STAGE III

Minimal(ist) derivation, memory & intervention (digesting the linguistic cake)

Minimalist Grammars

- Stabler's (1997) formalization of a Minimalist Grammar, **MG** (Chomsky 1995) as a 4-tuple (V, Cat, Lex, F) such that:

  V is a finite set of non-syntactic features, (P ∪ I) where

  P are phonetic features and I are semantic ones;

  Cat is a finite set of syntactic features,

  Cat = (base ∪ select ∪ licensors ∪ licensees) where

  base are standard categories {comp, tense, verb, noun ...},

  select specify a selection requirement {=x | x base}

  licensees force phrasal movement {–wh, –case ...},

  licensors satisfy licensee requirements {+wh, +case ...}

  Lex is a finite set of expressions built from V and Cat (the lexicon);

  F is a set of two partial functions from tuples of expressions to expressions : {merge, move};

Minimalist Grammars

V = P = {/what/, /did/, /you/, /see/},
I = {[what], [did], [you], [see]}

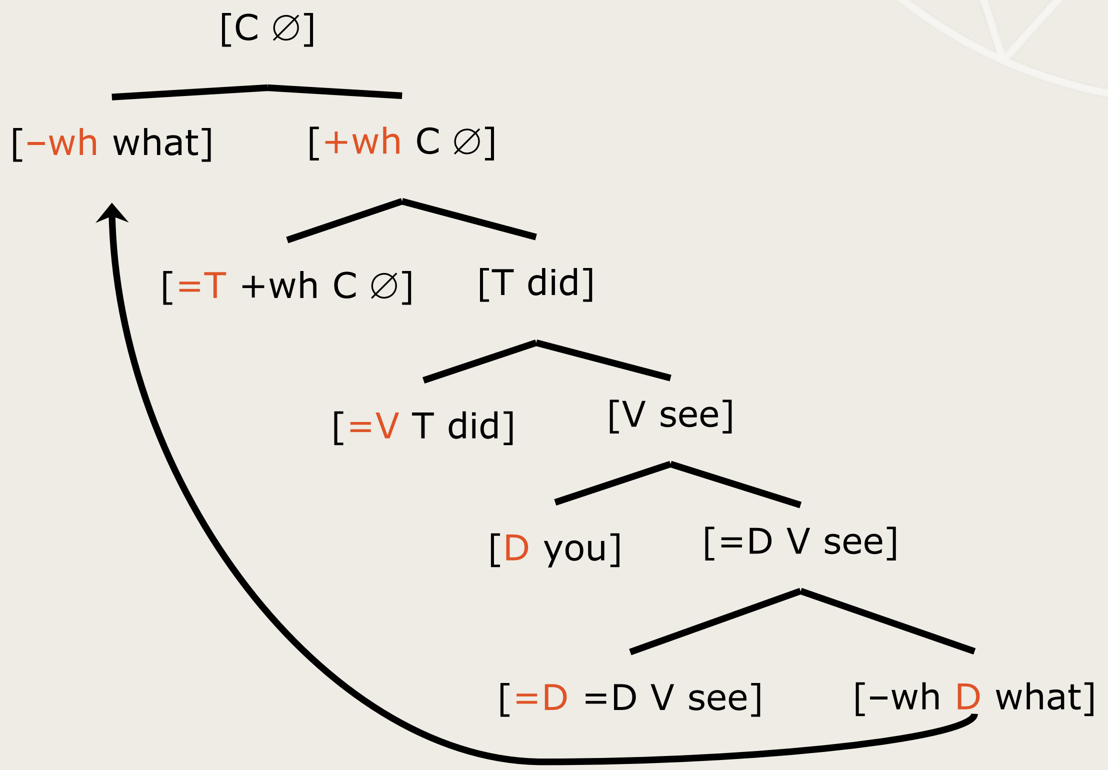
Cat = *base* = {D, N, V, T, C}
select = {=D, =N, =V, =T, =C}
licensors = {+wh}
licensees = {–wh}

Lex = { [–wh D what], [=V T did], [D you], [=D =D V see],
[=T +wh C ∅] }

F = {*merge*, *move*} such that:
merge ([$=F$ X] , [F Y]) = [$_X$ X Y]
("simple merge" on the right, "complex merge" on the left)
move ([$+g$ X] , [W [$-g$ Y]]) = [[$_X$ Y X] W, t_Y]

