A promise of language models for language science ? let’s chat

Large Language models for linguistics ?

Benoît Crabbé

21st March 2023

Laboratoire de Linguistique Formelle
Université Paris Cité

Claire’s Bio on the website

Before becoming a scientist, Claire trained race horses in Ireland, trapped possums in New Zealand and herded sheep on the German moors. In Geneva where she later went to become an interpretor, she met Maghi King who finally sent her to the UK to study for a Master in Intelligence Knowledge Based Systems at Essex University. She did her Master research under Ann de Roeck’s supervision and moved on to Edinburgh to do her PhD. There, she was hired by Gabriel Bès (U. de Clermont Ferrand, France) to work on the EU funded ACORD project on parsing and generating with Unification Categorial Grammar. In her thesis on the other hand she investigated the syntactic, semantic and computational properties of verbal ellipsis and was supervised by Ewan Klein and Robin Cooper.

After leaving Edinburgh, she worked in Amsterdam and Utrecht on EU funded projects and in Saarbrücken on German funded ones (LISA, InDiGen). Finally, in 2000, Jean-Marie Pierrel suggested her to apply for a CNRS position and since then she have been in Nancy working at LORIA. Along the way, she gave birth to three daughters: Jennifer, Gabrielle and Caroline. She taught Iyengar yoga at LORIA for 10 years. For this, she thank Clé Souren, her teacher in Amsterdam.

Chat GPT in action: programming, world knowledge

Please write a short bio for Claire Gardent and then write a program that will match her name in this text with a regular expression

Claire Gardent is a computer scientist and researcher who specializes in natural language processing (NLP) and computational linguistics. She is a research director at CNRS (the French National Center for Scientific Research) and a professor of computer science at Lorraine University in France. Her research focuses on developing computational models for natural language generation, semantic parsing, and discourse analysis.

Here is a Python program that matches the string ”Claire Gardent” in the text above using a regular expression:

```
import re

text = "Claire Gardent is a computer scientist and researcher [...]"

match = re.search(r"Claire Gardent", text)

if match:
            print("Match found!")
else:
            print("Match not found.")
```

Linguistic knowledge

Give me the rule of the agreement of the past participle with the direct object in French

In French, the past participle of a verb agrees in gender and number with the direct object of the verb if the direct object comes before the verb in the sentence. This is known as ”accord du participe passé avec le complément d’objet direct (COD)” or ”concordance du participe passé avec le COD.”

Here are the rules for the agreement of the past participle with the direct object in French:

1. *If the direct object is feminine singular, the past participle agrees with it in gender and adds an ”e” at the end of the participle. Example: J’ai mangé la pomme rouge. (I ate the red apple.) -* ”pomme” is feminine singular, so ”rouge” and ”mangé” agree with it and end in ”e”.
2. *If the direct object is masculine plural or feminine plural, the past participle agrees with it in gender and adds an ”s” at the end of the participle. Example: Nous avons mangé les pommes rouges. (We ate the red apples.)* - ”pommes” is feminine plural, so ”rouges” and ”mangé” agree with it and end in ”s”.
3. *If the direct object is masculine singular, the past participle does not change. Example: J’ai mangé le pain. (I ate the bread.) - ”pain” is masculine singular, so ”mangé” does not change.*
4. *If the direct object is a pronoun, the past participle agrees in gender and number with the pronoun. Example: Je l’ai mangée. (I ate it.) -* ”la” is feminine singular*, so ”mangée” agrees with it and ends in ”e”.*

It’s important to note that the agreement of the past participle with the direct object only applies when the direct object comes before the verb. If the direct object comes after the verb or is not present, there is no agreement.

Plan

1. How did we get there ?
2. Language models, Sesame street and GPT
3. Fine tuning language models
4. Language models and language structure

How did we get there ?

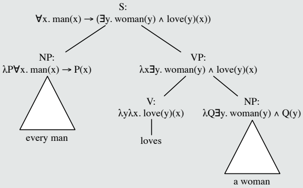
1956 The birth of AI

1956: Dartmouth summer school (John Mc Carthy, Marvin Minsky, Allan Newell, Claude Shannon, Herbert Simon...)

