Large language models and human language processing

Benoît Crabbé

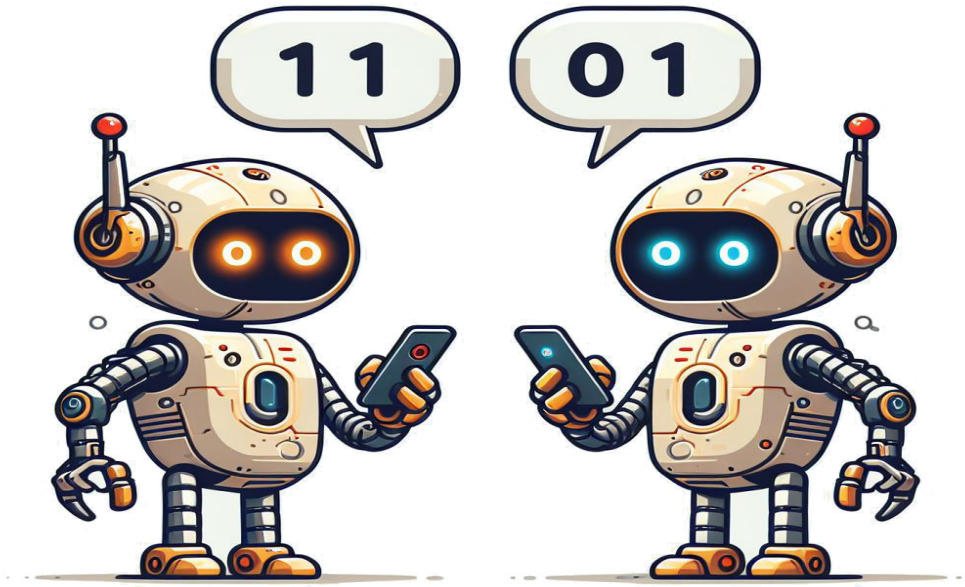
An information theoretic view

Academic year 2024-2025

Laboratoire de linguistique formelle

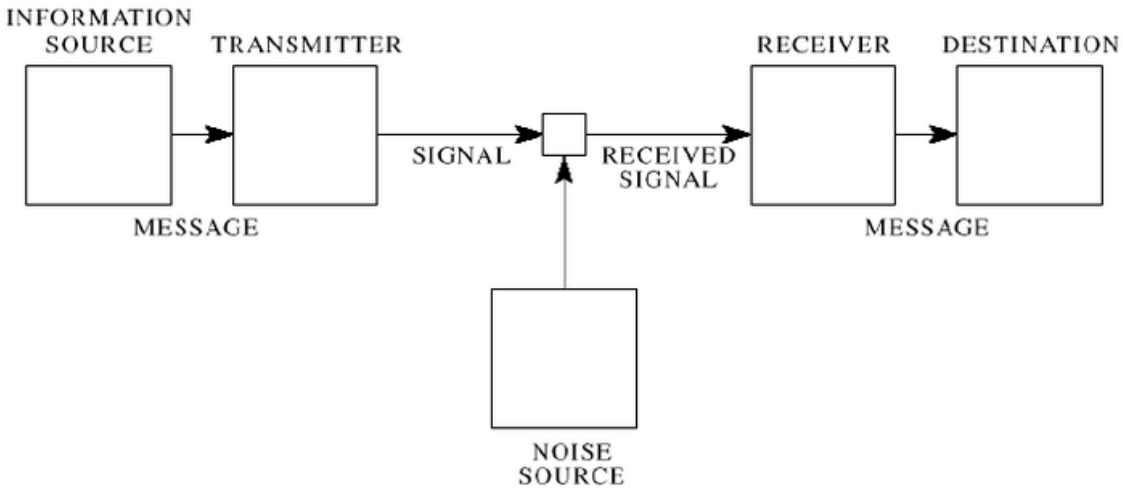
Outline

1. Entropy
2. The entropy of English
3. Relation to human processing
4. Memory limitations
5. Lossy memory models



Mathematical communication

- We consider the hypothesis. Natural language is a communication system in the sense of Shannon (1948):

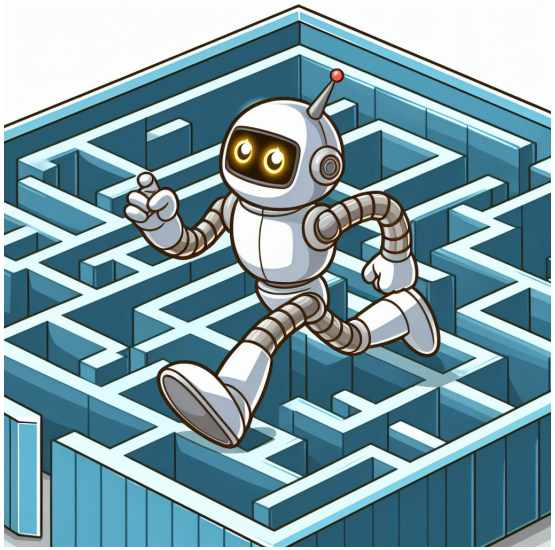


  Natural language codes messages that are sent from a transmitter to a receiver with some encoding such as written or spoken.

- By convention, and to abstract away from details of the actual coding alphabet, we rely on binary codes, that is sequences of 0 and 1

Efficient coding

We assume coding relies on a codebook that is a bijective function mapping each message to a code and vice-versa.
In information theory one seeks to design **optimally efficient codes**: more frequent messages should require cheaper transmission costs and should be coded with fewer bits. The length of a code is a function of the probability of the message.



Message	**s**tart	**e**nd	**f**orward	**b**ackward	**l**eft	**r**ight
Code	0001	0000	1	001	011	010

Application to natural language: working hypothesis

Information theory provides mathematical bounds for **optimal compression**. Yet we assume humans to be suboptimal communicators. Humans are limited by their biology and/or cognitive system.

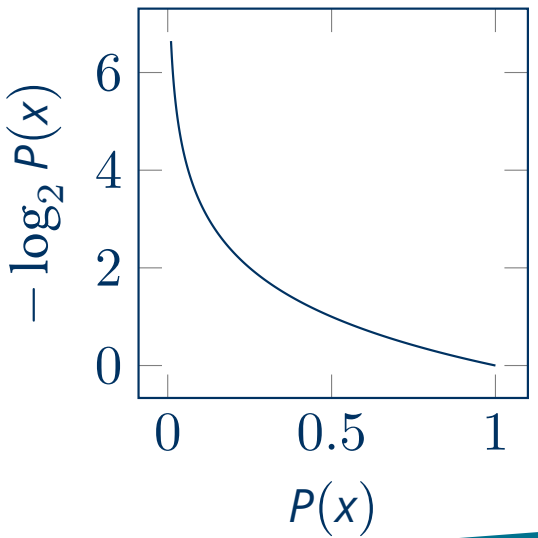
Information content or surprisal

- The quantity of information carried out by a message x is its **information content** or **surprisal**

$$S(x) = -\log_2 P(x)$$

Message	**s**tart	**e**nd	**f**orward	**b**ackward	**l**eft	**r**ight
$P(m)$	$\frac{1}{8}$	0	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
$S(m)$	3	∞	1	3	3	3

- The less probable the message, the more information it carries, the most probable the message the less information is carried:

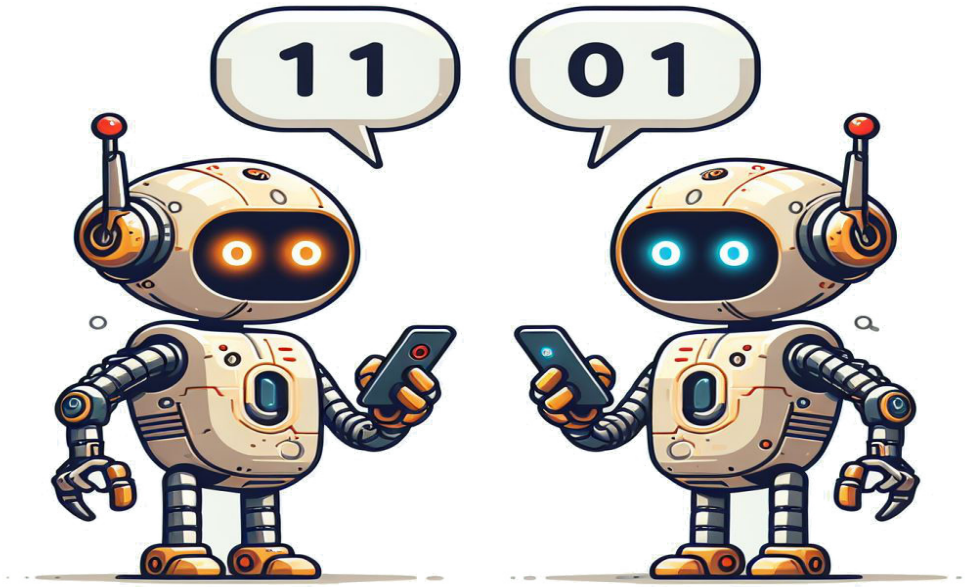


Interpretation

Surprisal represents the number of bits required for encoding this message optimally

Outline

1. Entropy
2. The entropy of English
3. Relation to human processing
4. Memory limitations
5. Lossy memory models



The entropy of English

- We view the English language as a sequence of word tokens, each of them is generated by a random variable $X_1 \ldots X_n$. That is, we see the English language as a sequence of randomly generated messages. Example:

X_1	X_2	X_3	X_4	X_5	X_6
the	cat	sleeps	on	the	mat

  we write its probability $P(X_1 = \mathit{the}, X_2 = \mathit{cat}, X_3 = \mathit{sleeps}, X_4 = \mathit{on}, X_5 = \mathit{the}, X_6 = \mathit{mat})$ or more compactly $P(\mathit{the}, \mathit{cat}, \mathit{sleeps}, \mathit{on}, \mathit{the}, \mathit{mat})$

Bibliography

Note that the methodology described here adapts to word sequences the character model of Shannon (1951).