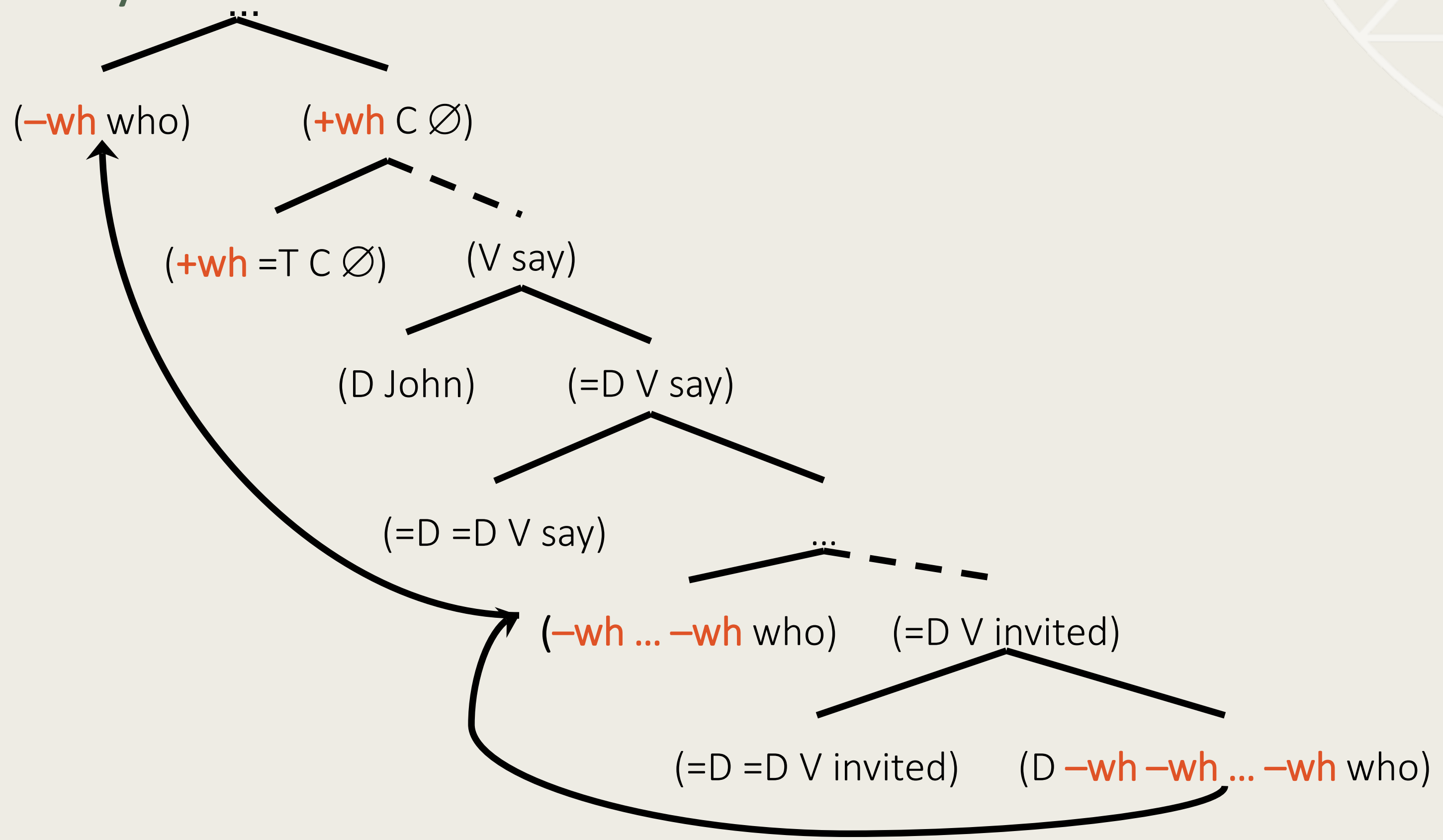
Minimalist Grammars

1. merge ([=D =D V see], [-wh D what]) → [$_{\text{see}}$ =D V see, –wh what]
2. merge ([D you], [=D V see, -wh what]) → [$_{\text{see}}$ you, [$_{\text{see}}$ V see, –wh what]]
3. merge ([=V T did], [$_{\text{see}}$ you, [$_{\text{see}}$ V see, -wh what]]) →
   ([$_{\text{did}}$ T did, [$_{\text{see}}$ you, [$_{\text{see}}$ see, –wh what]]]
4. merge ([=T +wh C ∅], [$_{\text{did}}$ T did, [$_{\text{see}}$ you, [$_{\text{see}}$ see, –wh what]]]) →
   ([$_{\text{C}}$ +wh C ∅, [$_{\text{did}}$ did, [$_{\text{see}}$ you, [$_{\text{see}}$ see, –wh what]]]])
5. move ([$_{\text{C}}$ +wh C ∅, [$_{\text{did}}$ did, [$_{\text{see}}$ you, [$_{\text{see}}$ see, –wh what]]]]) →
   [$_{\text{C}}$ What C ∅, [$_{\text{did}}$ did, [$_{\text{see}}$ you, [$_{\text{see}}$ see, t_{what}]]]]



