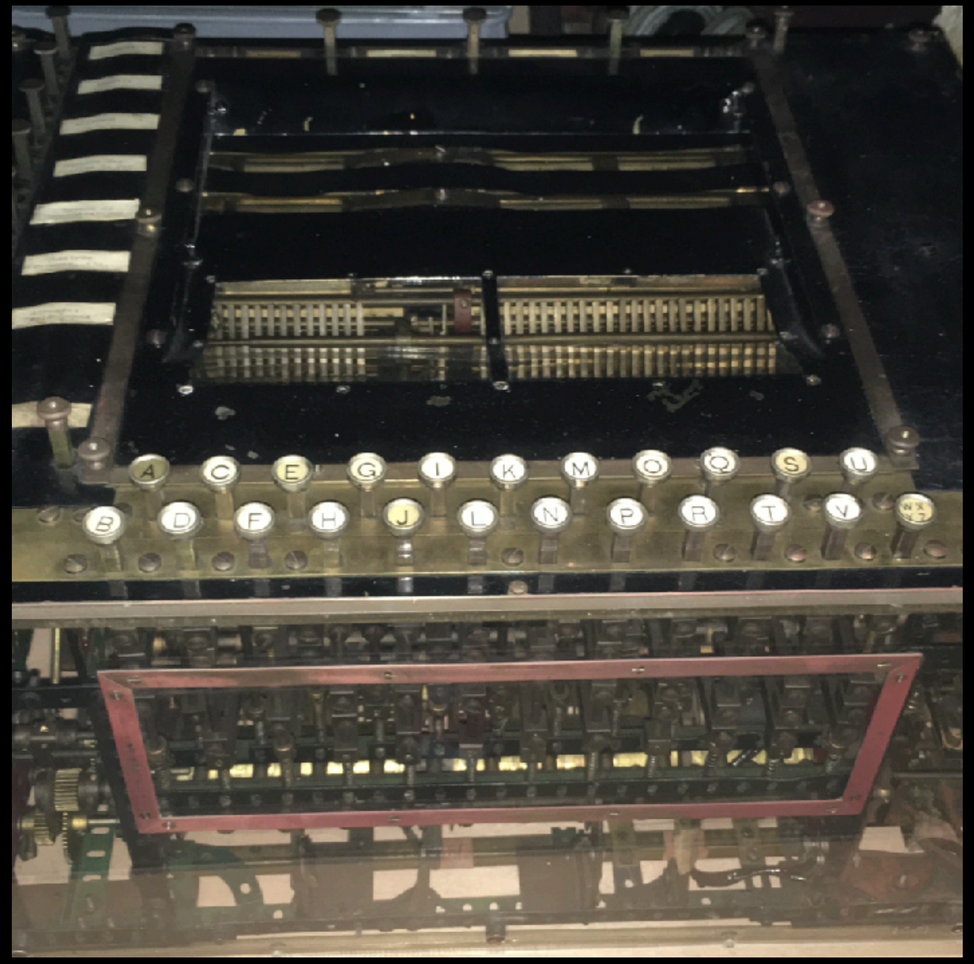
Representation learning and language modelling (1/2)