Language models

- The probability of a sequence of symbols $\mathrm{x} = x_1 \ldots x_n$ is given by the **chain rule** of probability:

$$P(\mathrm{x}) = P(x_1, \ldots x_n) = \prod_{i=1}^{n} P(x_i | x_1 \ldots x_{i-1})$$

  The conditional $P(x_i | x_1 \ldots x_{i-1})$ reads as "the probability of **generating** x_i given that we know $x_1 \ldots x_{i-1}$". The longer the context the easier it is to guess x_i, example:
  - the ...
  - the cat ...
  - the cat chases the ...

- A **language model** is a distribution of probability over a set $\mathcal{L}$ of variable length sequences, a formal language, such that:

$$0 \leq P(\mathrm{x}) \leq 1$$

$$\sum_{x \in \mathcal{L}} P(x) = 1$$

Markov assumption

- The number of parameters of a joint model grows exponentially with the size of the sequence, example:
  - Consider the family of languages $\mathcal{L} = \{a, b, c\}^n$. In case $n = 1$, we need 3^1 parameters. In case $n = 2$ we need 3^2 parameters and in general we need 3^n parameters.
- To reduce the number of parameters, a **markov assumption** sets the boundary of the context and simplifies the chain rule:

$$P(\mathrm{x}) \approx \prod_{i=1}^{n} P(x_i | x_{i-k} \ldots x_{i-1})$$

The conditionals now have a limited context of order k.

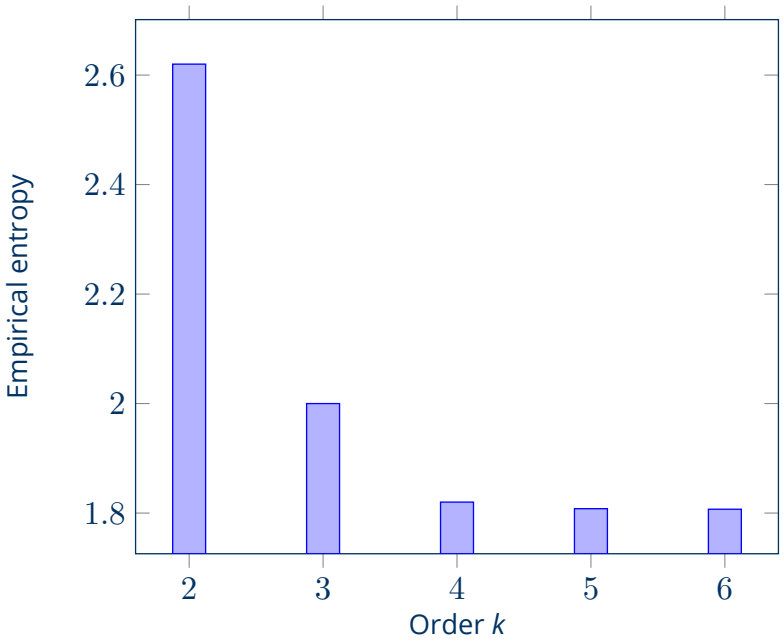
Example probability estimation from corpora

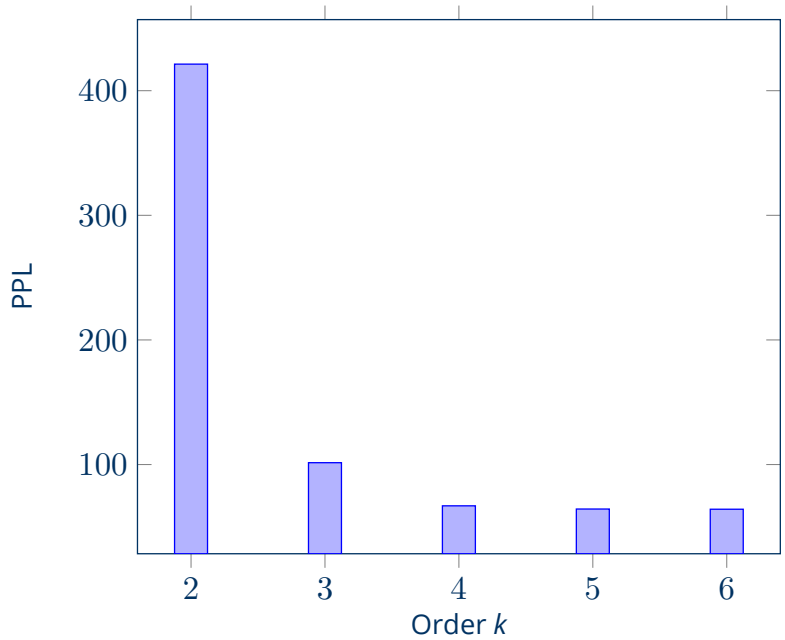
When $k = 2$ the factors are of the form $P(x_i | x_{i-2} x_{i-1})$ and are computed with the usual definition of conditional probabilities: $P(x_i | x_{i-2}, x_{i-1}) = \frac{P(x_{i-2}, x_{i-1}, x_i)}{P(x_{i-2}, x_{i-1})}$

Effect of the length of the context

Entropy of Shakespeare

- To illustrate the effect 'the longer the context the easier it is to predict the next word' we compute the empirical entropy and the perplexity of Shakespeare works texts with different sizes of context





$$H[\text{shakespeare}] = -\frac{1}{n}\sum_{i=1}^{n} \log_2 P(x_i|x_{i-k}\cdots x_{i-1})$$

$$PPL[\text{shakespeare}] = 2^{-\frac{1}{n}\sum_{i=1}^{n} \log_2 P(x_i|x_{i-k}\cdots x_{i-1})}$$

Effect of the length of the context

Random generation

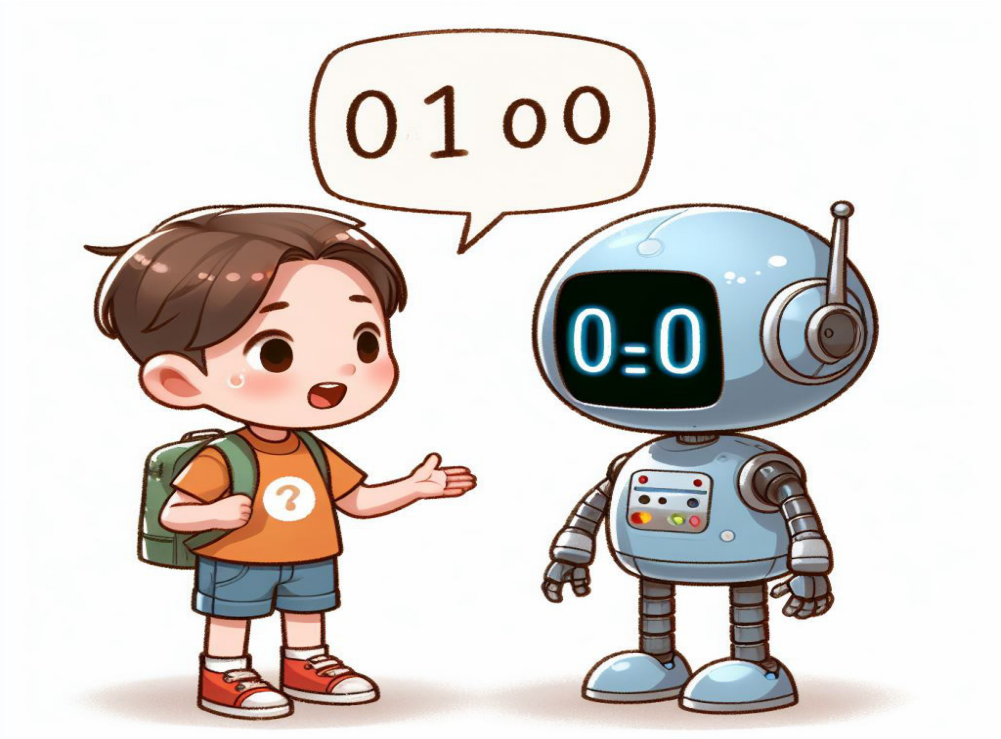
- k = 1 (bigram model) : The which he is in the high and our words. I would have a brave me that with a black in
- k = 2 (trigram model) : The Duke of Clarence; next to the Moor, and others as many as you to the point. O, the Lord of
- k = 3 (4-gram model) : The time to come. This you may know by my size of words. I marvel thy master hath not the life
- k = 4 (5-gram model) : The holy suit which fain it would convince, for the love of a guinea hen, I will not do them the
- k = 5 (6-gram model): The more of you 'twas felt, the more it shaped for sportive tricks, which are their own right by the law
- k = 6 (7-gram model): The Prince of Wales doth join with all the world to nothing with this answer, Hamlet. These are my ministers and