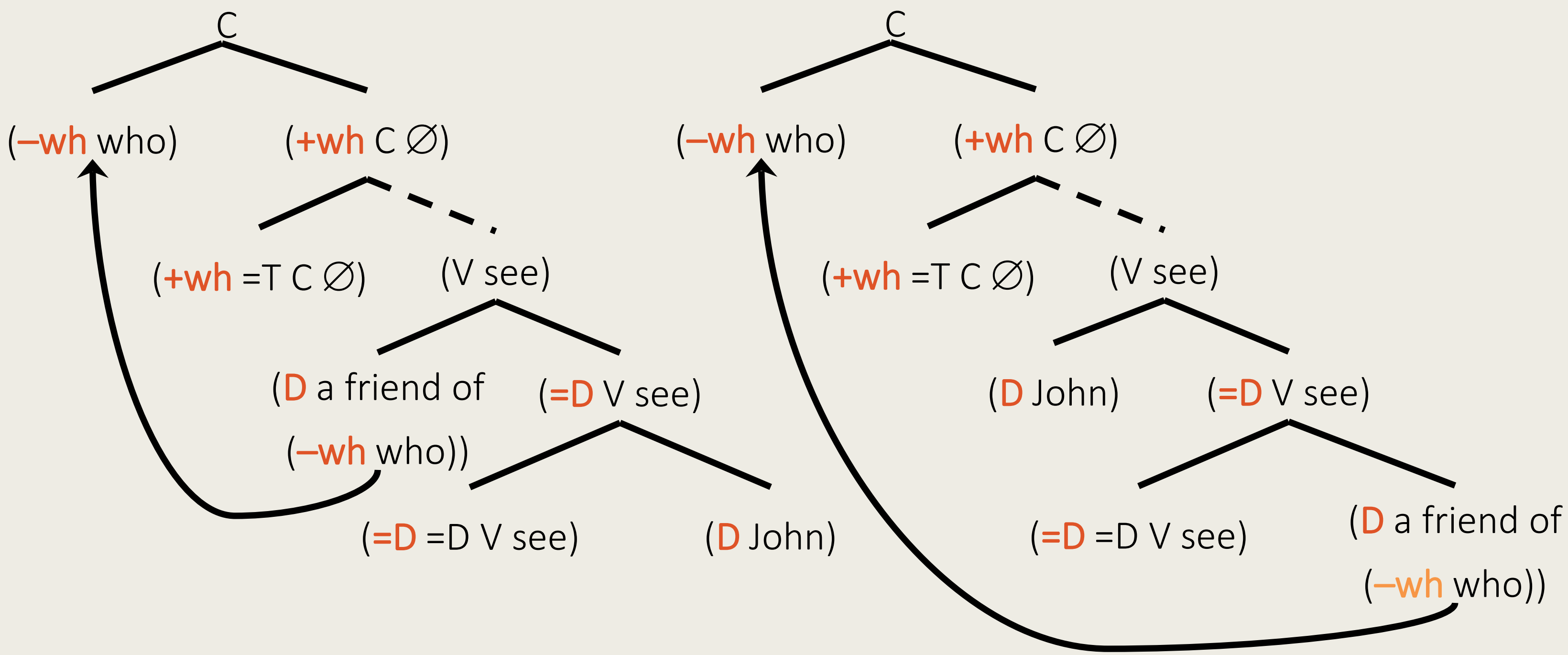
MG: problems with successive cyclicity

- *Wh-* successive cyclic movement



MG: how explaining islandhood?

- No difference in picking up an element from a **subject** or an **object** (idem for **RCs** and **Adjuncts**)



Representations vs. Derivations

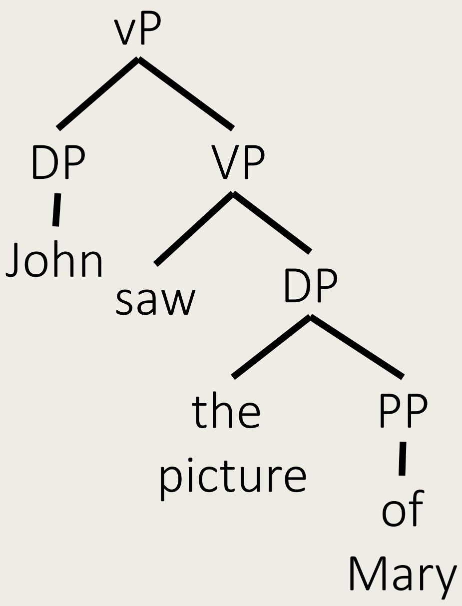
- "the computational system takes representations of a given format and **modifies them**" (Chomsky 1993:6)