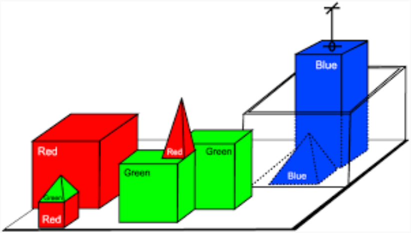
1956: Three models for the description of language (Chomsky)

Symbolic AI

That's the golden age of symbolic AI that begins. In NLP, beyond Chomsky's work on formalizing language **structure**, works from R. Montague allow to build bridges with the semantics of formal languages



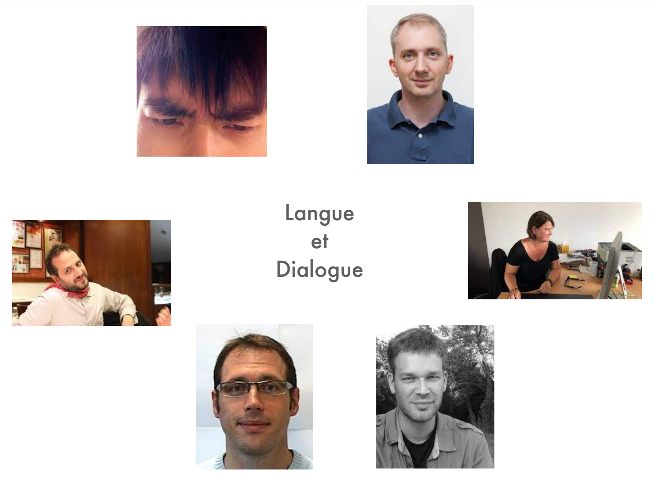
A key milestone...



```
HUMAN: Pick up a big red block.
SHRDLU: OK.
HUMAN: Grasp the pyramid.
SHRDLU: I don't understand which pyramid you mean
HUMAN: Find a block which is taller than the one you are holding
       and put it into the box.
SHRDLU: by "it", i assume you mean the block which is taller than the one
          I am holding.
SHRDLU: ok.
HUMAN: What does the box contain?
SHRDLU: The blue pyramid and the blue block.
```

(SHRDLU, Winograd 1972)

When we met Claire !



The problem

- SHRDLU is one of the first chatbots operating on a **closed world**. With few ambiguities and such a small world, it is possible to formalize a language to talk with the bot and the physics and implicits of the world.

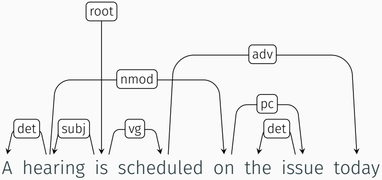
The problem

SHRDLU does not generalize

- Modeling language in an open world requires modeling a larger grammar but also world knowledge, common sense, wider discourse context...and it becomes pretty hard to model with symbolic rules only.

Statistical models

- From the linguist point of view, there is the key hypothesis that language has **structure**. For instance:



- **From the beginning of the 1990**, to address ambiguity problems, and with the augmentation of storage and computing power, the first machine learning models for natural language processing start to emerge.

The problem

Generally speaking, annotated data for learning structured models of language is very costly/time consuming to get. Annoted corpora are generally small and not representative (natural language is far from being iid)

Summary

Two main problems in modeling language:

1. Modeling the structure of language
2. Modeling the rest : world knowledge, context, common sense ...

Language models, Sesame street and GPT

Markovian language models

- Markovian language models do not need costly annotations and can be trained on very large amounts of raw text:

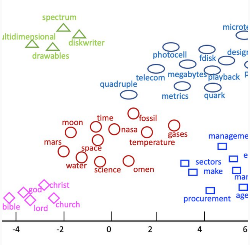
$$P(\mathbf{w}) \approx \prod_{i=1}^{n} P(w_i | w_{i-k} \ldots w_{i-1})$$

- Historically, these models were related to some emblematic tasks such as machine translation and automatic speech recognition
- Probability estimation from corpora, **by counting**, becomes a tricky art as soon as the order of the model (size *k* of the context) increases.