Sampling procedure

$x_i \sim P(x_i | x_{i-k} \dots x_{i-1})$

Outline

1. Entropy
2. The entropy of English
3. Relation to human processing
4. Memory limitations
5. Lossy memory models



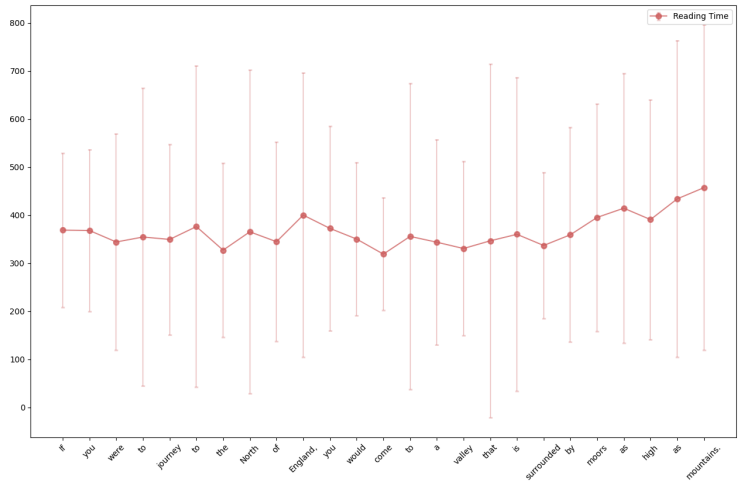
Self paced reading

- Self paced reading is a task where readers are asked to read a text where words are hidden except the current word. To read the next word, the reader has to push a button.

  Here __ __ _____ __ _____ _____

- Reading times are measured by timing the button pushes

Example



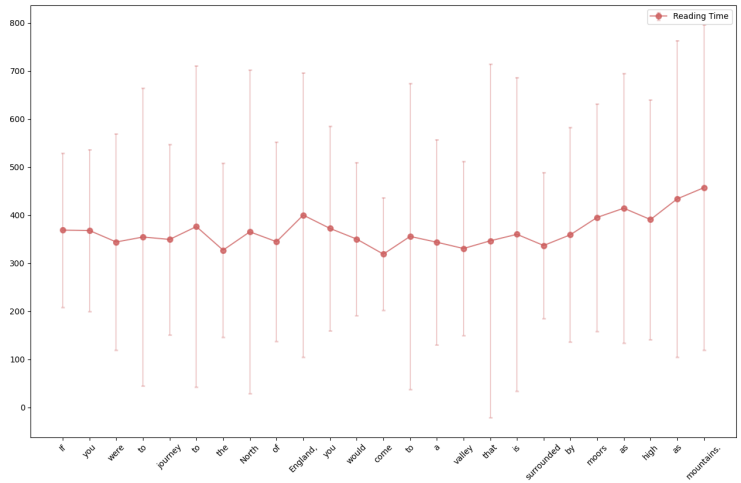
Self paced reading

- Self paced reading is a task where readers are asked to read a text where words are hidden except the current word. To read the next word, the reader has to push a button.

_____ is __ _____ __ _____ _____

- Reading times are measured by timing the button pushes

Example



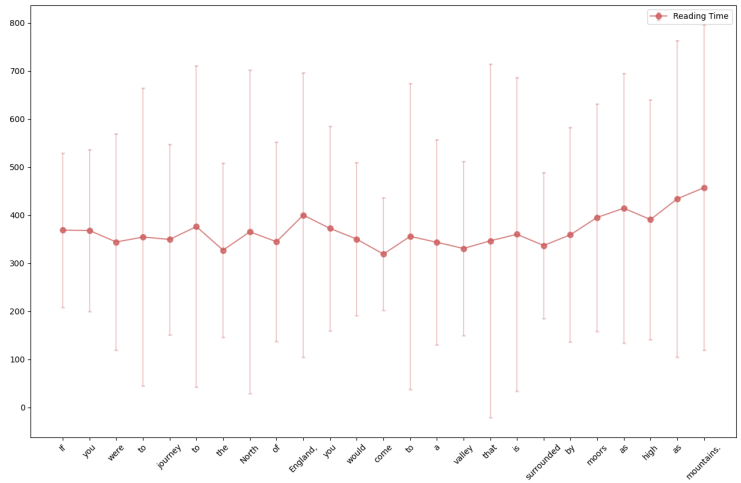
Self paced reading

- Self paced reading is a task where readers are asked to read a text where words are hidden except the current word. To read the next word, the reader has to push a button.

_____ __ an _____ ___ _____ _____

- Reading times are measured by timing the button pushes

Example



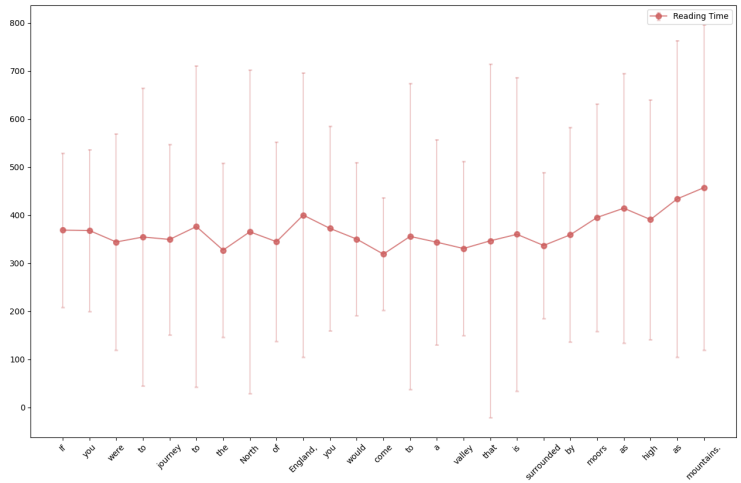
Self paced reading

- Self paced reading is a task where readers are asked to read a text where words are hidden except the current word. To read the next word, the reader has to push a button.

_____ __ ___ example __ _____ _____

- Reading times are measured by timing the button pushes

Example



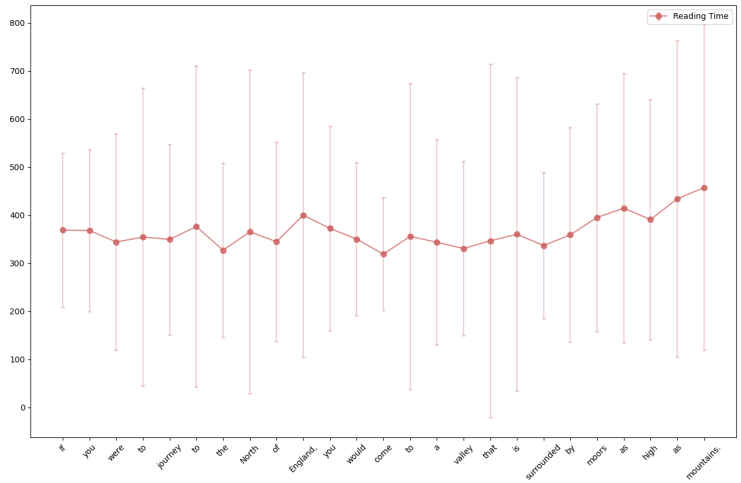
Self paced reading

- Self paced reading is a task where readers are asked to read a text where words are hidden except the current word. To read the next word, the reader has to push a button.