- The order of **Structure Building Operation** is abstract with "no temporal interpretation implied" (Chomsky 1995:380)

- **Derivation by Phase** (Chomsky 2005-08): a phase is a Syntactic Object built assuming Structure Building Operations (Merge and Move) over a finite set of Lexical Item (Lexical Array, aka **Numeration**) **CP** and **vP** are phases (maybe **DP**)

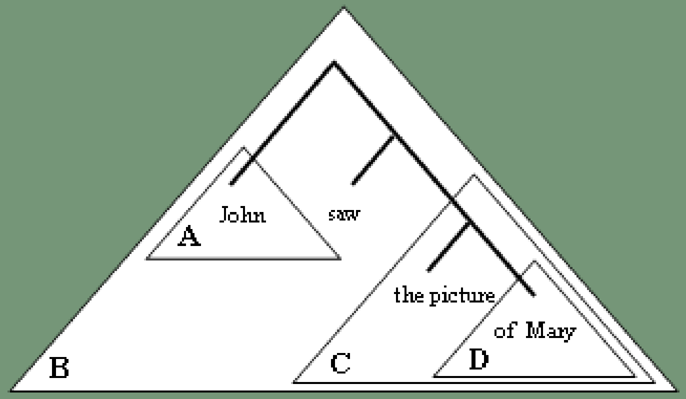
Derivations: some logical possibilities

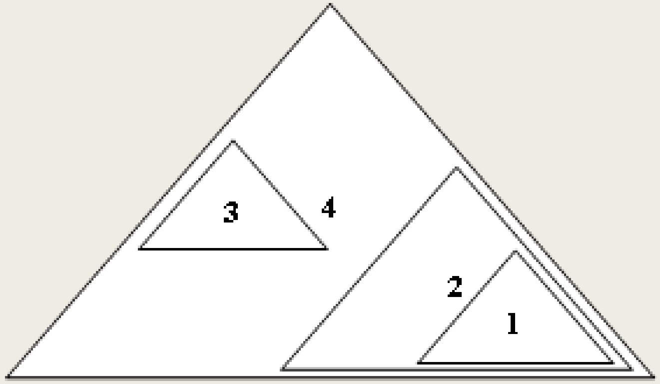
- ((John) saw ((the picture) (of Mary)))



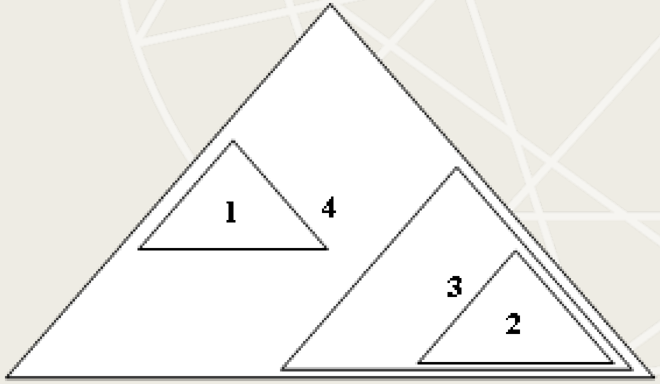
Derivations: Local Relations

$({}_B\ ({}_A$ John) saw $({}_C$ the picture $({}_D$ of Mary)))