Neural language models (word embeddings)

The breakthrough takes place in two times:

- **Distributional hypothesis:** words appearing in similar contexts have similar meanings
- **(Mikolov 2013, static word embeddings)** Learn a mapping from words to dense vectors in such a way that words occuring in similar contexts have similar embeddings



- **Word embeddings =** static dictionary (strings→ vectors)
- What about lexical ambiguities in context ?

  ex. *avocat:fr* (lawyer,advocado) ? *est:fr* (east, is) ?

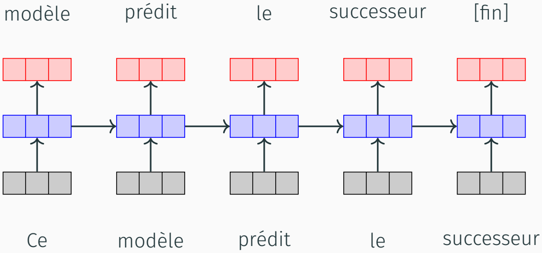
Recurrent neural language models

(2011, Mikolov et al.)

- Markovian hypotheses are relaxed:

$$P(\mathbf{w}) \approx \prod_{i=1}^{n} P(w_i | w_1 \ldots w_{i-1})$$

- The conditional probabilities are now dependant of **contextualised embeddings**



The story so far...

Historically, the estimation of conditionals $P(w_i|context)$ is made by counting n-grams:

- Example :

$$P(cat|the, black) = \frac{\#(the, black, cat)}{\#(the, black)}$$

Zipf and Context Sharing

- n-grams are Zipf distributed ⇒ most n-grams occur only once (or zero) even in very large corpora.
- Probability estimation naturally suffers from a sparseness problem
- Using embeddings for representing contexts allows to share information between similar contexts : *the black, cat* ; *the white cat*; *the gray cat* ...

Current state of the art

Ernie, Elmo and Bert, characters from Sesame Street

OpenAI (GPT-3, chatGPT, GPT-4)

Properties

- Contextualized language models
- Trained on very large data sets
- Designed to be efficiently run on GPU (parallelism)
- First outstanding success of unsupervised (self-supervised) methods in natural language processing
- **not a single bit of linguistics in the design**

Two games for learning a language model

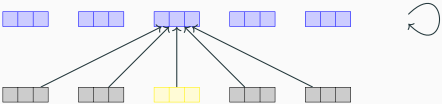
- Fill in the gap (BERT) :
  The latest breakthrough in ________ intelligence
- Guess the next word (Elmo, GPT-3,GPT-4):
  The latest breakthrough in artificial ________

Optimisation objective

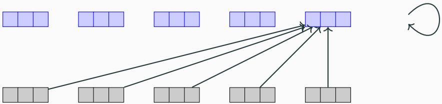
This child's ”game” is framed as an optimisation problem that aims to maximise the probability of the word to guess given the context

Attention inside !

- Masked language models (BERT)



- Autoregressive models (GPT)



Attention in a nutshell

Contextualised embeddings are a weighted sum of the sentence embeddings. The weights between positions *i* and *j* are a function of the similarity between input vectors $\mathbf{x}_i$ and $\mathbf{x}_j$

Learning representations

- The main interest in current language models is to take advantage of their **representations** for words, sentences, text...rather than actually paying much attention to their capacity of predicting the missing word.
  - A language model is naturally used as a **function mapping strings to contextualized vectors**
  - By contrast with static embeddings, the contextualisation performs some sort of word sense disambiguation.
- Training them on huge amounts of data is likely to help them to model context and acquire correlates of world knowledge and common sense without having to create explicit symbolic data bases.
  - But this likely remains a somewhat approximate and incomplete knowledge