_____ __ __ _____ of _____ _____

- Reading times are measured by timing the button pushes

Example



Laboratoire de linguistique formelle

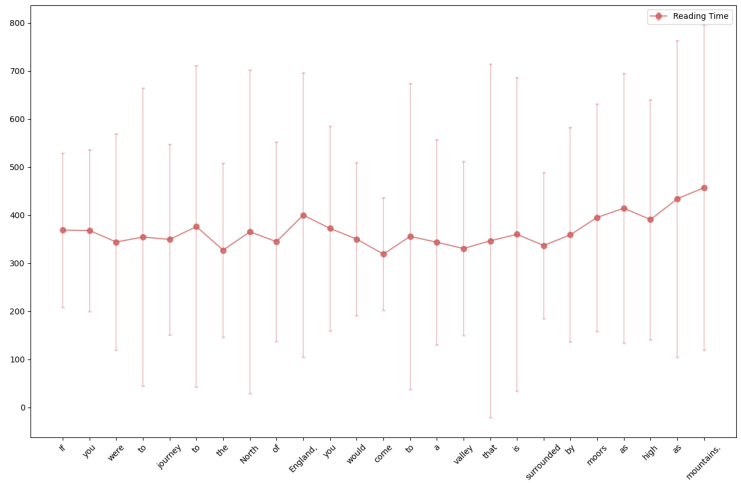
Self paced reading

- Self paced reading is a task where readers are asked to read a text where words are hidden except the current word. To read the next word, the reader has to push a button.

_____ __ __ _____ __ self-paced _____

- Reading times are measured by timing the button pushes

Example



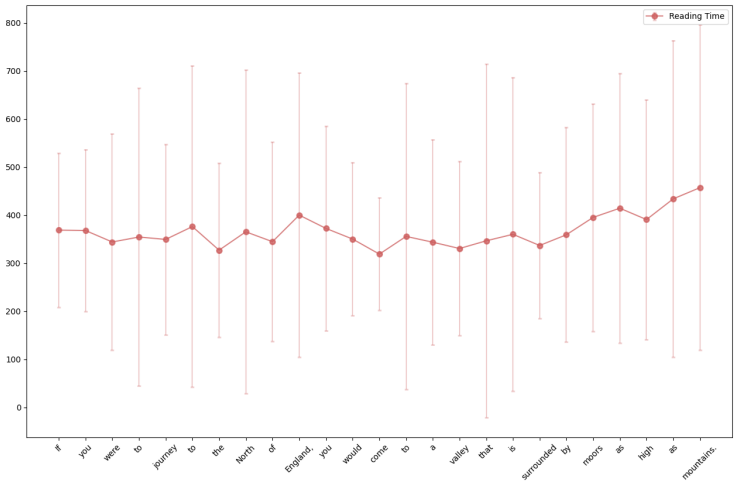
Self paced reading

- Self paced reading is a task where readers are asked to read a text where words are hidden except the current word. To read the next word, the reader has to push a button.

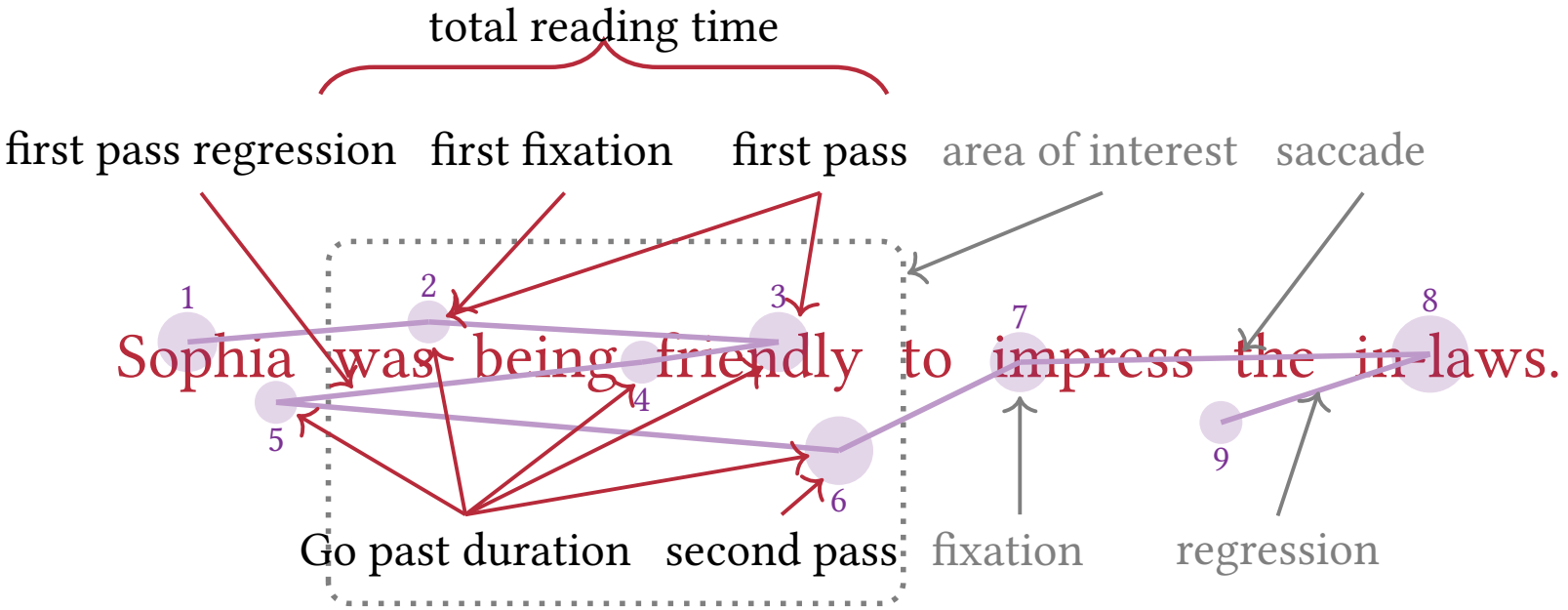
_____ ___ ___ _____ ___ _____ reading

- Reading times are measured by timing the button pushes

Example



Eye tracking



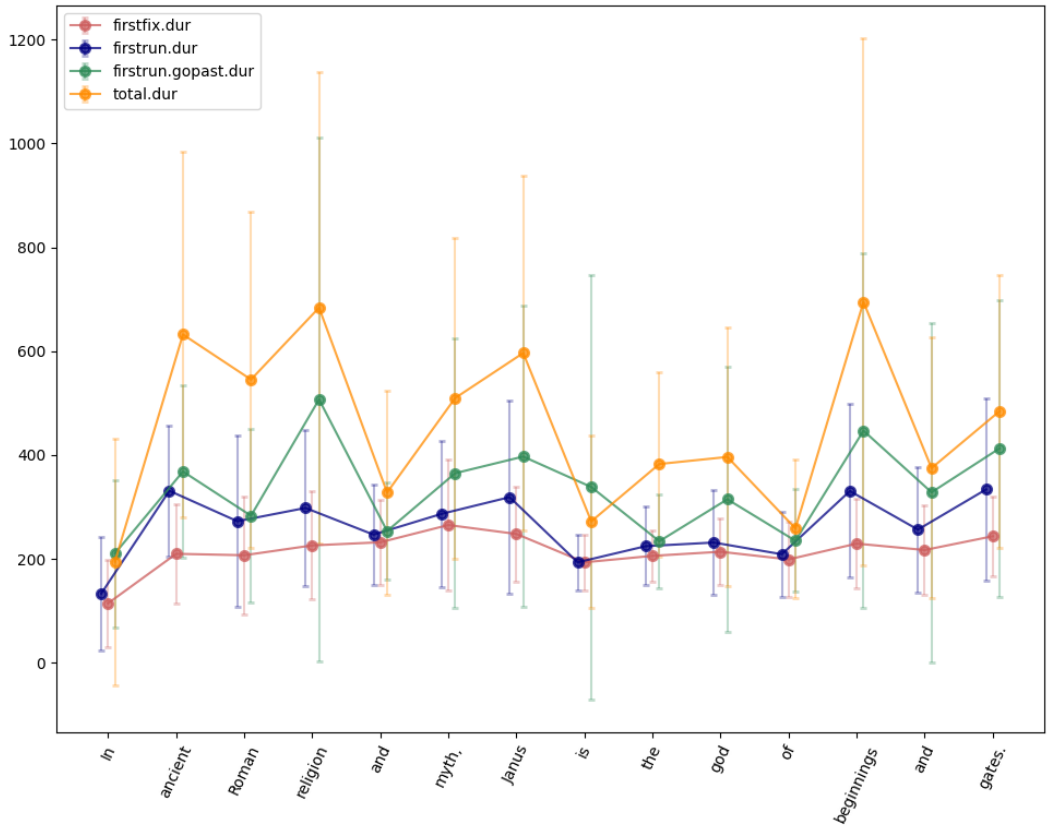
Common eye tracking metrics

The eye tracker naturally generates a time series of x, y coordinates on a screen. For reading purposes, here are common metrics usually computed from the raw measures:

Name	Description	Example
First fixation duration	First fixation in the AOI	2
First pass duration	Sum of first pass fixations in the AOI	2+3
First pass go past duration	Sum of all fixations before leaving the AOI rightwards	2+3+4+5+6
Total duration	Sum of all fixations in the AOI	2+3+4+6

Example

From the MECO corpus (Siegelmann et al. 2022)

