<td>I</td>
<td>ï</td>
</tr>
<tr>
<td>J</td>
<td>ÿ</td>
</tr>
<tr>
<td>K</td>
<td>ì</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Plain</th>
<th>Cipher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logograms</td>
<td>Æ Ø Δ X</td>
</tr>
<tr>
<td></td>
<td>ß</td>
</tr>
<tr>
<td></td>
<td>ë</td>
</tr>
<tr>
<td></td>
<td>ñ</td>
</tr>
<tr>
<td></td>
<td>ß</td>
</tr>
<tr>
<td></td>
<td>û</td>
</tr>
<tr>
<td></td>
<td>ï</td>
</tr>
<tr>
<td></td>
<td>ÿ</td>
</tr>
<tr>
<td></td>
<td>ì</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Plain</th>
<th>Cipher</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>ì</td>
</tr>
<tr>
<td>X</td>
<td>ß</td>
</tr>
<tr>
<td>Y</td>
<td>û</td>
</tr>
<tr>
<td>Z</td>
<td>ë</td>
</tr>
<tr>
<td>SCH</td>
<td>ñ</td>
</tr>
<tr>
<td>SS</td>
<td>ò</td>
</tr>
<tr>
<td>ST</td>
<td>ñ</td>
</tr>
<tr>
<td>CH</td>
<td>õ</td>
</tr>
<tr>
<td>repeat</td>
<td>:</td>
</tr>
<tr>
<td>EN/EM</td>
<td>ã ë ã</td>
</tr>
<tr>
<td>space</td>
<td>abcde fghijklmnopqrstuvwxyz</td>
</tr>
</tbody>
</table>
1. Do we get a message?

2. Language identification

3. Authorship attribution

4. Sequence prediction

5. Capturing word meaning
The corpus

A few books from Gutemberg
http://www.gutenberg.org

- **Stendhal**
  - Le rouge et le noir (1830, 212Kmots)
  - La chartreuse de Parme (1839, 219Kmots)

- **Jules Vernes**
  - Voyage au centre de la terre (1864, 87Kmots)
  - 20000 lieues sous les mers (1870, 175Kmots)
  - Le tour du monde en 80 jours (1873, 100Kmots)

- **Gaston Leroux**
  - Le mystère de la chambre jaune (1907, 109Kmots)
  - Le fauteuil hanté (1909, 66Kmots)

- **Maurice Leblanc**
  - Arsène Lupin gentleman-cambrioleur (1907, 73Kmots)

- **Marcel Proust**
  - Du côté de chez Swann (1913, 201Kmots)
  - Le côté de Guermantes (1921-22, 85Kmots)
Naive segmentation into **token**: whitespace, punctuations, apostrophes (in front of vowels)

```bash
> perl ./analyze.pl pg13765.l1.txt
```