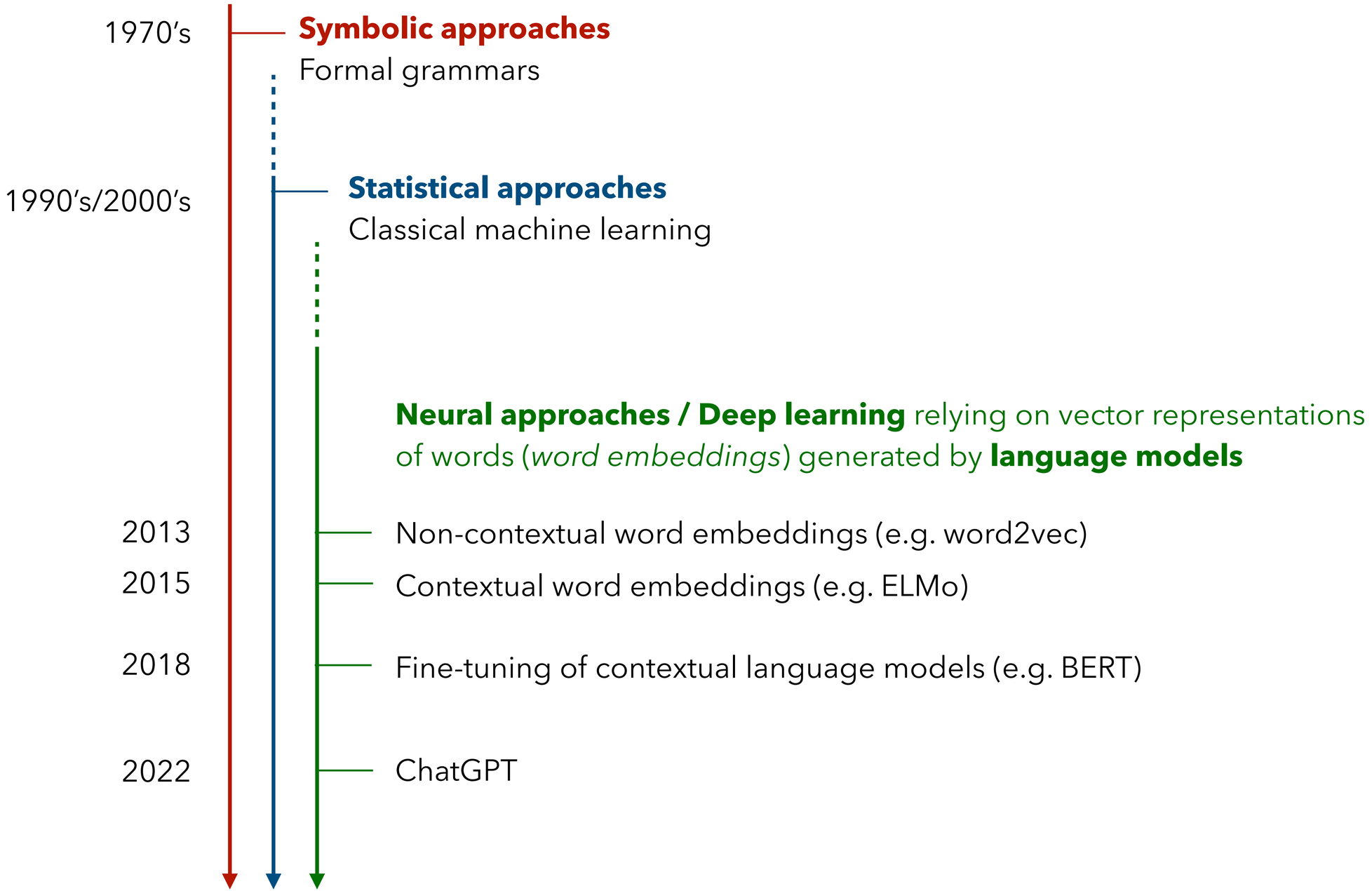
Benoît Sagot

MVA – Speech and Language processing – Class #1.2 – 15th January 2026

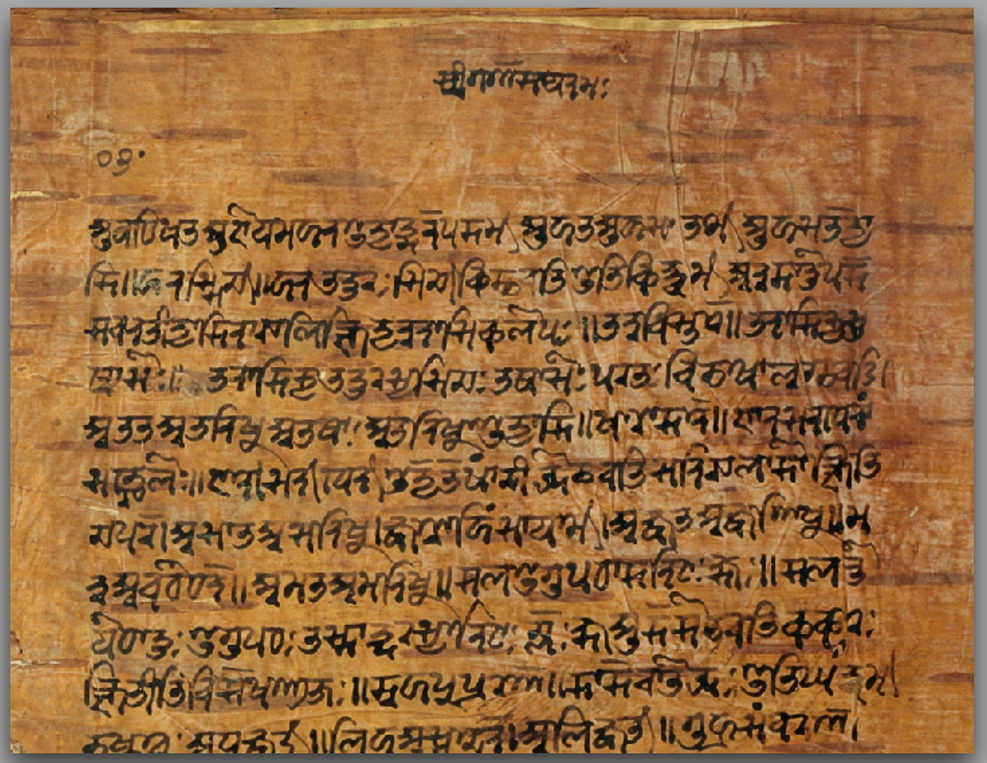
A brief history of NLP



Artsrouni's second "mechanical brain" (1933), preserved at the Musée des Arts et Métiers


1970's
Symbolic approaches
Formal grammars
1990's/2000's
Statistical approaches
Classical machine learning
Neural approaches / Deep learning relying on vector representations of words (word embeddings) generated by language models
2013
Non-contextual word embeddings (e.g. word2vec)
2015
Contextual word embeddings (e.g. ELMo)
2018
Fine-tuning of contextual language models (e.g. BERT)
2022
ChatGPT

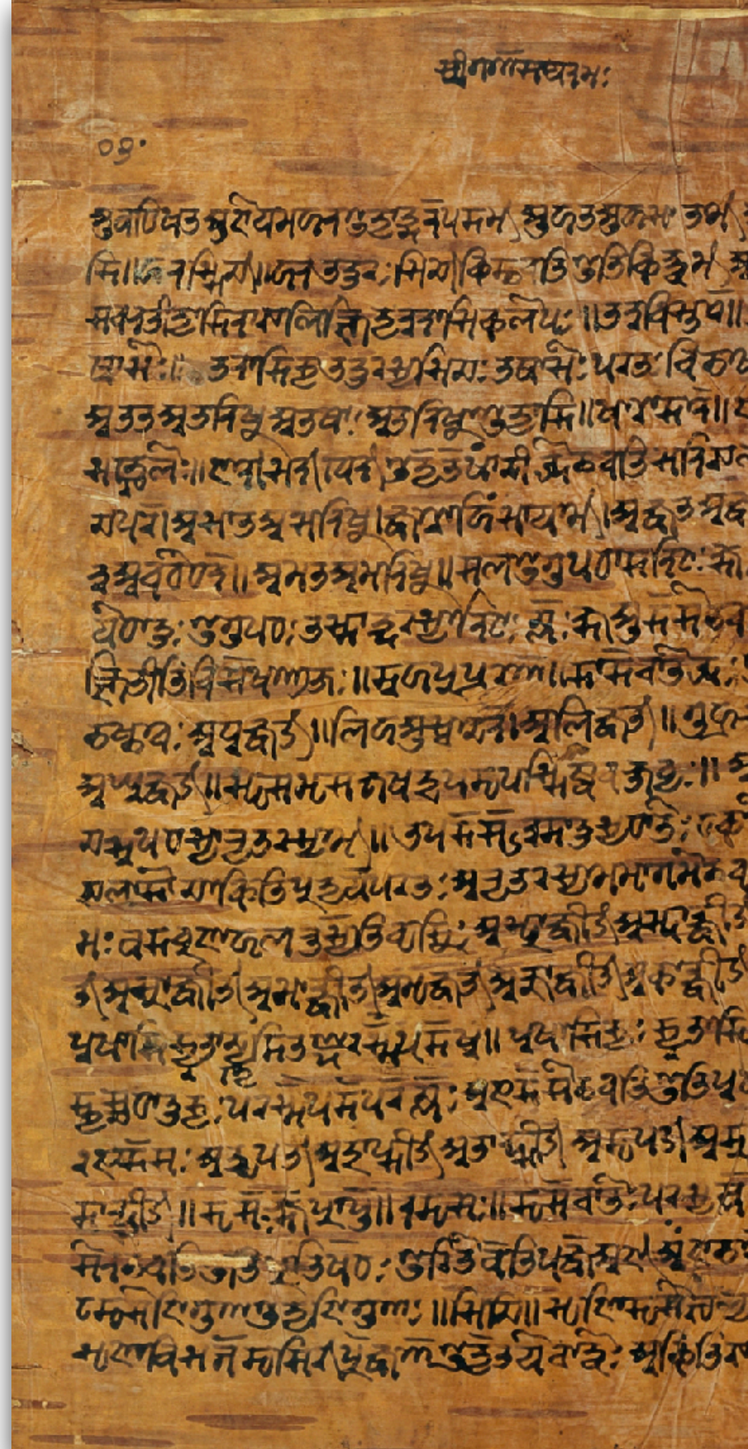
Before NLP



A birch bark manuscript from Kashmir of the Rupavatara, a grammatical textbook based on the Sanskrit grammar of Pāṇini (dated 1663). Source: Wikipedia