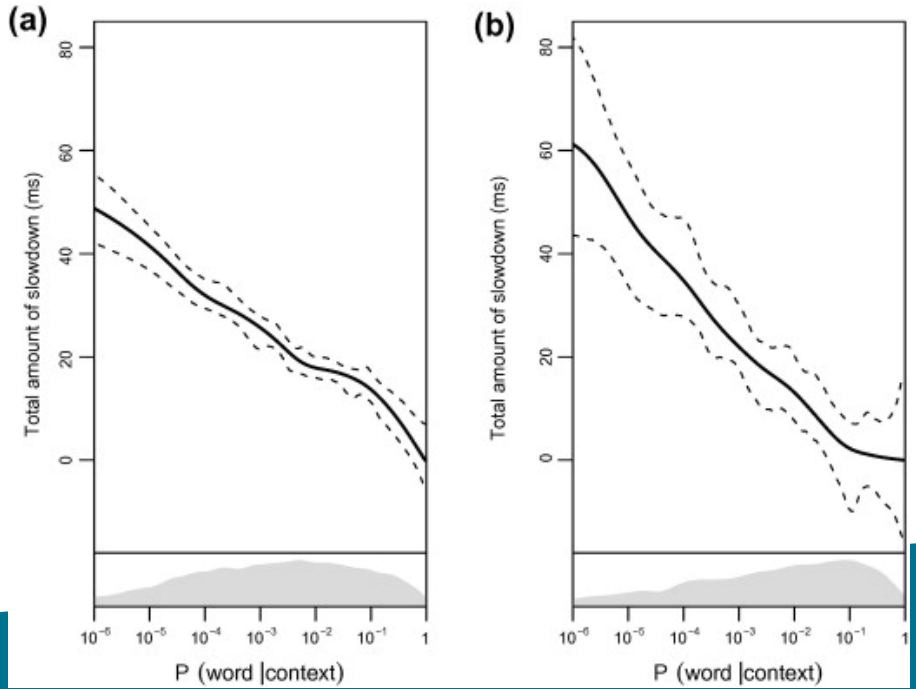


Reading times correlate with surprisal

(Smith and R. Levy 2013)

- Human language processing is predictive:
  1. My brother came inside to …
  2. The children went outside to…

  In (2) most people will predict *play* and will read it quicker than in (1) where the prediction is also less determined.
- Crucially **reading times are** known to be a function of word frequencies (Just and Carpenter 1982) and are linearly **correlated to surprisal** (Smith and R. Levy 2013).



Eye tracking times, first pass gaze duration, from the Dundee corpus (a) and self paced reading times read from the Brown corpus (b) as a function of $P(x_i|x_{<i})$

Physiological observations

Common instruments

- **EEG** (electroencephalogram): measures voltage differences from electrodes set on the head.
- **MEG** (magnetoencephalogram): measures changes in magnetic field from sensors around the head. EEG is a more noisy observation method than EEG: the skull distorts the signal.
- **fMRI** (functional magnetic resonance imaging): measures the BOLD signal (Blood Oxygen Level Difference) for every voxel as a function of time

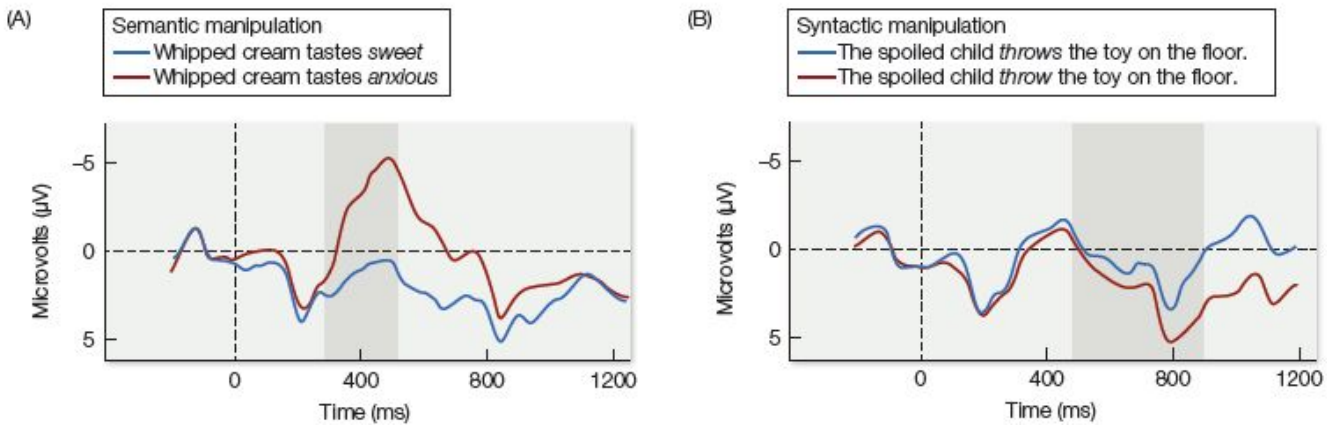
	Spatial	Temporal
EEG	-	+
MEG	+/-	+
fMRI	+	-

EEG data analysis

Event Related Potentials (ERP)

- EEG signal is essentially a time dependant signal for each of the n electrodes (e.g. $n = 64$). The signal is typically averaged over the relevant electrodes
- To ease data analysis and interpretation, a common practice is to align the signal with external interpretable **events** (such as the word onset/word is shown on a screen). Some intervals, relative to the event onset, are known to exhibit a peak in the signal under some interpretable circumstances, example:

ERP	interval(ms)	description	figure
N400	300-500	semantic surprisal	(A)
P600	500-900	syntactic surprisal	(B)



More examples

N400 (Kutas and Hillyard 1980)

- John buttered his bread with **socks** / with butter
- I take my coffee with cream and **dog** / milk.

One observes a negative peak around 400 miliseconds after the word event by contrast with the expected case

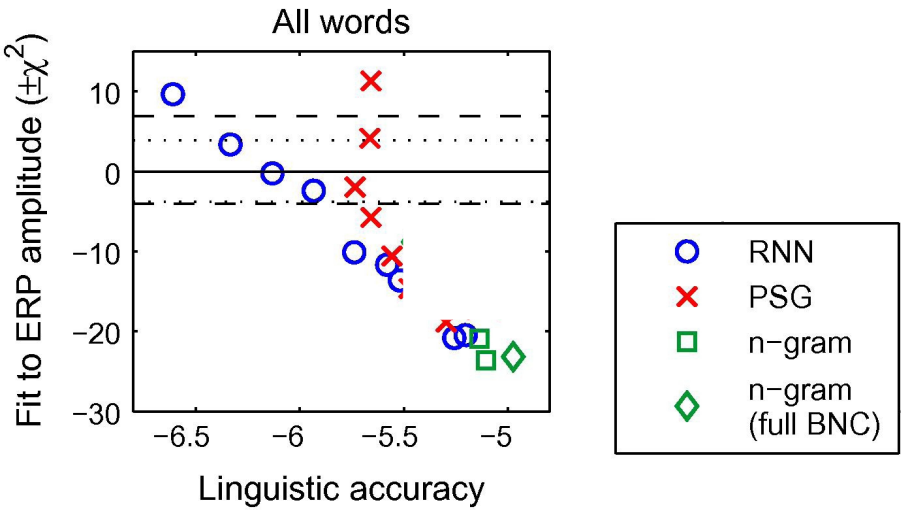
P600 (Osterhout and Olcomb 1992)

- The spoiled child **throw**/throws the toys on the floor.
- Charles doesn't shave because **him**/he tends to cut himself.

One observes a positive peak around 600 miliseconds after the word event by contrast with the expected case

Correlation between N400 amplitude and surprisal

- The ERP amplitude is defined as the average value of the signal on the ERP interval $y = \frac{1}{b-a} \int_a^b s(x)dx$. For the N400: $a = 300$ and $b = 500$
- Frank et al. (2015) identifies a correlation between surprisal and the amplitude of the N400. Two models are compared. Both predict ERP amplitude: one includes the surprisal (Linguistic accurracy) and control predictors (word frequency, word length, position in the sentence) and one omits surprisal.
- The resulting plot illustrates the difference in goodness of fit as function of surprisal.