Language models and downstream tasks

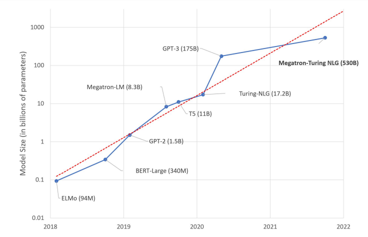
- Large language models are mostly used to map linguistic strings to vectors.
- They are used in downstream applications, for instance in question answering (SQUAD):

  > *Apollo ran from 1961 to 1972, and was supported by the two-man Gemini program which ran concurrently with it from 1962 to 1966. Gemini missions developed some of the space travel techniques that were necessary for the success of the Apollo missions. Apollo used Saturn family rockets as launch vehicles. Apollo/Saturn vehicles were also used for an Apollo Applications Program, which consisted of Skylab, a space station that supported three manned missions in 1973–74, and the Apollo–Soyuz Test Project, a joint Earth orbit mission with the Soviet Union in 1975.*
  >
  > Q: What space station supported three manned missions in 1973–1974 ?
  >
  > A: Skylab

- Such a system builds a representation (e.g. a vector, a list of vectors...) for the text, a representation for the question (e.g. a vector) that are used as input for predicting the answer.
- Most of the tasks in NLP may be framed with a similar setup (sentiment analysis, summarization, machine translation...)

Scaling up

- The time-dependant architecture of recurrent models (GRU, LSTM) is impractical for scaling up
- Current language models are architecturally state-less (GPT simulates temporal states)
- Instead they are designed to run in parallel without time dependant states
- GPUs can be used efficiently and models can scale up



Model	Training size (Gb)
BERT	16
XLNET	126
Megatron	174
T5	745
TuringNLG	174
GPT3	570

A word on chat GPT

- Although unpublished, Chat GPT is essentially GPT-3 with additional **reinforcement learning with human feedback**. This likely to correct weaknesses of GPT-3 (Bang et al. 2023):
  - Hallucinations
  - Deficient at reasoning (mathematical, multi-hop) but has acquired some commonsense reasoning
- The users of chat GPT may indicate whether the generated text is good or bad
- Given its success and the number of users, openAI is likely to collect an unprecedented amount of human annotations in the history of the field.

Fine tuning language models

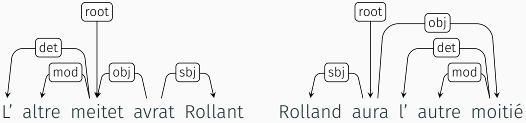
The evolution of French syntax (Grobol et al. 2022)

- Modern French is an SVO language
- Old French has relatively free word order with dominance SOV, SVO
- Latin is essentially an SOV language

Exemple

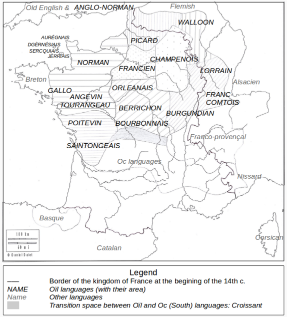
1. L'altre meitet avrat Rollant (OF : OVS)
2. Roland aura l'autre moitié (MF : SVO)

- The **question** : what is the evolution mechanism that explains the evolution from Old French to Modern French ?
- The **method** requires to make the syntax explicit: it requires to annotate corpora across time



Difficulties for annoting Old French

- Pretty costly to annotate text in syntax, and particularly for Old French where you can hardly find native speakers
- Training a parser on Old French is hard: very limited annotated data sets (in size)
- The language is non normalized ⇒ Variation
  - Genre (verse or prose)
  - Geographic



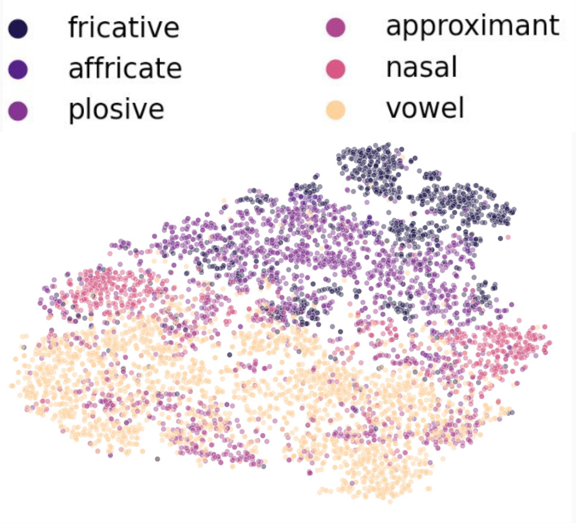
Method (Grobol et al. 2022)