Formalising thought and language

- 5th century BC: **Pāṇini formalises the grammar of Sanskrit** using 4,000 algebraic rules in his *Aṣṭādhyāyī*
- The idea of a **universal language** that could be used to combine concepts in a mechanical way is investigated throughout the second millennium (Llull 13th century; Descartes 1629; Leibniz 1666; Wilkins 1668)
- Systems allowing to **translate from one language to another** using an intermediate representation generalise this idea and prefigure word-for-word machine translation (Kircher 1663)
- 1934: **Carnap** (Vienna Circle) introduces a **formal logic-based theory of syntax** (*Logische Syntax der Sprache*)



A birch bark manuscript from Kashmir of the Rupavatara, a grammatical textbook based on the Sanskrit grammar of Pāṇini (dated 1663). Source: Wikipedia

Two forerunners of machine translation

- Two inventors, in 1933, independently patented **translation machines**
  - Georges Artsrouni, for his *mechanical brain* that could process information – including for word-for-word translation
  - In the USSR, Peter Petrovitch Trojanskii

The second machine built by Artsrouni (1933), now at the Musée des Arts et Métiers (Paris)

Symbolic approaches

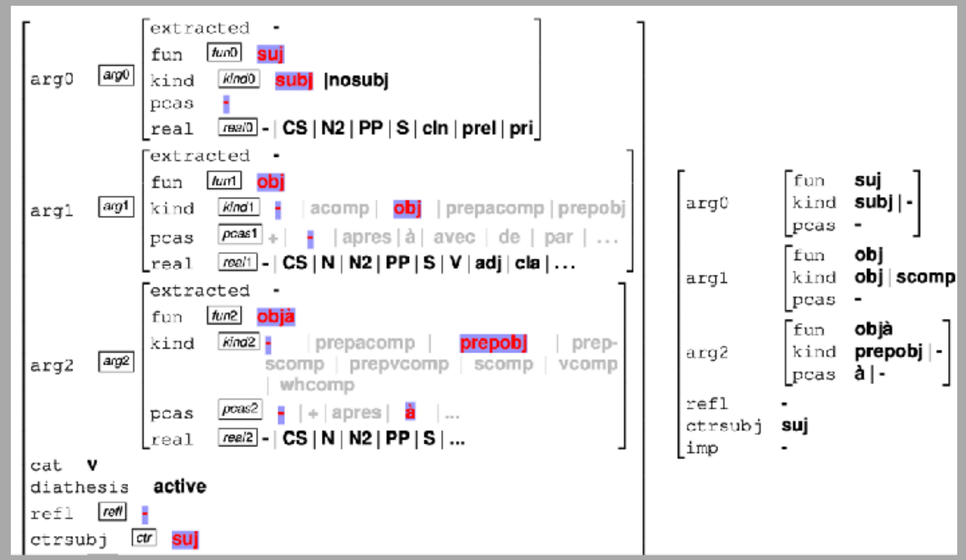
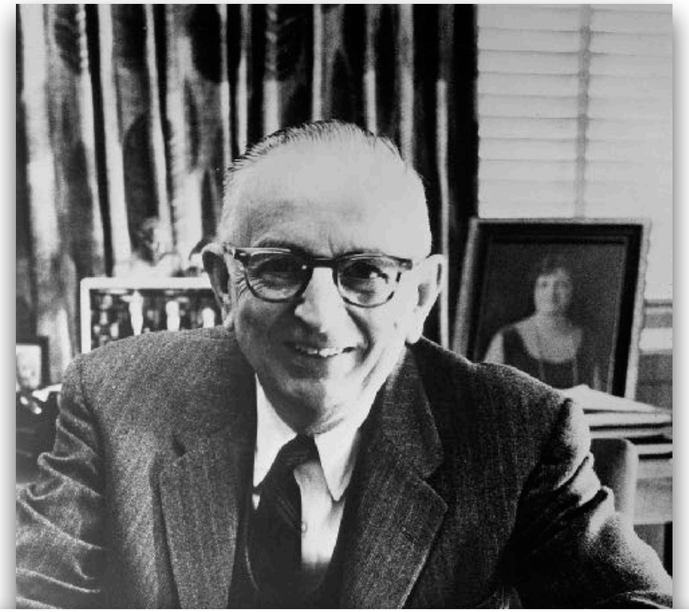


Illustration (Tolone *et al.* 2012): Hypertag from the FRMG grammar (de La Clergerie 2005) and hypertag extracted from the Le*fff* lexicon (Sagot 2010)

The beginnings of machine translation

- From the birth of computing in the 1940s, **machine translation was one of its first applications** (Booth, Weaver, Richens, 1946–1948)
  - **Warren Weaver's *Translation* memorandum (1949)** had a major scientific and political impact
- Machine translation was initially funded by the US military (main objective: Russian → English translation)

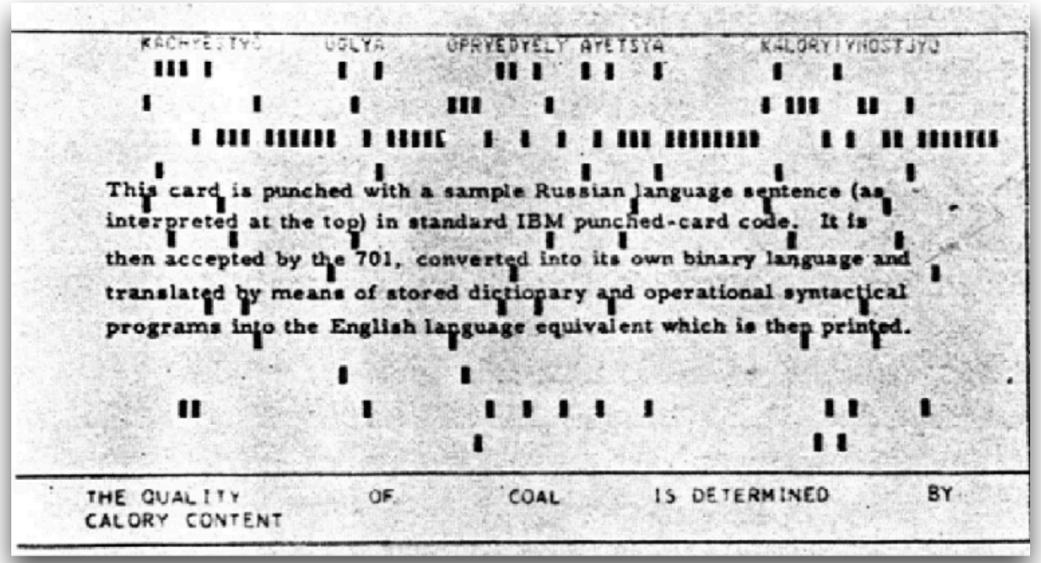


Warren Weaver. Source: Rockefeller Archive Center

[…] one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."
(Weaver, letter to Wiener, 1947)