Correlation between P600 and surprisal

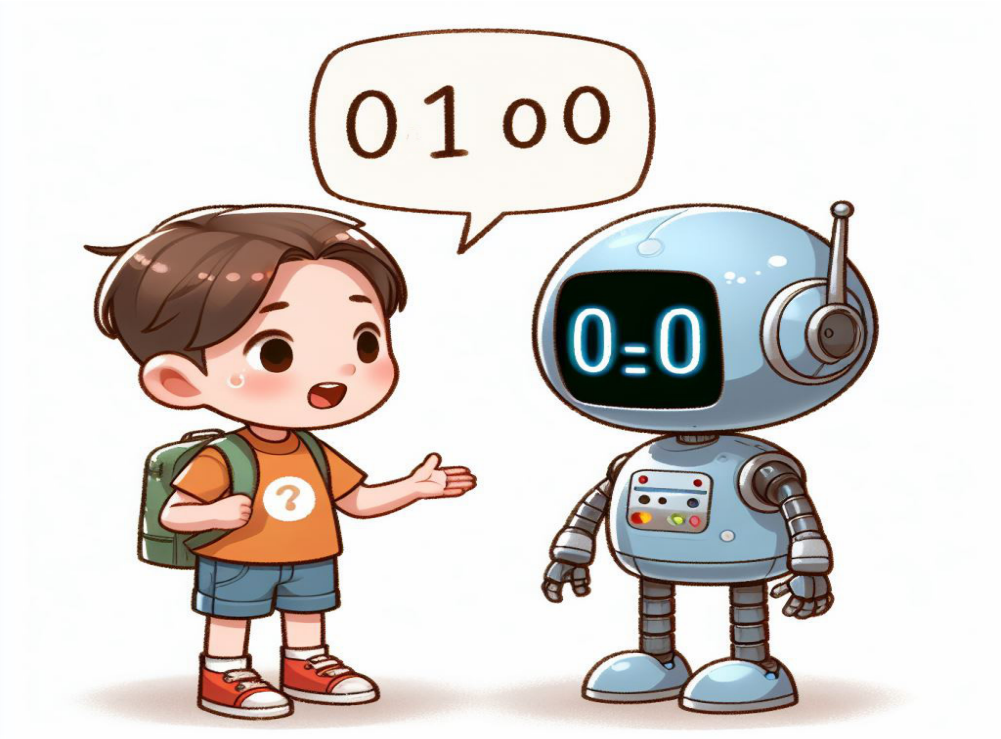
(Aurnhammer et al. 2021)

@see https://pubmed.ncbi.nlm.nih.gov/34582472/ for P600

@see Hale et al. on neurocomputational models of language

Outline

1. Entropy
2. The entropy of English
3. Relation to human processing
4. Memory limitations
5. Lossy memory models



LLMs are not limited like humans

- Language models have lower perplexity and lower average surprisal as the context size increase.
- Large Language Models have an almost unbounded context size and are supposedly close to optimal predictors
- Humans are context/memory limited so we expect them to use a ressource limited form of memory:

$$P(x_i|M) \approx P(x_i|x_1 \dots x_{i-1})$$

where M is an approximate (or lossy) representation of the context.

Main hypothesis : language has structure

Rather than memorizing the context perfectly, humans make sense of language by structuring it and by creating meaning compositionnaly. Thus memory limitations fundamentally shape human language

Human language cannot be reduced to a formal communication system

- A mathematical communication system features agents that are trying to communicate as **efficiently** and as faithfully as possible between each other : the encoding should be optimally short and lossless.
- In a communication system the code is not related to the meaning of the message, it is a function of the frequency of the form: the more frequent the form the shorter the code.
- In a communication system the codes are static: the codebook is designed before exchanging messages. In human languages there are dynamics taking advantage of the context (e.g. by using pronouns or definite expressions)

Human language is not a mathematical communication system

Contrary to an optimal code generated by a machine, human language is subject to biological or cognitive constraints: memory is limited and encoding is lossy.

Human language is shaped by memory limitations

- Humans have **limited memory capacity**: ”7 chunks” (Miller 51)
- Humans sometimes have difficulties to retrieve words already integrated because of **memory decay** or **interferences** (Lewis and Van Dyke xxx)
- There is some relation between form and meaning, human language has structure, it is compositional, systematic and productive: Human language is **compositional** (Fodor and Phylyshin 1988): we **structure** sentences and compute their meaning rather than store a photographic memory
  - **Productivity** (Chomsky 1956) We can create longer phrases by combining their parts
  - **Systematicity** If we can understand *John saw Mary* then we can also understand *Mary saw John* without being explicitly taught the form meaning mapping

Illustration (Recall task)

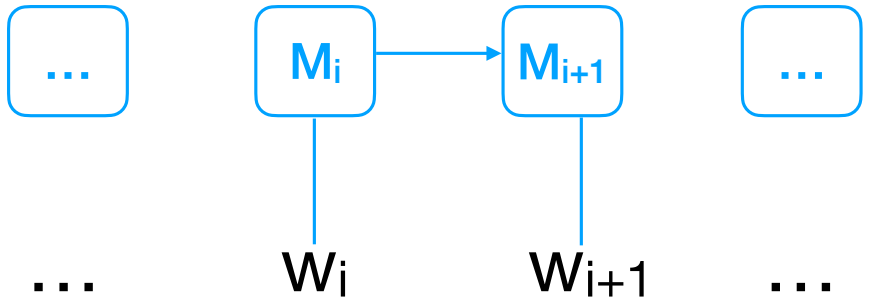
The number sequence recall task is harder as the sequence size increases while recalling a long natural sentence remains easy:

- 77 -3 9 17 -13 12 51 73 -87 10
- The cat sleeps on the mat and the rat is in the cellar

The memory processing model

(Futrell, Gibson, and R. P. Levy 2020)

- The memory processing model assumes **incremental** processing: language is a discrete **time dependant process** where at each time step the word w_i is **integrated** into the memory M.

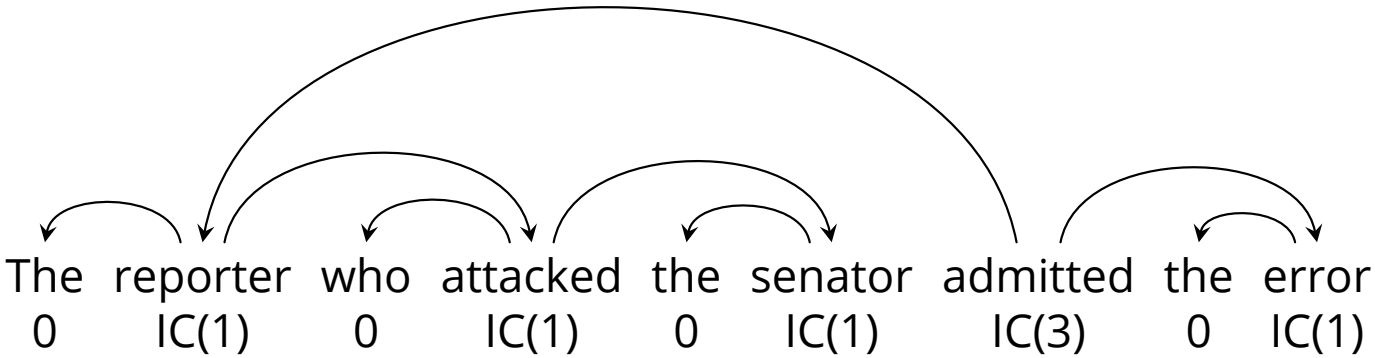


- The memory has two essential properties:
  - It is limited in capacity (Miller 1951)
  - The longer a **chunk** lives in memory the harder it is to retrieve it: either **memory decay** (Ebbinghaus 1913;Thorndike 1914) or **interference** (Van Dyke and Lewis 2003)

Dependency length theory (DLT)

(Gibson 1998)

- DLT models memory limitations and illustrates a memory based account of complexity. It assumes memory to be a **structured** object akin to a dependency tree built incrementally.
- To **integrate** a word in the memory, we link it to its dependants already in memory. The longer the dependencies the higher the integration cost. The integration cost is a function of the number of referential words the edges are spanning over.



- At some time *t*, the storage cost is the number of syntactic heads required to get a full parse tree.

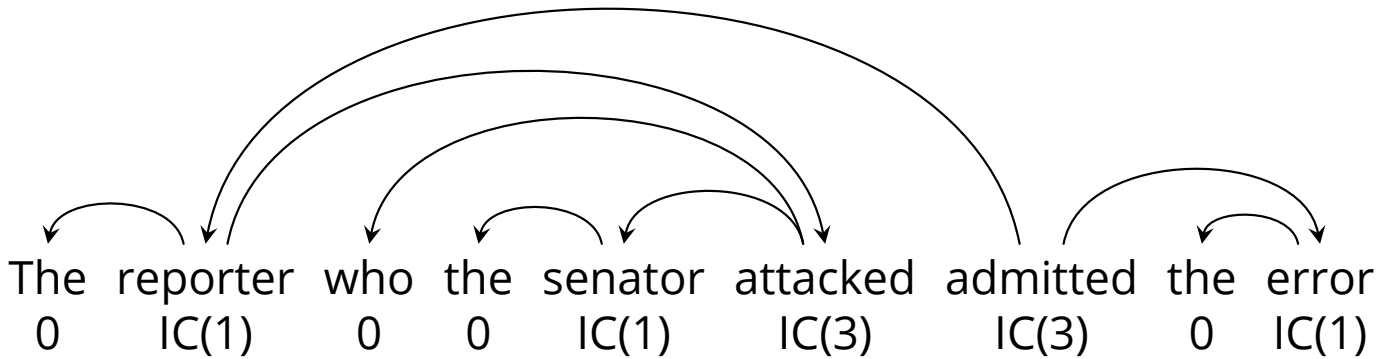
DLT and reading times

DLT predictions are meant to be compared to behavioral measures (reading times).
The tree structure is mapped to a time series by means of a **linking hypothesis**.