<table>
<thead>
<tr>
<th>mot</th>
<th>#occ</th>
<th>freq (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>13,693</td>
<td>6.80</td>
</tr>
<tr>
<td>de</td>
<td>7,734</td>
<td>3.84</td>
</tr>
<tr>
<td>.</td>
<td>4,485</td>
<td>2.23</td>
</tr>
<tr>
<td>la</td>
<td>3,846</td>
<td>1.91</td>
</tr>
<tr>
<td>à</td>
<td>3,603</td>
<td>1.79</td>
</tr>
<tr>
<td>et</td>
<td>3,491</td>
<td>1.73</td>
</tr>
<tr>
<td>que</td>
<td>3,107</td>
<td>1.54</td>
</tr>
<tr>
<td>le</td>
<td>2,945</td>
<td>1.46</td>
</tr>
<tr>
<td>il</td>
<td>2,803</td>
<td>1.39</td>
</tr>
<tr>
<td>qu’</td>
<td>2,747</td>
<td>1.36</td>
</tr>
<tr>
<td>l’</td>
<td>2,476</td>
<td>1.23</td>
</tr>
<tr>
<td>un</td>
<td>2,462</td>
<td>1.22</td>
</tr>
<tr>
<td>d’</td>
<td>2,455</td>
<td>1.22</td>
</tr>
<tr>
<td>les</td>
<td>2,276</td>
<td>1.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>mot</th>
<th>#occ</th>
<th>freq (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>13,912</td>
<td>7.92</td>
</tr>
<tr>
<td>de</td>
<td>7,860</td>
<td>4.48</td>
</tr>
<tr>
<td>.</td>
<td>6,238</td>
<td>3.55</td>
</tr>
<tr>
<td>le</td>
<td>3,243</td>
<td>1.85</td>
</tr>
<tr>
<td>et</td>
<td>3,066</td>
<td>1.75</td>
</tr>
<tr>
<td>la</td>
<td>2,958</td>
<td>1.68</td>
</tr>
<tr>
<td>à</td>
<td>2,762</td>
<td>1.57</td>
</tr>
<tr>
<td>les</td>
<td>2,336</td>
<td>1.33</td>
</tr>
<tr>
<td>l’</td>
<td>2,011</td>
<td>1.14</td>
</tr>
<tr>
<td>des</td>
<td>1,968</td>
<td>1.12</td>
</tr>
<tr>
<td>un</td>
<td>1,708</td>
<td>0.97</td>
</tr>
<tr>
<td>que</td>
<td>1,556</td>
<td>0.89</td>
</tr>
<tr>
<td>d’</td>
<td>1,493</td>
<td>0.85</td>
</tr>
<tr>
<td>–</td>
<td>1,432</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Comparing the distributions

We compare the variations of distributions for the $n$ most frequent words.

Need a distance or a similarity measure between the word rankings

$$\text{rank-distance}(d_a, d_b) = \sum_w |r_a(w) - r_b(w)|$$

Other (normalized) measures are available:
Spearman correlation measure $\rho \in [-1, 1]$, Kendall coefficient $\tau$

$$\rho = 1 - \frac{6\sum_w(r_a(w) - r_b(w))^2}{n(n^2 - 1)}$$
### Rank-distance matrix for $n = 50$

- perl ./rankdis.pl *.voc

<table>
<thead>
<tr>
<th></th>
<th>Du Côté de Chez ...</th>
<th>La Chartreuse ...</th>
<th>Le mystère de ...</th>
<th>Le fauteuil hanté</th>
<th>Tour Du Mond 80 ...</th>
<th>Voyage au Centre ...</th>
<th>20000 Lieues ...</th>
<th>Le Rouge et le ...</th>
<th>Le Côté de Guermantes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Du Côté de Chez ...</td>
<td>0</td>
<td>62</td>
<td>106</td>
<td>92</td>
<td>84</td>
<td>108</td>
<td>120</td>
<td>118</td>
<td>68</td>
</tr>
<tr>
<td>La Chartreuse</td>
<td>0</td>
<td>100</td>
<td>92</td>
<td>84</td>
<td>78</td>
<td>100</td>
<td>90</td>
<td>36</td>
<td>66</td>
</tr>
<tr>
<td>Le mystère de</td>
<td>0</td>
<td>68</td>
<td>100</td>
<td>122</td>
<td>136</td>
<td>122</td>
<td>100</td>
<td>112</td>
<td>100</td>
</tr>
<tr>
<td>Le fauteuil hanté</td>
<td>0</td>
<td>76</td>
<td>108</td>
<td>134</td>
<td>122</td>
<td>88</td>
<td>100</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>Arsène Lupin</td>
<td>0</td>
<td>84</td>
<td>88</td>
<td>88</td>
<td>84</td>
<td>82</td>
<td>100</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>Tour Du Mond 80</td>
<td>0</td>
<td>72</td>
<td>62</td>
<td>86</td>
<td>112</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voyage au Centre</td>
<td>0</td>
<td>46</td>
<td>104</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20000 Lieues</td>
<td>0</td>
<td>98</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Le Rouge et le</td>
<td>0</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Le Côté de Guermantes</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Regroup close books into **clusters**