- Large Language models for modern French do exist (CamemBERT, FlauBERT)
- Fine-tuning them on 55Mb of Old French raw corpora gets you BERTrade
- We can learn a very accurate parser on Old French using the BERTrade language models
- Manual annotation/verification is substantially reduced

Base model	UPOS	UAS	LAS
Vanilla	93.51	87.60	81.54
BERTrade-mBERT	96.95	93.33	89.60
BERTrade-CamemBERT	97.16	93.75	90.06
BERTrade-FlauBERT	96.94	93.75	90.07

Discoveries in typology ?

- Speech language models (e.g. `wav2vec`) have representations that predict phonological features:



Deep Typo (Wisniewski et al.)

- There are about 5000 languages spoken in the world and current models cover a few dozens of them...
- **Question** Can we design methods to transfer annotations from well resourced languages to low resourced languages ?
- Can we take advantage of speech models to:
    - Help annotation (phonological)
    - Suggest new hypotheses in typology by computing distances accross languages

Pangloss (Lacito)

The Pangloss collection gathers speech recordings from rare languages. These documents result from careful data collection made by linguists that aim to preserve cultural heritage

Language models and language structure

What does Chat GPT know about language structure ?

Basic Anaphora

- **What is the antecedent of she in the sentence ”Claire has been awarded the CNRS silver medal and is now a fellow of the ACL”**

  The antecedent of ”she” in the sentence ”Claire has been awarded the CNRS silver medal and is now a fellow of the ACL” is ”Claire”.

- **What is the antecedent of she in the sentence ”Claire has been awarded the CNRS silver medal and she is now a fellow of the ACL”**

  The antecedent of ”she” in the sentence ”Claire has been awarded the CNRS silver medal and is now a fellow of the ACL” is ”Claire”.

What does Chat GPT know about language structure ?

Basic Syntax

- **What is the object in the sentence ”Claire has been awarded the CNRS silver medal and is now a fellow of the ACL”**

  The object in the sentence ”Claire has been awarded the CNRS silver medal and is now a fellow of the ACL” is ”the CNRS silver medal”.
- **What is the object in the sentence ”Claire is now a fellow of the ACL”**

  The object in the sentence ”Claire is now a fellow of the ACL” is not clear because there is no direct object in this sentence. Instead, ”a fellow of the ACL” is a predicate nominative that renames the subject ”Claire” and describes her status or position.

Language models and cognition

- Large Language models are not only good for providing representations/text generation, they are also good predictors of human language behavior and human physiology.
- Example stimuli (written/speech):

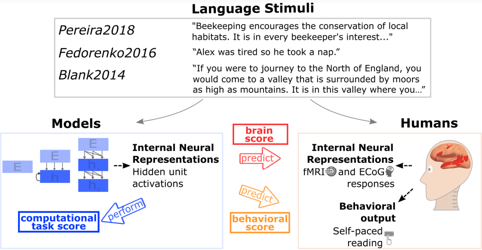
  > *If you were to journey to the North of England you would come to a valley that is surrounded by moors as high as mountains. It is in this valley...*

- Measures
  - from behavioral tasks: self paced reading times, eye tracking
  - Measures from physiological tasks: EEG, MEG, fMRI...
- But also model predictions!