Example

In terms of reading times, DLT predicts that the head verb in the relative clause *attacked* is harder to read when the relative is an object relative.

1. The reporter who attacked the senator admitted the error
2. The reporter who the senator attacked admitted the error



Note that the DLT is annotation-scheme dependant (!)

Testing humans for sensitivity to syntactic structure

- Gibson (1998) observes that humans reading times increase as dependency length increase by contrasting subject and object relatives
  - The reporter who sent the photographer to the editor hoped for a story
  - The reporter who the photographer sent to the editor hoped for a story
- Reading times are increased when processing the object governor in the object relative by contrast with the subject relative
- Reading times increase on the main verb

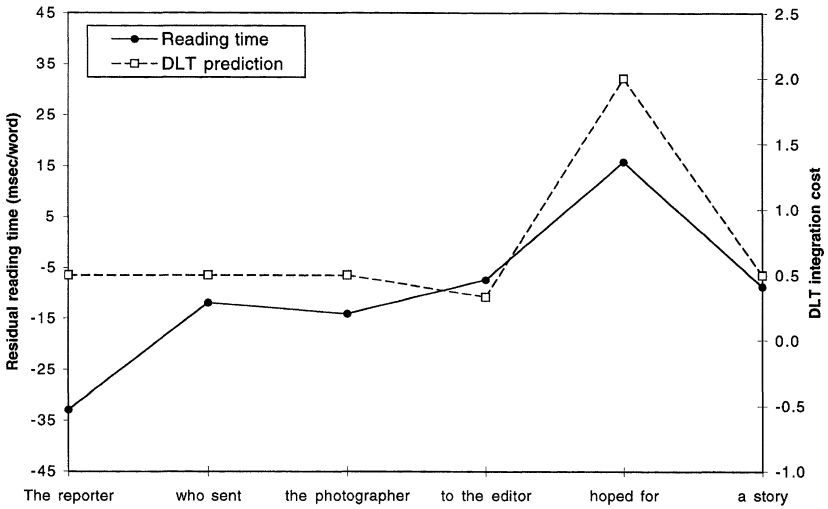


Figure 5.4
A comparison between residual reading times and locality-based integration costs in a subject-extracted RC structure.

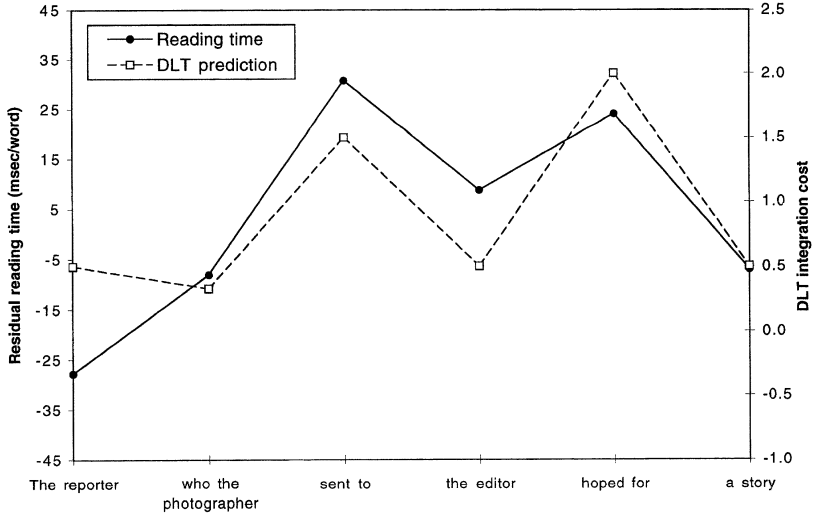


Figure 5.3
A comparison between residual reading times and locality-based integration costs in an object-extracted RC structure.

Further observations

Vasishth and Drenhaus (2011)

- **Dependency Length effects** Non relative clause specific observations are reported by Grodner and Gibson (2005).
  - The nurse supervised **the administrator** while ...
  - The nurse from the clinic supervised **the administrator** while ...
  - The nurse who was from the clinic supervised **the administrator** while ...

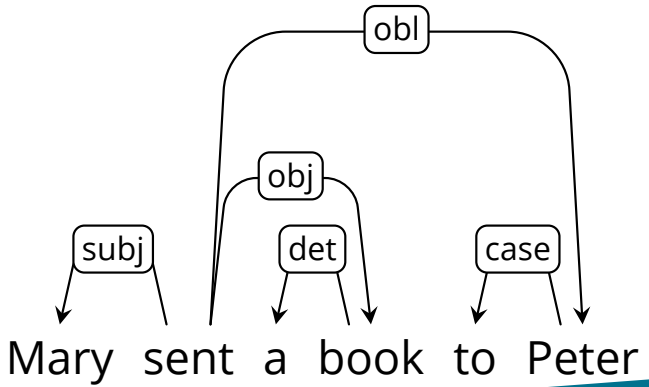
  Human reading times are expected to increase on the dependant verb.
- **Antilocality effects** are known for German where the final verb is read faster when the subject verb distance gets longer (Konieczny 2000).
- In English, Jaeger et al (08) also report such antilocality effects in:
  - The player [RC that the coach met at 8 o'clock] **bought** the house...
  - The player [RC that the coach met by the river at 8 o'clock] **bought** the house...
  - The player [RC that the coach met near the gym by the river at 8 o'clock] **bought** the house...

Dependency length minimization

- Experimental evidence remains fragile, yet corpus studies from typologically varied treebanks (Futrell et al. 2019) and further mathematical analysis (Temperley 2007) strengthens the hypothesis
- **Dependency Length Minimization** is the study of dependency lengths in **annotated corpora** without any **incremental** perspective and without any explicit connection to time series representations

Universal Dependencies

Statistics and counts come from the Universal Dependencies project, a multilingual treebank of dependency trees with shared annotation convention across languages



An old story...

The "laws" of Behagel (1930)

- **The highest law** "That which belongs together mentally is placed close together": covers cases such as adjectives modifying nouns are close to their head nouns, determiners are close to the noun...
- **The law of growing constituents** Of two sentence components, the shorter goes before longer when possible:

...and it reduces the global dependency length !

Dependency Length

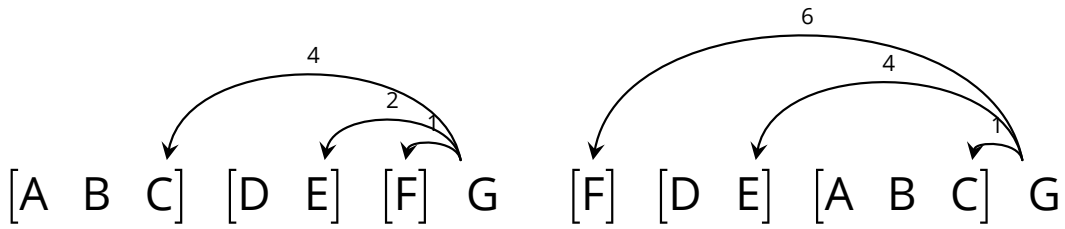
Let $i \to j$ be an annotated dependency, then $DL(i \to j) = |j - i|$ and the **dependency length of a tree** is $DL(\text{tree}) = \sum_{i \to j \in \text{tree}} DL(i \to j)$

Short before long or Long before short

(Hawkins 1994)

- **Short before long** Head initial languages (English, French…):

- **Long before short** Head final languages (Japanese, German…)



Proof (Temperley 2007)

DL(tree) is minimized if the word order follows one of the patterns:

Head initial	short before long
Head final	long before short

Consistency in head direction

(Temperley 2007; Gildea and Temperley 2010)

- Consistency in head direction is creates shorter dependency length than inconsistencies:

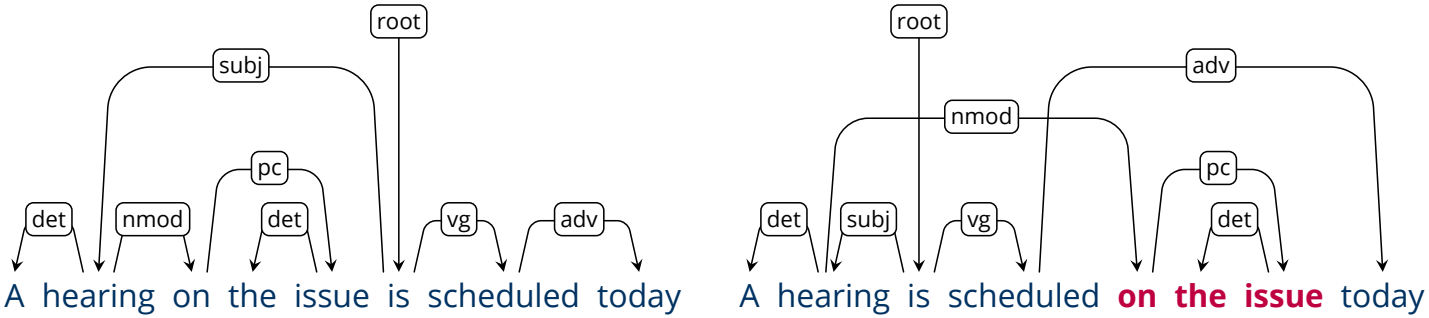
The left pattern is typical of a Verb Object Preposition Noun structure in English or French

- Except when we have multiple dependants

Greenberg (1963) universals

When a language is VO it has high chances to use prepositions, when a language is OV it has high chances to use postpositions

Projectivity and non projectivity



Left: a projective English sentence. Right: a non projective English sentence where a phrase "has moved".