Use an **Agglomerative Hierarchical Clustering**

1. [init] each book forms a cluster
2. [iterate] at each step, group the two **closest** clusters

\[
(c_1^*, c_2^*) = \arg\min_{c_1, c_2} \frac{\sum_{a \in c_1} \sum_{b \in c_2} d(a, b)}{|c_1| \cdot |c_2|}
\]

3. [end] stop when only one remaining cluster

**Note:** Many other clustering algorithms

Hierarchical Clustering \(\Rightarrow\) tree visualization as a **dendogram**
20000 Lieues sous les mers
Voyage au Centre de la Terre
Tour Du Mond 80 Jours
Le fauteuil hanté
Le mystère de la chambre jaune
Arsène Lupin gentleman-cambrioleur
Le Rouge et le noir
La Chartreuse de Parme
Le Côté de Guermantes
Du Côté de Chez Swann
- Rank Distance as a Stylistic Similarity
  Marius Popescu & Liviu P. Dinu
  starting point for this experiment

- Inter-textual distance and authorship attribution Corneille and Moliere
  Labbé, Cyril and Dominique Labbé. 2001.
Outline

1. Do we get a message?
2. Language identification
3. Authorship attribution
4. Sequence prediction
5. Capturing word meaning
Language models

Already explored for entropy computation over (char or) word sequences:

- word n-grams \( p(w_n|w_{1:n-1}) = p(w_n|w_1 \cdots w_{n-1}) \)

Use of chain rule and Markov assumption (with implicit \( w_i = \langle S \rangle \), for \( i \leq 0 \))

\[
p(w_1 \ldots w_N) = p(w_1) \prod_{i=2}^{N} p(w_i|w_{1:i-1}) \approx \prod_{i=1}^{N} p(w_i|w_{i-n+1:i-1})
\]

Maximum Likelihood Estimate \( p_{\text{MLE}} \) of \( p(w_n|w_{1:n-1}) \) computed over large corpora,

\[
p(w_n|w_{1:n-1}) \approx p_{\text{MLE}}(w_n|w_{1:n-1}) = \frac{c(w_{1:n})}{c(w_{1:n-1})}
\]

e.g., with bigrams,

\[
p(w_1 \ldots w_N) \approx \prod_{i=1}^{N} p_{\text{MLE}}(w_i|w_{i-1})
\]

**Note:** better approximation of \( p \) with some smoothing over \( p_{\text{MLE}} \)
**Task:** Given a model and a sequence, propose the most probable computations auto-adaptation of the model to an author (SwiftKey on smartphones)

Extending a sequence, by sampling accordingly to $p(w_N | w_{N-n+1:N-1})$

```
shell > cat pg13765.l1.txt | perl ./entropy.pl 8 4
...
> 100 il se précipite vers
il se précipite vers le pavillon m’empêcher son poste
d’observation de la hauteur. Qui dit: «Joseph Rouletabille qui con

> word 20 il pense que
il pense que c’est le «diable» ou la «Bête du Bon Dieu», la mère Agenoux, une vieille sorcière de Sainte–Geneviève– des–Bois, son miaulement
```

See also online https://www.cs.toronto.edu/~ilya/fourth.cgi
Smoothing

Principle:
- remove some probability mass from observed events (discounting)
- distribute this mass among unseen events

Questions:
- how much to remove?
- how to distribute?

Laplace smoothing (on unigrams): assume at least one occurrence

\[ p_L(w_i) = \frac{c(w_i) + 1}{N + V} = \frac{c^*(w_i)}{N} \text{ with } c^*(w_i) = \left(\frac{c(w_i) + 1}{N + V}\right) \]

On bigrams,

\[ p_L(b|a) = \frac{c(a, b) + 1}{c(a) + V} \]
**Intuition:** Smooth the count $c$ of n-gram $x$ through the number of n-grams with count $c + 1$.