Observation (Schrimpf et al. 2022)

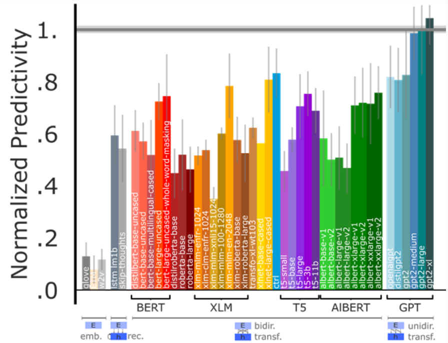
Large Language Models run on the same stimuli are good predictors of human behavior/physiolology

Large language models and cognition (Schrimpf et al 2021)



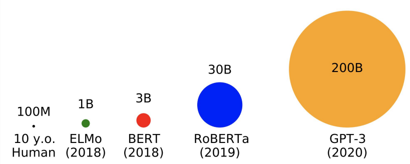
Language Stimuli

Pereira2018	"Beekeeping encourages the conservation of local habitats. It is in every beekeeper's interest..."
Fedorenko2016	"Alex was tired so he took a nap."
Blank2014	"If you were to journey to the North of England, you would come to a valley that is surrounded by moors as high as mountains. It is in this valley where you..."



Language models and language structure

- Large Language Models (transformers) are not motivated by linguistic theory (biases are likely to be very different)
- Contrary to linguistic theory, no distinction between language and common sense, world knowledge (Mahowald et al. 2023)
    - Results from neurosciences tend to highlight there is a language-specific network in the brain (Blank et al. 2016, Fedorenko et al. 2018)
- Require an unrealistic amount of training data:



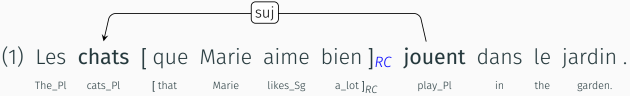
- But they have impressive empirical adequacy and (apparently) relevant internal representations
- New/alternative insights for language sciences ?

Structural properties (ongoing work)

Modern language models have rich internal structure:

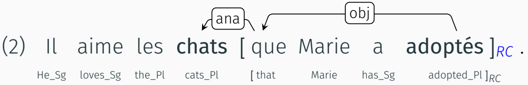
- To which extent do they capture anything similar to the linguistic structure of language ?
- To which extent do they generalize rather than memorize ?
- To which extent can they be modified/biased to reflect linguistic hypotheses

Two cases of agreement in French (Li, Wisniewski, Crabbé 2023)



(1) Les **chats** [que Marie aime bien]$_{RC}$ **jouent** dans le jardin .

The_Pl cats_Pl [that Marie likes_Sg a_lot]$_{RC}$ play_Pl in the garden.



(2) Il aime les **chats** [que Marie a **adoptés**]$_{RC}$.

He_Sg loves_Sg the_Pl cats_Pl [that Marie has_Sg adopted_Pl]$_{RC}$

Why this example ?

- In surface, both cases look similar: a noun precedes a verbal form and agreement should hold between the two
- The linguistic structure is however quite different: in the first case it requires to identify an embedded relative clause; in the second case there is an anaphoric clause and a ”moved” object

The model used...

- Small scale autoregressive transformer model/LSTM
- Trained on French Wikipedia ($\approx$ 9Gb)
- Tokens are traditional words rather than tokens ($V = 50000$ tokens)
- Tested on French Gutenberg extracted examples ($\approx$ 95000 examples overall)

Structure encoding

Question : Do language models capture a non superficial structure ?

Test task : We ask to an **incremental** (autoregressive) model to predict the verbal form (binary choice, two potential forms)

Les chats que Marie aime bien ______ {joue, jouent}

Test examples are ranked as a function of the number *h* of heuristics they satisfy:

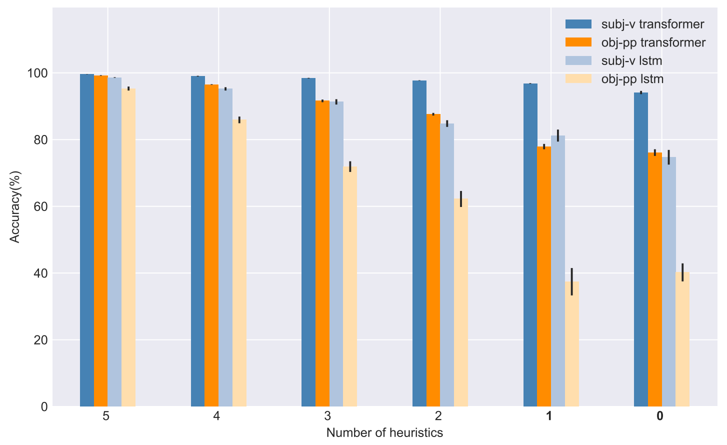
- agreement with the first noun in the sentence gives the correct prediction
- agreement with the noun closest to the verb gives the closest prediction
- agreement with the token closest to the verb gives the closest prediction
- agreement with the noun before the closest *que*
- agreement with the majority number in the sentence gives the closest prediction

The higher the *h* the easier is the example

Rating examples difficulty with heuristics

Number of heuristics	Difficulty of agreement	Examples
5	- - -	Si les **idées** que ces mots représentent ne **sont** pas ... (5) *If the ideas*$^{(1)}_{(4)}$ *that these words*$_{(2)}$ *represent*$_{(3)}$ *are not...*
4	- -	Les **choses** que nous avions vues cent fois avec indifférence nous **touchent**... (5) *The things*$^{(1)}_{(4)}$ *that we had seen a hundred times with indifference us*$_{(3)}$ *touch*
3	-	Un philosophe est curieux de savoir si les **idées** qu' il a semées **auront**... *A philosopher is curious to know if the ideas*$^{(4)}_{(2)}$ *that he has sown*$_{(3)}$ *have...*
2	+	Les **emblèmes** qu' on y rencontre à chaque pas **disent** ... *The emblems*$^{(1)}_{(4)}$ *that we meet at each step say ...*
1	+ +	Les **qualités** qui t'ont fait arriver si jeune au grade que tu as **doivent** te porter ... *The qualities*$_{(1)}$ *that made you arrive so young at the rank you have must bring you ...*
0	+ + +	Ce soir les **hommes** que j'ai postés sur la route que doit suivre le roi **prendront** ... *Tonight the men that I have posted on the road that the king must follow will_take ...*

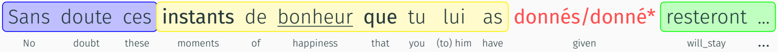
Results



In non trivial cases ($h = 0$) incremental transformers are still above chance

Where is encoded the number

- **Protocol :** We use a probe to predict the correct agreement number (antecedent,verb) on all the tokens
- **Observation :**



Sans	doute	ces	**instants**	de	bonheur	**que**	tu	lui	as	données/donné*	resteront	…
No	doubt	these	moments	of	happiness	that	you	(to) him	have	given	will_stay	…

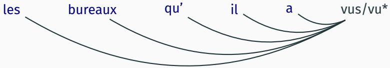
	Mean probing Accuracy	
	Object-pp	Subject-verb
prefix	$58.6\%_{\pm0.1}$	$59.5\%_{\pm0.2}$
context	$92.3\%_{\pm0.2}$	$93.0\%_{\pm0.1}$
suffix	$73.6\%_{\pm0.2}$	$78.1\%_{\pm0.2}$

- The number is essentially encoded within the context

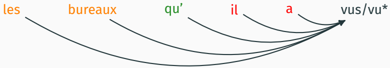
Differences between the two cases of agreement ?

Protocol:

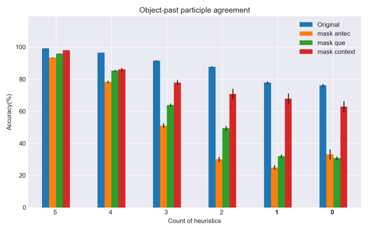
- An incremental transformer predicts the verbal form verbale with full access to the previous context

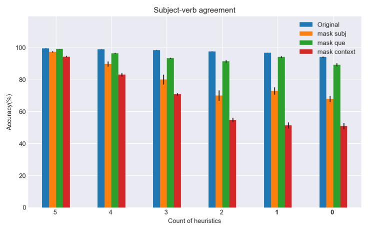


- Here we prevent the model to have access to some key elements of the context (ex ‘que’):