- Projective sentence have shorter dependency lengths
- Projective sentences are massively more frequent in natural languages. Even in free word order languages such as German and Latin, non projective sentences are not that frequent
- It can be proven mathematically that minimizing dependency lengths by reordering the words of a sentence creates a more projective structure (Ferrer-i-Cancho 2016)

Assessing Dependency Length Minimization

(Futrell et al. 2015)

- The empirical assessment of dependency length minimization is made from Universal Dependencies and by comparing the actually annotated dependency trees with alternatives where words are reordered. More specifically the fact is established from comparisons between:

1. Reordered sentences where words are reordered in order to minimize DL(tree).
2. Annotated sentences
3. Reordered sentences with random permutations (randomness is controlled)

Annotated sentences mean dependency length is very close to that of (1)

A strong tendency

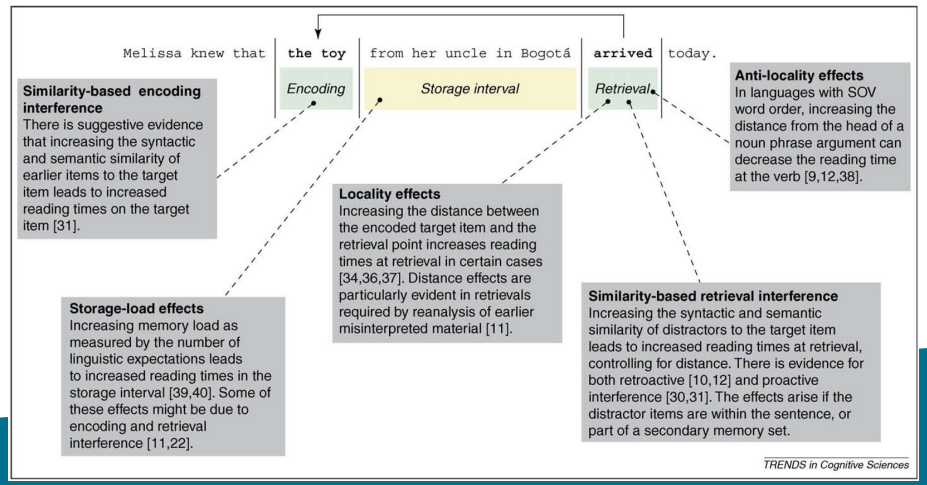
Dependency length minimization is not a ”law” of natural languages, there are plenty of cases where dependency length is not minimized. Rather Dependency length minimization should be viewed as a pressure, it is part of the general principles that contribute to explain word order

Summary

- There are two main lines/theories for explaining reading times measures:
  1. **Predict the future based theories**: this line assumes that human processing involves predicting the future given a context $P(x_i|C)$ and is grounded on information theory. The linking hypothesis with reading times rests on surprisal: $S(x_i|C) = -\log_2 P(x_i|C)$
  2. **Memory based theories**: this line assumes that human processing involves integrating words in the memory at each time step and that reading times are dependant of the memory structure (dependendency length)

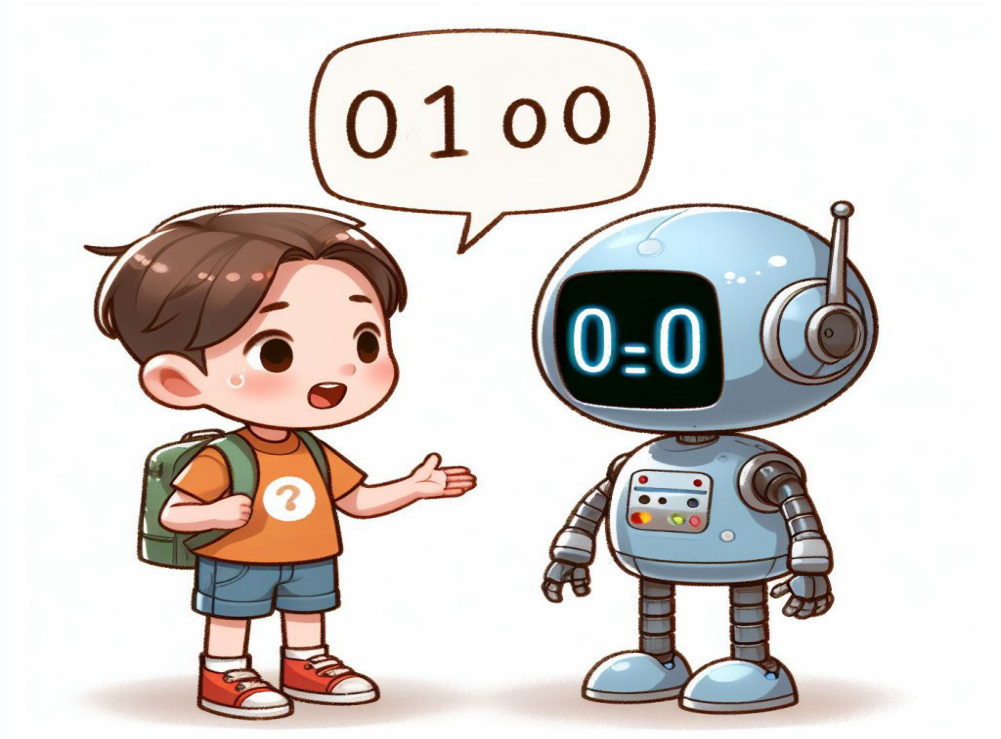
Main difficulty

Predictive effects and memory/structural effects are correlated and dependant of multiple factors (Lewis et al. 2006):



Outline

1. Entropy
2. The entropy of English
3. Relation to human processing
4. Memory limitations
5. Lossy memory models



Information-based model subject to memory constraints

- Rather than considering independently an information theoretic predictor and memory-based constrained predictor, the lossy memory model joins them together
- The memory is imperfect and chunks that are too old/too distant are dropped. Losing critical long distance information entails that $P(x_i|M)$ decreases by contrast with the perfect memory $P(x_i|C)$

Memory structure and lossy memory

Memory can be represented by different data structures and the lossy part can be represented by different means:

- The RNNs hidden vectors can play the role of a size constrained memory
- The GPT state can play the role of an incremental but non dynamic memory
- Classical n-grams models can play the role of a limited approximative memory too
- The state of an incremental (generative) parser (Stack, Buffer) can play the role of a memory
- Other graph-based methods are suggested by the literature

The case of the parser memory and ambiguity

- Let $\mathrm{x} = x_1 \ldots x_n$ be a sequence of tokens, then the probability $P(\mathrm{x}) = \prod_{i=1}^{n} P(x_i|M)$ is given directly by the model if the model in all cases except for the parser.
- In case the model is a generative parser, the probability is of the form $P(\mathrm{x}, d)$ where d is the specific parse derivation that generated the sentence. To get the probability $P(\mathrm{x})$ of a sequence we must marginalize over all derivations $P(\mathrm{x}) = \sum_{d \in D} P(\mathrm{x}, d)$ and the conditional takes the form:

$$P(x_i|M) = \frac{\sum_{d \in D} P(x_1 \ldots x_i, d)}{\sum_{d' \in D'} P(x_1 \ldots x_{i-1}, d')}$$

- Contrary to most language models, a generative parser explicitly encodes ambiguity in its different derivations at the price of enumerating the combinatorics. This may cause the interpretation of the memory harder than for language models.
- Yet (**levy-2008**) provide an information theoretic interpretation of surprisal in terms of derivation reranking in the parser.

A resource-rational model of human processing of recursive linguistic structure

Hahn, Futrell, Levy, PNAS, 2022

- Idée: GPT-2 est à peu près un ”communicant optimal”
- On peut lui enlever de manière controlée et théoriquement motivée (information mutuelle) des élements de la mémoire
- Tests sur des temps de lecture

Varia: LLMs/Memory and generalisation

- How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN (McCoy et al. 2023)
- ”Energy Transformers” (with V. Segonne). We explore the hypothesis that LLMs do **not** generalize rather they only retrieve from memory. (Idea from Hopfield networks)