In particular for unseen one ($c = 0$)

$$N_c = \sum_{x: c(x) = c} 1 \implies N = \sum_c cN_c$$

For $x$ seen, with $c(x) = c$, new estimator $c^*$

$$c^*(x) = (c + 1) \frac{E(N_{c+1})}{E(N_c)} \approx (c + 1) \frac{N_{c+1}}{N_c} \land p_{GT}(x) = \frac{c^*(x)}{N}$$

For $x$ unseen in training data ($c = c(x) = 0$)

$$p_{GT}(x) = \frac{E(N_1)}{N} \approx \frac{N_1}{N}$$

For some (large) values of $c$, $E(N_c)$ has to be estimated (by interpolation)
**Interpolation** and **backoff**

**Interpolation**: linear combining of several models, including simpler (denser) ones

\[
\hat{p}(c|ab) = \lambda_1 p(c|ab) + \lambda_2 p(c|b) + \lambda_3 p(c) \text{ with } \Sigma_{i=1}^{3} \lambda_i = 1
\]

\(\lambda_i\) learned on some development data set (while \(p\) learned on a training set)

**backoff**: when 0-counts at \(n\), back off to shorter n-gram models \((n-1)\), and so forth

\[
p_{\text{katz}}(c|ab) = \begin{cases} 
p_{\text{GT}}(c|ab) & \text{if } c(abc) > 0 \\
\alpha(ab)p_{\text{katz}}(c|b) & \text{if } c(ab) > 0 \\
p_{\text{GT}}(c) & \text{otherwise}
\end{cases}
\]

\[
p_{\text{katz}}(c|b) = \begin{cases} 
p_{\text{GT}}(c|b) & \text{if } c(bc) > 0 \\
\alpha(b)p_{\text{GT}}(c) & \text{otherwise}
\end{cases}
\]

\(\alpha\) parameters learned over development data set
1. Do we get a message?
2. Language identification
3. Authorship attribution
4. Sequence prediction
5. Capturing word meaning
The relation between a word and its meaning is arbitrary, but . . .

*Meanings of words are (largely) determined by their distributional patterns* (Harris 1968)

*You shall know a word by the company it keeps* (Firth 1957)

Practically, each word $w$ has an associated vector of weighted contexts $v_w$

**principle:** words semantically close have close vectors (e.g. $\cos(v_a, v_b)$)

Very large sparse vectors may be replaced by smaller dense vectors
Part III

A more traditional view of Linguistics
Paul, je t’ai dit que François Flore est sorti faché de chez son banquier car celui-ci lui avait ex abrupto refusé son prêt pour sa future maison ?

**Pragmatic**: context & knowledge
references: celui-ci=banquier, lui=son=sa=François, t’=Paul
discourse: refusal explains anger
scenarii, implicits

**Semantic**: meaning of sentences and words
predicative structures, roles (agent, patient, ...), scope
refuser(agent=celui-ci,patient=lui,theme=prêt)

**Syntax**: sentence structure and relations between words
syntactic functions (subject, object, ...): **celui-ci**=subject,
**prêt**=object, **lui**=indirect obj of **refusé**

**Morphology**: the words and their structure (**lubéronisation**)
segmentation into words, syntactic categories:
celui/pro -ci/adj lui/ld avait/aux ex_abrupto/adv ...
flexion (conjugaison): **avait**=avoir+3s+Ind+Imparfait
named entities (persons, locations, ...) : (François Flore) **PERSON_m**
Constituency vs dependencies

Paul mange un délicieux gâteau

S

NP

vp

v

un

det

NP

N

adj

nc

gâteau
From constituents to dependencies: using constituent heads

\[ h(S) = h(\text{VP}) = v \]

\[ h(\text{NP}) = h(\text{N}) \in \{ \text{nc}, \text{pn} \} \]

however, no perfect consensus over constituent and dependency schemes!
Main difficulties for NLP

- diversity and creativity $\implies$ NLP robustness
- implicit knowledge
- $\sim$ ambiguities: everywhere!
A never ending flow of new words!