The beginnings of machine translation

- From the birth of computing in the 1940s, **machine translation was one of its first applications** (Booth, Weaver, Richens, 1946–1948)
  - **Warren Weaver's *Translation* memorandum (1949)** had a major scientific and political impact
- Machine translation was initially funded by the US military (main objective: Russian → English translation)
- The **IBM–Georgetown experiment** (1954) left a strong impression at the time (Hutchins 2006)
  - Around sixty Russian sentences translated into English
  - A program with a vocabulary of 250 words and 6 syntactic rules



Punch card and demonstrators (Hurd, Dostert, Watson) of the IBM–Georgetown experiment (1954)

A wave of optimism

- **This experience gave some people the impression that machine translation was just around the corner.**
  - *"**five, perhaps three, years** hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact."* (IBM press release)
  - *"M. Hurd estime que la traduction parlée est déjà théoriquement réalisable et que **dans quinze ans des traducteurs électroniques** pourront être utilisés dans les assemblées internationales, par exemple aux Nations Unies."* (*Le Monde*, 9th January 1954, C.-G. B.)
- **A parallel can be drawn with some recent predictions about Artificial General Intelligence (AGI)**
  - *"I would be surprised if there is not digital superintelligence in roughly the **five- or six-year timeframe**."* (Musk, July 2023)
  - *"I think we could be just a few years, maybe within **a decade away**[, from the arrival of AGI]"* (Hassabis, May 2023)

Doubts and discouragement

- **Some nuance this enthusiasm**
  - *"of course, it must be emphasized that **a vast amount of work is still needed**, to render mechanically translatable more languages and wider areas of a language. For 250 words and 6 syntactical structures are simply a "Kitty Hawk" flight."* (MacDonald 1954)
  - *"the most widely accepted definition of **AGI** is AI to do any intellectual tasks that the human can. And I do see many companies redefining AGI to other definitions. So for the original definition, **I think we're decades away**."* (Andrew Ng, November 2023, Ground Truths)
- Machine translation in the 1950s was barely different from word-for-word machine translation: **this is insufficient**
- Gradual disappointment of funders and decision-makers, who commission experts to assess the situation
  - **1960: Bar-Hillel report** on machine translation
  - **1966: ALPAC committee report** (*Automatic Language Processing Advisory Committee*, headed by J. R. Pierce)

The machine translation winter and the emergence of NLP

- Conclusion: **high-quality machine translation is an illusion**
    - More linguistic research is required, particularly in syntactic analysis.
    - The “new linguistics”, which is formal and computational **(Chomsky 1957, *Syntactic Structures*)**, is legitimated
  - The only technically feasible and economically viable objective is computer-assisted translation
- **Funding for machine translation dries up**
    - The “new linguistics” and NLP gain autonomy with respect to machine translation
    - Central role of parsing (= syntactic analysis)


LANGUAGE AND MACHINES

COMPUTERS IN TRANSLATION AND LINGUISTICS

A Report by the
Automatic Language Processing Advisory Committee
Division of Behavioral Sciences
National Academy of Sciences
National Research Council

Publication 1416

National Academy of Sciences National Research Council

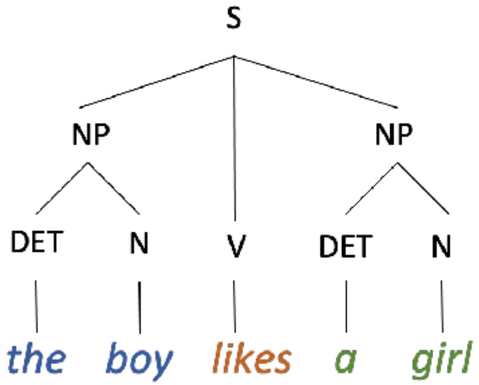
Washington, D. C. 1966

Cover page of the ALPAC report (1966)
National Academies Press

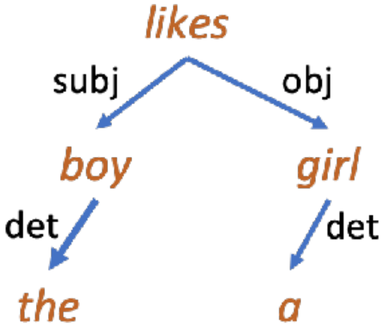
Between linguistics and computer science

Formal grammar and parsing

- A parser takes a sentence as an input and produces a formal representation of its syntactic structure
  - It relies on a **syntactic theory** based on a **formal grammar** that generally models such structures as/based on **trees**
- Most syntactic theories and parsers produce structures based on either **constituents** or **dependencies**
  - Constituents hierarchically structure the tokens in the sentence (leaves of the tree are part-of-speech tags)
  - Dependencies relate each token to its governor
- **Before the 90s, parsers were purely rule-based** (grammatical formalism, grammar, lexicon, disambiguation rules)
  - They often require a pre-processing by a part-of-speech tagger (e.g. Brill 1995, one of the last and best symbolic taggers)