Results


Object-past participle agreement


Subject-verb agreement

- ‘que’ is important for achieving object past participle agreement, but much less pour for subject-verb agreement
- For object-past participle agreement, decisions rely on antecedent and que
- For subject verb agreement, decisions rely on subject and local context

⇒ predictions rely on different mechanics, one for each agreement type

Memorization or Generalization ? (COGS, Kim and Linzen 2020)

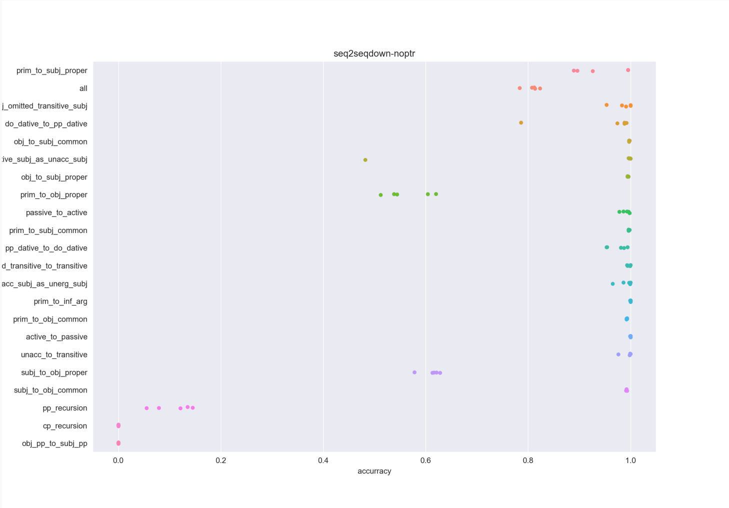
- **Task:** learn on a dataset with some missing **structural** generalization patterns
- Predict on a dataset where the model has to generalize for **substitutivity** and **productivity** patterns (Fodor, Phylyshin 1988)

Training	Generalization
A hedgehog ate the cake $*\mathit{cake}(x_4); \mathit{hedgehog}(x_1)$ $\wedge \texttt{eat.agent}(x_2, x_1) \wedge \texttt{eat.theme}(x_2, x_4)$	The baby liked the hedgehog $*\mathit{baby}(x_1); \mathit{hedgehog}(x_4)$ $\wedge \texttt{like.agent}(x_2, x_1) \wedge \texttt{like.theme}(x_2, x_4)$
Ava saw a ball in a bowl on the table $*\mathit{table}(x_9); \texttt{ball}(x_3) \wedge \texttt{bowl}(x_6)$ $\texttt{see.agent}(x_1, \mathit{ava}) \wedge \texttt{see.theme}(x_1, x_3)$ $\texttt{ball.nmod.in}(x_3, x_6) \wedge \texttt{ball.nmod.on}(x_6, x_9)$	Ava saw a ball in a bowl on the table on the floor $*\mathit{table}(x_9); *\texttt{floor}(x_{12}); \texttt{ball}(x_3) \wedge \texttt{bowl}(x_6)$ $\texttt{see.agent}(x_1, \mathit{ava}) \wedge \texttt{see.theme}(x_1, x_3)$ $\texttt{ball.nmod.in}(x_3, x_6) \wedge \texttt{ball.nmod.on}(x_6, x_9)$ $\texttt{table.nmod.on}(x_9, x_{12})$

Observation

For standard seq2seq, total failure on productivity patterns. More or less accurate on substitutivity patterns

Example (seq2seq)



Models that generalize on productivity patterns are neural parsers.

Conclusion and perspectives

- New insights from existing models ? (Baroni 2020)
- Models with built-in structural, linguistic or cognitive biases:
  - (Vasishth and Engelmann 2022) Sentence Comprehension as a Cognitive Process
  - Graph Neural networks (Scarselli 2009)
  - Time dependant (recurrent) models (Futrell et al 2020)
- New perspectives in linguistic typology
- … can we expect an AlphaFold for language sciences ? the question remains open

Last but not least

thank you !