- by borrowing and appropriation of foreign (and technical) words: googliser, tweeter, selfie

- by creation of neologisms, often using derivational morphology: lubéronisation, hippocotomomonstrosesquipédaliophobia, ou peur des mots trop longs

- by shortening/abbreviating existing words
Real-life documents have many occurrences of:

- **named entities** such as Persons, Organizations, Locations, Dates, Products, ...  
  some follow easy patterns (dates) but many don’t!

C’est la principale innovation d’Assassin’s creed : unity, le dernier-né de la franchise du géant français

- **terms**, often as multi-word expression (MWE)  
  Usually syntax-compliant, but not always

l’effarante invasion des “fils et filles de”

- (semi) frozen multi-word expressions  
  Usually syntax compliant, but not semantically compositional

il a pris le taureau par les cornes
Language evolves and specializes, and also one may play with language:


Carthographie des Nuages – D. Mitchell

@IziiBabe C mm pa élégant wsh tpx mm pa marshé a coté dsa d meufs ki fnt les thugs c mm pa leur rôle wsh

Ce n’est même pas élégant voyons, tu ne peux même pas marcher à coté de sa petite amie qu’ils font les voyous, ce n’est même pas leur rôle voyons.

It is not even elegant. One cannot even walk besides his girl friend, they already start bullying people. It is not even their role

Tweet / French Social Media Bank
More than a way to express a same idea, often through transformations at syntactic level (+ morphological adjustments).

*Les enfants allument la télé. La télé est allumée par les enfants.*

*Il donne un livre à Paul. Il donne à Paul un livre.*

*Il le lui donne. donne-le-lui ! ne le lui donne pas !*

*Tu dois parler à ton père. C’est à ton père que tu dois parler.*

(*) À ton père parler tu dois

*La critique est aisée. Critiquer est aisé. Il est aisé de critiquer!*

*Se connaître soi-même nécessite une bonne connaissance de soi.*
Part of syntactic diversity may be seen as transformations over a canonical representation.

e.g. active voice (canonical) $\rightarrow$ passive voice $\rightarrow$ wh-sentence $\rightarrow^* \ldots$

$\sim$ transformational grammars:

- a base grammar (say CFG) for building canonical constructions
- a finite set of transformations over syntactic trees

Peters & Ritchie (1973) Transformation grammars are too complex (power of Turing-machine)

reason: unbounded sequences of erasing/increasing transformations

No longer considered but influential for other formalisms such as TAGs, metagrammars,\ldots

idea: pre-computation at grammar level a finite set of transformation sequences
Ambiguity is present everywhere in language, but mostly invisible to humans

*il observe une maman avec ses jumelles*

- lexical ambiguity on *jumelles*
- syntactic ambiguity on PP-attachment of *avec ses jumelles*
- anaphora ambiguity on *ses*

At least 8 interpretations (2 at syntactic level)
for a chain of $k$ PPs, exponential number of syntactic trees wrt $k$

la Chambre des communes reprendra l’examen du projet de loi de ratification du traité de Maastricht dès la reprise de la session du soir dans la salle principale du batiment.
Implicit and Ambiguities

- Paul mange la pomme

- Paul mange le soir

Note: Prosody may help in this specific case (argument vs modifier)
Implicit and PP-attachments

- Il mange une tarte avec ses amis
- Il mange une tarte avec de la chantilly
- Il mange une tarte avec sa bière
- Paul mange une [ pomme de terre ] cuite

Conclusion we need some knowledge about words and world
By using distributional techniques to capture meanings and contexts

\[
\begin{align*}
\text{tartelette} & \quad \text{tarte} & \text{semantically close} \\
\text{quetsche} & \quad \text{kind of fruit} \\
\text{aux_fruits} & \quad \text{frequent context for tarte}
\end{align*}
\] ⇒ \text{tartelette à la quetsche}

il mange une tartelette maison à la quetsche.
Using very local knowledge

One may have ellipsis in a sentence to be filled by local information for instance, coordination with ellipse