constituency tree

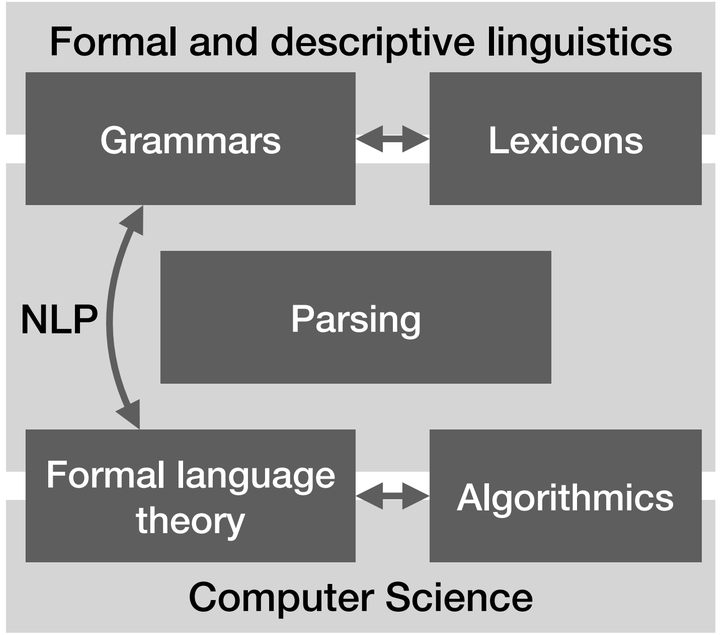


dependency tree

Between linguistics and computer science

Formal grammar and parsing

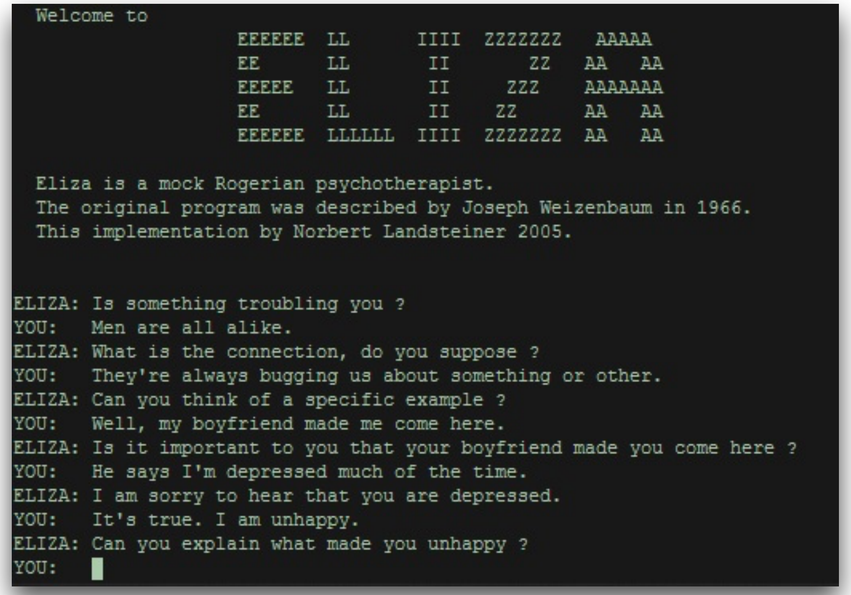
- **Over the following decades, formal syntax, formal language theory and parsing algorithmics fed off each other.**
  - The linguist **Chomsky** introduces as early as 1956 the **hierarchy** named after him, which defines four classes of formal grammars
  - Since then, NLP has significantly contributed to progress in **formal language theory**
- **Increasingly complex parsers** are developed for an increasing number of languages, including French
  - Large-coverage **grammars** and rich morphological and syntactic **lexicons**
  - Efficient **algorithms** and **implementations** (logic programming, optimisations)



Symbolic approaches

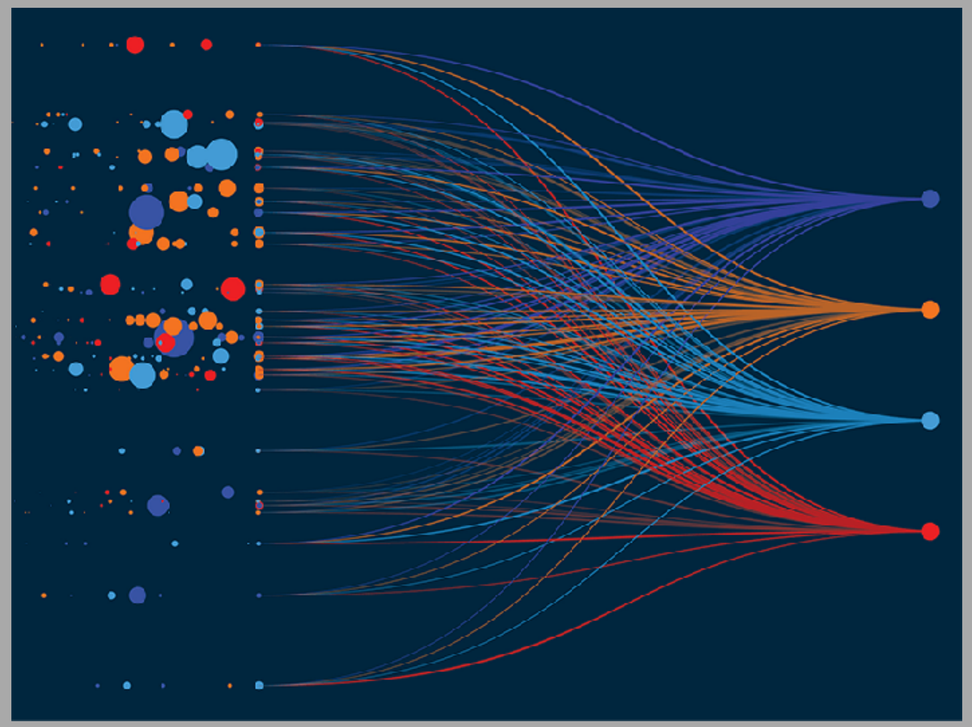
Beyond parsing and machine translation

- The first **conversational agents (or *chatbots*)**, including ELIZA (Weizenbaum 1966) and PARRY (Colby 1972)
  - Anthropomorphisation: the "**ELIZA effect**"
- **Expert systems**, which rely on large quantities of business knowledge encoded formally in the machine
  - 1980s: the golden age of expert systems