 İl boit un café et elle ε un thé.
Chomsky hierarchy (1959): Classify grammars \((\mathcal{N}, \Sigma, S, \mathcal{P})\) with \(\mathcal{P}\) finite set of productions over terminal set \(\Sigma\) and non-terminal set \(\mathcal{N}\), notations: \(a \in \Sigma, \, A, B \in \mathcal{N}, \, \alpha, \beta, \gamma \in (\Sigma \cup \mathcal{N})^*\)

Type 3: Regular languages
\[
A \rightarrow a, \, A \rightarrow aB
\]

Type 2: Context-free languages
\[
A \rightarrow \gamma
\]

Type 1: Context-sensitive languages
\[
\alpha A \beta \rightarrow \alpha \gamma \beta, \, |\gamma| > 0
\]

Type 0: recursively enumerable languages
\[
\alpha \rightarrow \beta
\]
Chomsky (1957): “English is not a regular language”

The cat likes tuna fish
The cat [the dog chased] likes tuna fish
The cat [the dog [the rat bit] chased] likes tuna fish
The cat [the dog [the rat [the elephant admired] bit] chased] likes] tuna fish

⇒ analogous to $n^n v^n$ language (not a regular one)
A Context-Free Grammar $G = (\mathcal{N}, \Sigma, S, \mathcal{P})$ with

- $\mathcal{N}$ a finite set of non-terminals such as $S$, NP, VP
- $\Sigma$ a finite set of terminals such as nc, pn, v
- $S$ a distinguished non-terminal
- $\mathcal{P}$ a finite set of productions $A \rightarrow \gamma$ with $\gamma \in (\mathcal{N} \cup \Sigma)^*$

The context-free language $L(G)$ generated by $G$ defined as

$$L(G) = \{ w \in \Sigma^* | S \Rightarrow^* w \}$$

with $\Rightarrow^*$ transitive closure of

$$\alpha A \beta \Rightarrow \alpha \gamma \beta \text{ iff } A \rightarrow \gamma \in \mathcal{P}$$

Membership of $w \in L(G)$ may be checked in $O(|w|^3)$
CFGs seems sufficient for many syntactic phenomena, including embedding. In particular, $a^n b^n$ is a CFL.

The derivations may be represented by parse trees (or proof trees) similar to linguist’s syntactic trees:

$$S \Rightarrow NP \ VP \Rightarrow pn \ VP \Rightarrow pn \ VP \ PP \Rightarrow pn \ v \ NP \ PP \Rightarrow^* pn \ v \ det \ nc \ prep \ det \ nc$$

Diagram:

```
S  =>  S   =>  S   =>  *
    |     |     |     |
NP  VP  NP  VP  NP  VP  VP  PP
    |     |     |     |     |     |
pn  NP  PN  NP  VP  NP  np  det  nc  prep  det  nc
    |     |     |     |     |     |     |     |     |     |     |
v  NP  prep  NP  det  nc  det  nc
```

Production rules:
- $S \rightarrow NP \ VP$
- $NP \rightarrow pn$
- $NP \rightarrow det \ n$
- $NP \rightarrow NP \ PP$
- $VP \rightarrow v \ NP$
- $VP \rightarrow VP \ PP$
- $PP \rightarrow prep \ NP$
Are CFLs enough?

2 aspects:

- How do we check that a language is not context-free?
  - Use of pumping lemma

Theorem (Bar Hillel’s pumping lemma)

$L$ is a CFL iff

$$\exists N > 0, \forall w \in L, |w| > N \implies \exists u, v, w, x, y, \wedge$$

$$w = uvwxy \quad |vwx| \leq N \wedge |vx| > 0$$

$$\forall n \geq 0, uv^nwx^n y \in L$$

In particular, language $a^n b^m c^n d^m, \quad n, m \geq 0$ is not context-free

(cross-serial dependencies)

Can we find a linguistic counter-example? Not so easy!
Swiss-German example (Shieber 1985)

We can iterate, embedding more verbs (at the end) requiring case-marked arguments (accusative & dative).

Verbs should follow nouns, but dative nouns may be stacked before acc. nouns, and idem for verbs
Swiss German is not context-free

...das mer (d’chind)$^n$ (em Hans)$^m$ es huus (lönd)$^n$ (hälfd)$^m$ asstriiiche.

...that we (the children-ACC)$^n$ (Hans-DAT)$^m$ the house-ACC (let)$^n$ (helped)$^m$ pair.

We take homomorphism $h$ such that:

\[
\begin{align*}
    h(d’chind) &= a & h(säit das mer) &= \epsilon \\
    h(em Hans) &= h(noun-DAT) = b & h(es huus) &= \epsilon \\
    h(lönd) &= c & h(asstriiiche) &= \epsilon \\
    h(hälfd) &= h(v-DAT) = d & h(w) &= \epsilon \text{ otherwise}
\end{align*}
\]

and intersect $h(L_{SW})$ with regular language $L_R = a^*b^*c^*d^*$

\[
l = h(L_{SW}) \cap L_R = a^nb^mc^nb^m
\]

if $L_{SW}$ is a CFL, then $l$ is a CFL
(closures by homomorphism and intersection with regular language)

but $l$ is not CFLs, and therefore $L_{SW}$ is not CFL
Theorem

Swiss German is not a context-free language

No context-free grammar can generate the strings of Swiss-German language
\[ \Rightarrow \text{SG} \Rightarrow \text{notion of weak generative capacity} \]

\[ G_1 \equiv_{\text{weak}} G_2 \iff L(G_1) = L(G_2) \]

Actually, linguists are mostly interested by the parse trees
\[ \Rightarrow \text{notion of strong generative capacity} \]

\[ G_1 \equiv_{\text{strong}} G_2 \iff \text{trees}(G_1) = \text{trees}(G_2) \]

Easier to be persuaded than CFGs lack strong generative capacity to model some expected syntactic trees
Dutch exhibits similar phenomena than for Swiss-German, but without visible case-marking.

If we require parse trees reflecting these crossing dependencies, then the resulting set of parse trees can’t be generated by a CFG.

Dutch is not strongly CFG (but seems to be weakly CFG)
What about French?

There are several syntactic phenomena for French for whose “natural” syntactic trees do not correspond to CFG parse trees.

For instance, the comparative construction:

Paul est un plus grand joueur que toi!
We will need to explore new classes of languages (slightly) beyond CFLs.

Each class of language have an associated class of automata, that may be used for parsing.

<table>
<thead>
<tr>
<th>grammars</th>
<th>automata</th>
</tr>
</thead>
<tbody>
<tr>
<td>regular grammars</td>
<td>finite-state automata</td>
</tr>
<tr>
<td>context-free grammars</td>
<td>push-down automata</td>
</tr>
<tr>
<td>context-sensitive grammars</td>
<td>linear-bounded automata</td>
</tr>
<tr>
<td>unrestricted grammars</td>
<td>Turing machine</td>
</tr>
</tbody>
</table>

Efficient parsing is often related to modeling computations with an adapted class of automata.
Chomsky opposes a syntax-based view of language with a probabilistic one:

\[
\text{Colorless green ideas sleep furiously} \\
\text{Furiously sleep ideas green colorless}
\]

The two sentences should not occur \( \Rightarrow p(s_1) = p(s_2) = 0 \)
But \( s_1 \) is grammatical while \( s_2 \) is not

However, F. Pereira (2000) using (smoothed) language models

\[
\frac{p(\text{Colorless green ideas sleep furiously})}{p(\text{Furiously sleep ideas green colorless})} \approx 2 \cdot 10^5
\]

where \( p(w_{1:n}) = p(w_1) \prod_{i=2}^{n} p(w_i|w_{i-1}) \) with \( p(w_i|w_{i-1}) = \sum_{c=1}^{C} p(w_i|c)p(c|w_{i-1}) \)
aggregated Markov model \( (C = 16) \)