Screenshot of a conversation with ELIZA. Source: Wikipedia

Statistical approaches



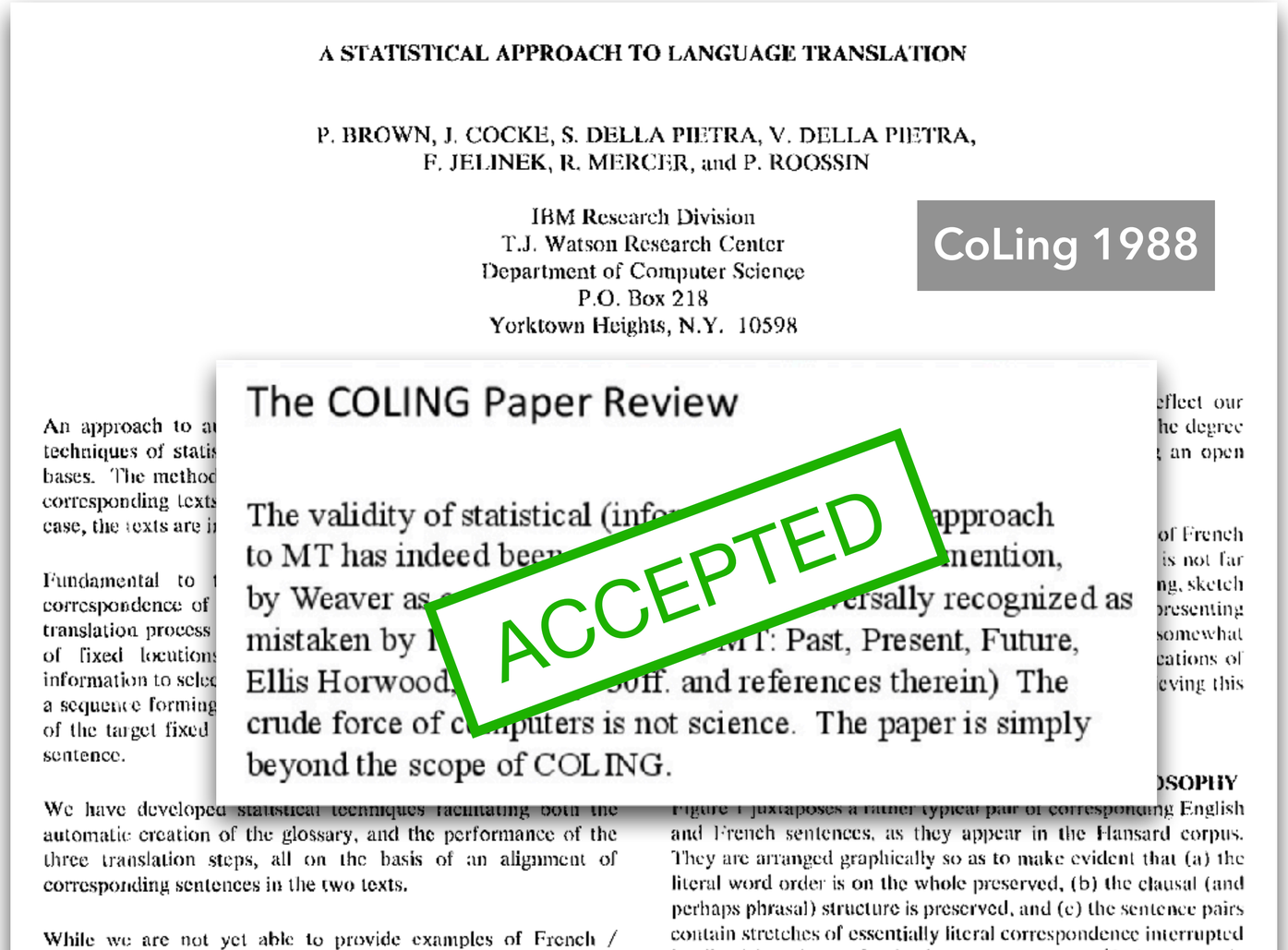
Credit: iStock.com/GarryKillian

Machine learning

The first revolution in NLP

- From speech processing (F. Jelinek at IBM)
- Then applied to machine translation
  - IBM models (Brown *et al.* 1988, 1990, 1993)
- Generalisation to NLP in general:
  - **1990s for English**
  - **2000s for French**

First page of (Brown et al. 1988). More information (and review extract) to be found in (Way 2009)



A STATISTICAL APPROACH TO LANGUAGE TRANSLATION

P. BROWN, J. COCKE, S. DELLA PIETRA, V. DELLA PIETRA,
F. JELINEK, R. MERCER, and P. ROOSSIN

IBM Research Division
T.J. Watson Research Center
Department of Computer Science
P.O. Box 218
Yorktown Heights, N.Y. 10598

CoLing 1988

The COLING Paper Review

The validity of statistical (infor… approach to MT has indeed been… mention, by Weaver as… versally recognized as mistaken by I… MT: Past, Present, Future, Ellis Horwood, …30ff. and references therein) The crude force of c…puters is not science. The paper is simply beyond the scope of COLING.

ACCEPTED

We have developed statistical techniques facilitating both the automatic creation of the glossary, and the performance of the three translation steps, all on the basis of an alignment of corresponding sentences in the two texts.

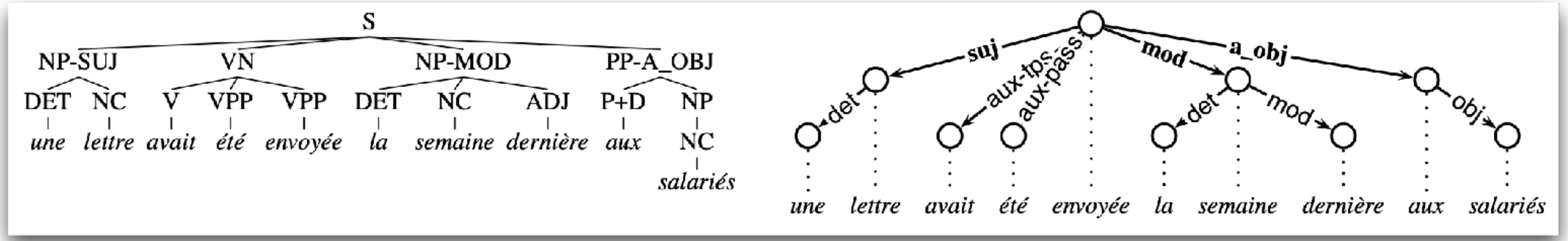
While we are not yet able to provide examples of French /

Figure 1 juxtaposes a rather typical pair of corresponding English and French sentences, as they appear in the Hansard corpus. They are arranged graphically so as to make evident that (a) the literal word order is on the whole preserved, (b) the clausal (and perhaps phrasal) structure is preserved, and (c) the sentence pairs contain stretches of essentially literal correspondence interrupted

Machine learning

The first revolution in NLP

- Two main factors:
  - Sustained increase in **computing power** of available computers
  - **Development of large text corpora** with linguistic annotations
    - First digital corpora: 1950s (Survey of English Usage, Brown Corpus, Trésor de la Langue Française)
    - Treebanks: Penn TreeBank (PTB) 🇺🇸 (Marcus *et al.* 1993), French TreeBank (FTB) 🇫🇷 (Abeillé & Barrier 2004)
    - Annotated corpora for other tasks (named entities, etc.)
- The **linguistic analysis** now lies in the data annotation process



Parse trees for the same sentence in two versions of French TreeBank (original constituent structure on the left, dependency structure on the right). Source: (Candito *et al*. 2010)

Statistical machine translation

(Brown *et al.* 1988, 1990, 1993 – at IBM)

- Let us imagine we want to translate from French to English
  - We want to find the best English translation y of a French sentence x
  - We model the quality of a translation by a probability $P(y \mid x)$
  - We are therefore looking for: $\text{argmax}_y P(y \mid x)$
- We use Bayes' formula to decompose this probability into two separate components:

$$= \text{argmax}_y P(x \mid y) \cdot P(y)$$

Translation model

Models how words and word sequences must be translated to preserve meaning.

Trained on parallel corpora.

Language model

Models how to produce sentences in correct English.

Trained on monolingual corpora.

Language models

- A language model is a **probability distribution over sequences of words**
  - It quantifies the fact that the sequence *Ceci est une phrase correcte.* is more probable that the sequence *a Le llkdhfeee correcte est est*
  - This can be used to
    - **Find the most probable next word** given an input sequence
      - Example: *En raison de la grève, Pierre est arrivé en* [ ]
    - **Find the most probable word to fill a gap** in a sequence
      - Example: *En raison de la* [ ] *, Pierre est arrivé en retard.*

Language models

- A language model is a **probability distribution over sequences of words**
  - It quantifies the fact that the sequence *Ceci est une phrase correcte.* is more probable that the sequence *a Le llkdhfeee correcte est est*
  - This can be used to
    - **Find the most probable next word** given an input sequence
      - Example: *En raison de la grève, Pierre est arrivé en* avance
    - **Find the most probable word to fill a gap** in a sequence
      - Example: *En raison de la* , *Pierre est arrivé en retard.*

Language models

- A language model is a **probability distribution over sequences of words**
  - It quantifies the fact that the sequence *Ceci est une phrase correcte.* is more probable that the sequence *a Le llkdhfeee correcte est est*
  - This can be used to
    - **Find the most probable next word** given an input sequence
      - Example: *En raison de la grève, Pierre est arrivé en* *marchant*
    - **Find the most probable word to fill a gap** in a sequence
      - Example: *En raison de la* *, Pierre est arrivé en retard.*

Language models

- A language model is a **probability distribution over sequences of words**
  - It quantifies the fact that the sequence *Ceci est une phrase correcte.* is more probable that the sequence *a Le llkdhfeee correcte est est*
  - This can be used to
    - **Find the most probable next word** given an input sequence
      - Example: *En raison de la grève, Pierre est arrivé en* *taxi*
    - **Find the most probable word to fill a gap** in a sequence
      - Example: *En raison de la* *, Pierre est arrivé en retard.*

Language models

- A language model is a **probability distribution over sequences of words**
  - It quantifies the fact that the sequence *Ceci est une phrase correcte.* is more probable that the sequence *a Le llkdhfeee correcte est est*
  - This can be used to
    - **Find the most probable next word** given an input sequence
      - Example: *En raison de la grève, Pierre est arrivé en retard*
    - **Find the most probable word to fill a gap** in a sequence
      - Example: *En raison de la , Pierre est arrivé en retard.*

Language models

- A language model is a **probability distribution over sequences of words**
  - It quantifies the fact that the sequence *Ceci est une phrase correcte.* is more probable that the sequence *a Le llkdhfeee correcte est est*
  - This can be used to
    - **Find the most probable next word** given an input sequence
      - Example: *En raison de la grève, Pierre est arrivé en retard*
    - **Find the most probable word to fill a gap** in a sequence
      - Example: *En raison de la grève , Pierre est arrivé en retard.*

Language models

- A language model is a **probability distribution over sequences of words**
  - It quantifies the fact that the sequence *Ceci est une phrase correcte.* is more probable that the sequence *a Le llkdhfeee correcte est est*
  - This can be used to
    - **Find the most probable next word** given an input sequence
      - Example: *En raison de la grève, Pierre est arrivé en* *retard*
    - **Find the most probable word to fill a gap** in a sequence
      - Example: *En raison de la* *panne* *, Pierre est arrivé en retard.*

Language models

- A language model is a **probability distribution over sequences of words**
  - It quantifies the fact that the sequence *Ceci est une phrase correcte.* is more probable that the sequence *a Le llkdhfeee correcte est est*
  - This can be used to
    - **Find the most probable next word** given an input sequence
      - Example: *En raison de la grève, Pierre est arrivé en retard*
    - **Find the most probable word to fill a gap** in a sequence
      - Example: *En raison de la circulation, Pierre est arrivé en retard.*

Language models

- A language model is a **probability distribution over sequences of words**
  - It quantifies the fact that the sequence *Ceci est une phrase correcte.* is more probable that the sequence *a Le llkdhfeee correcte est est*
  - This can be used to
    - **Find the most probable next word** given an input sequence
      - Example: *En raison de la grève, Pierre est arrivé en retard*
    - **Find the most probable word to fill a gap** in a sequence
      - Example: *En raison de la pluie , Pierre est arrivé en retard.*

Conditional language models

- The probability assigned by a language model to a sequence can depend from another sequence
  - **Conditional language models**
  - There are different types of link between the main sequence and the conditioning sequence
    - Translation, summarisation, question answering
  - Example:
    Conditioning sequence: *Because of the strike, Pierre arrived by bike*
    Main sequence: *En raison de la grève, Pierre est arrivé en* [redacted]

Conditional language models

- The probability assigned by a language model to a sequence can depend from another sequence
  - **Conditional language models**
  - There are different types of link between the main sequence and the conditioning sequence
    - Translation, summarisation, question answering
  - Example:
    Conditioning sequence: *Because of the strike, Pierre arrived by bike*
    Main sequence: *En raison de la grève, Pierre est arrivé en* *retard*

Conditional language models

- The probability assigned by a language model to a sequence can depend from another sequence
  - **Conditional language models**
  - There are different types of link between the main sequence and the conditioning sequence
    - Translation, summarisation, question answering
  - Example:
    Conditioning sequence: *Because of the strike, Pierre arrived by bike*
    Main sequence: *En raison de la grève, Pierre est arrivé en* *moto*

Conditional language models

- The probability assigned by a language model to a sequence can depend from another sequence
  - **Conditional language models**
  - There are different types of link between the main sequence and the conditioning sequence
    - Translation, summarisation, question answering
  - Example:
    Conditioning sequence: *Because of the strike, Pierre arrived by bike*
    Main sequence: *En raison de la grève, Pierre est arrivé en bicyclette*

Conditional language models

- The probability assigned by a language model to a sequence can depend from another sequence
  - **Conditional language models**
  - There are different types of link between the main sequence and the conditioning sequence
    - Translation, summarisation, question answering
  - Example:
    Conditioning sequence: *Because of the strike, Pierre arrived by bike*
    Main sequence: *En raison de la grève, Pierre est arrivé en vélo*

Intrinsic evaluation: perplexity

- **A good model assigns high probabilities to correct sequence it has not seen during training**
  - This shows a good generalisation ability

$$PP(w_1 \ldots w_n) = \exp\left(\frac{1}{n}\sum_{i=1}^{n} -\log P\left(w_i \,|\, w_i\text{'s context}\right)\right)$$

- Can be used to compare two language models: **lower perplexity is better**
- But:
  - It only really works on data that is similar to the training data
  - You can use it to compare models only if they share everything apart from the weights (e.g. the vocabulary…)

Assessing the probability of a sequence

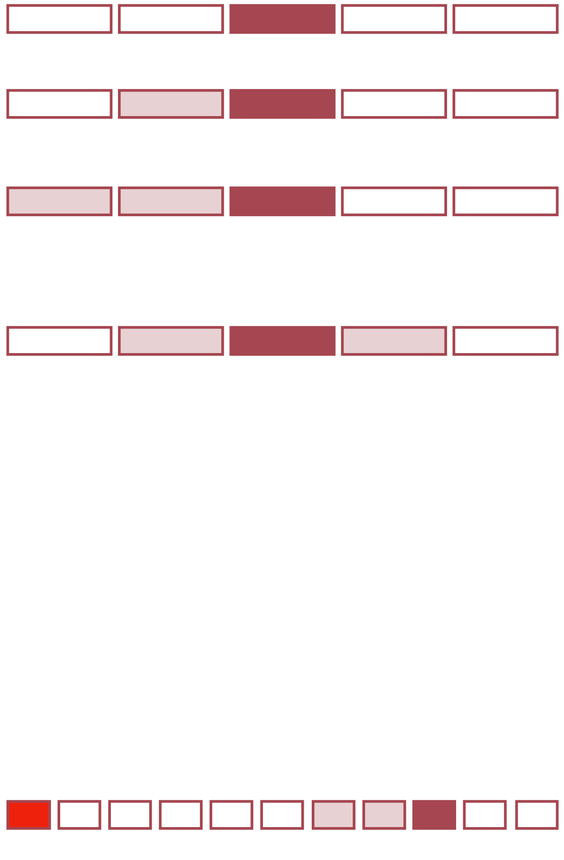
- Even with a gigantic corpus, most sequences will be so rare that their probability cannot be extracted directly
  - And what about sequences never before seen in the corpus?

- **The probability $P(w_1 \ldots w_k)$ of a sequence $w_1 \ldots w_k$ is modelled based on subsequences**
  - Chain rule: $P(w_1 \ldots w_k) = P(w_1|\#)\ P(w_2|\#w_1)\ P(w_3|\#w_1w_2)\ \ldots\ P(w_k|\#w_1 \ldots w_{k-1})$
  - Many ways to compute $P(w_i|w_1 \ldots w_{i-1})$, often using an approximation

n-gram Markov models

(Shannon 1948)

- Unigram (non contextual) model : $P_1(w_1w_2w_3) = P(w_1)P(w_2)P(w_3)$
- Bigram (Markov) model: $P_2(w_1w_2w_3) = P(w_1|\#)P(w_2|w_1)P(w_3|w_2)$
- *n*-gram (Markov) model: $P_n(w_i|\#w_1\ldots w_{i-2}w_{i-1}) = \ldots P(w_i|w_{i-n+1}\ldots w_{i-1})$
- Bidirectional Markov models:
  $P_{2,BD}(w_1w_2w_3) = P(w_1|\#)P'(w_1|w_2)\ P(w_2|w_1)P'(w_2|w_3)\ P(w_3|w_2)P'(w_3|\#)$
- Probabilities can be computed by **counting** occurrences of *n*-grams in a large corpus:
  $P(w_i) = occ(w_i) / N \quad ; \quad P(w_i|w_{i-n+1}\ldots w_{i-1}) = occ(w_{i-n+1}\ldots w_i) / occ(w_{i-n+1}\ldots w_{i-1})$
  - Maximum likelihood
  - Very fast to train (one pass over training data) and to execute, easy to interpret
  - Local context is not always enough
    Example: *Alice went to the veterinary to pick up [her/his] dog.*



n-gram language models

- N-gram language models (generally improved with smoothing, interpolation and smoothing techniques):
  - State-of-the-art until the early 2010's (Goodman 2001)
- Software packages – widely used in machine translation until the mid-2010s, still used today when extreme speed is needed (e.g. to filter huge corpora)
  - SRILM (Stolcke 2002)
  - KenLM (Heafield 2011)
  - IRSTLM (Federico *et al.* 2008)

Statistical part-of-speech tagging and lemmatisation

- Amongst the main successful task-specific models are **part-of-speech tagging and lemmatisation models**
  - **Sequence labelling problems**
    - PoS tagging = assign a part-of-speech to each word (noun, verb, prep…)
    - Lemmatisation = assign the dictionary header of each word (its lemma)

books are useful

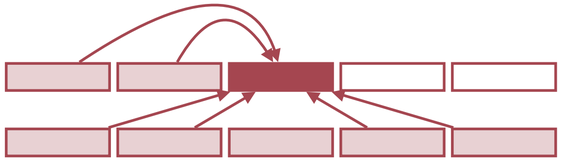
books.N are.V useful.A

book.N be.V useful.A

  - **Commonly used architectures**
    - Hidden Markov Models (HMMs)
    - Maximum Entropy (MaxEnt) models (Ratnaparkhi 1996, 1999, Toutanova *et al.* 2003)
    - Conditional Random Fields (CRFs) (Lafferty *et al.* 2001)
  - Usual features: the word itself, surrounding words, previous tags, sometimes lexical information

Predictive language models

- **Prediction-based approach**: Given a position and a context, a model can be trained to predict which word would best fit said position
  - It will associate a weight (or a probability) to each possible word
  - Context restricted to the left context (autocompletion)
    → **generative language model**
  - Full context
    → **masked language model**
  - The language model can be **conditional**
    - Most conditional language models are generative



Towards neural approaches

- **Underlying rationale** (Manning 2024):
  - Language is the symbolic system *par excellence*; we should study and make use of its symbolic structure
  - This does not show that the main processor of these symbols, the human brain, is implemented as a physical symbol system
  - We need not design NLP systems as physical symbol systems
  - The brain is more like a neural network model
  - Artificial neural network models scale better and can capture the world represented by symbols
- Yet **artificial neural networks are only very distantly inspired by the human brain**
  - Interface between research on neural language models and on neuroscience = active research field

DALL·E 3 (via ChatGPT, GPT-4, 19/11/2023)