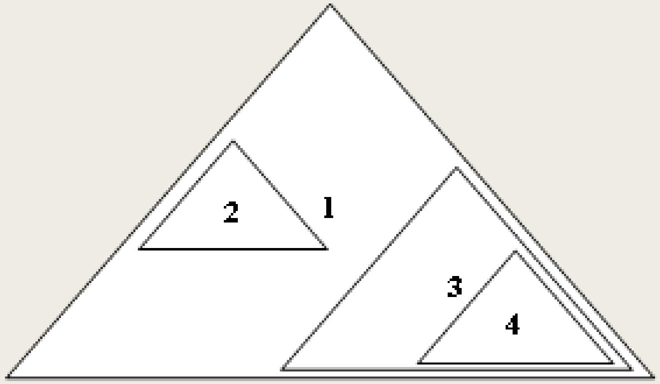




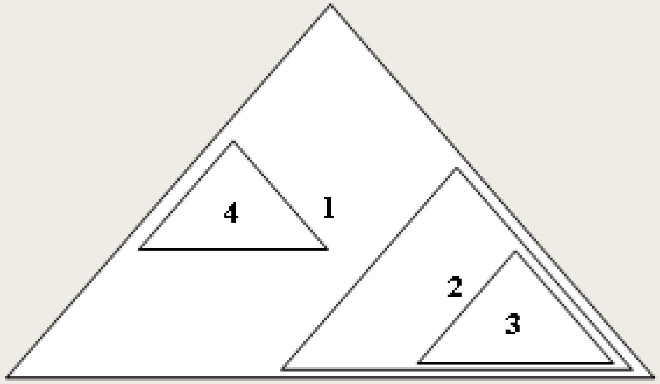
bottom-up, right left



bottom-up, left-right



top-down, left-right



top-down, right-left

Processing Object Relatives (ORs)

- Bever (1970)
  double embedding is not always nearly impossible to process (Miller & Chomsky 1963):

  - **The reporter the politician** **the commentator** met trusts said the president won't resign.
  - **The reporter everyone** **I** met trusts said the president won't resign.

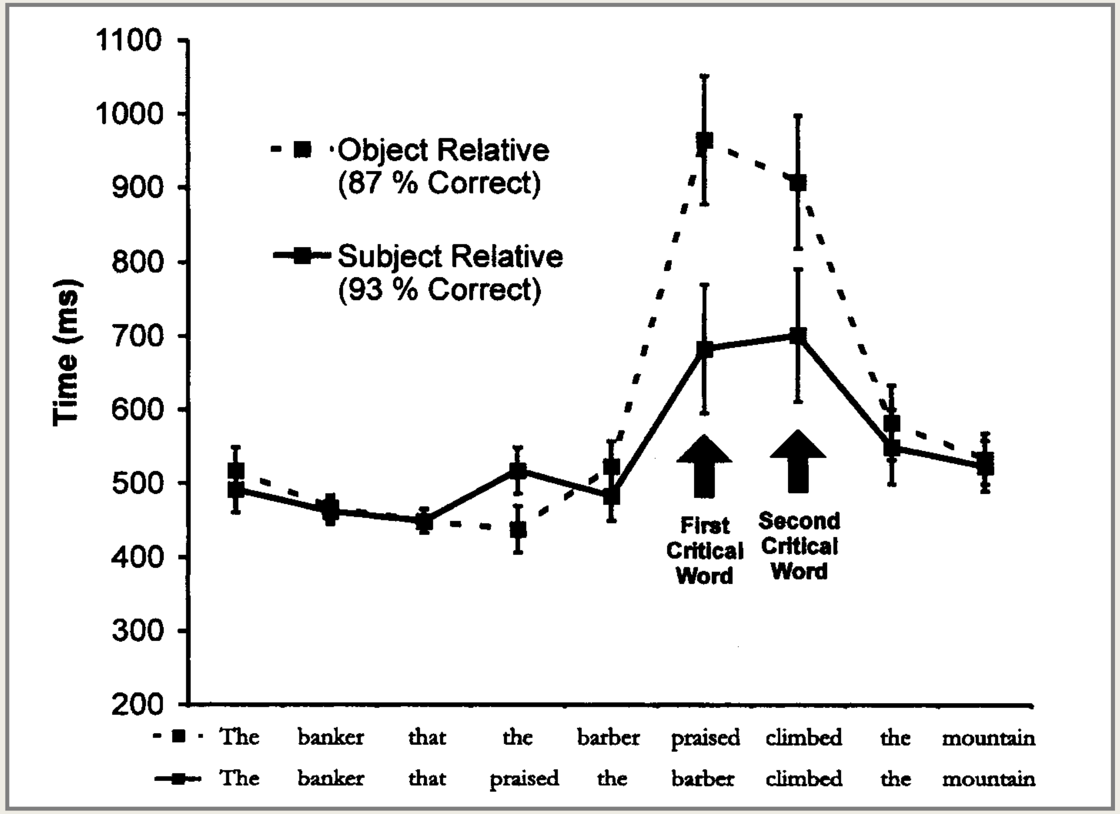
Processing Object Relatives (ORs)

- Gordon, Hendrick & Johnson (2001)
  working memory request is evaluated by studying **reading time** (**RT**) and **comprehension accuracy** in **self-paced reading** experiments comparing critical regions of various kinds of **Relative Clauses**:

- **Experiment 1** (materials): **SRs** (a) and **ORs** (b)
  - The banker [that _ praised the barber] climbed the mountain
  - The banker [that the barber praised _] climbed the mountain

Processing Object Relatives (ORs)

- Gordon et al. (2001) - **Experiment 1** (results)



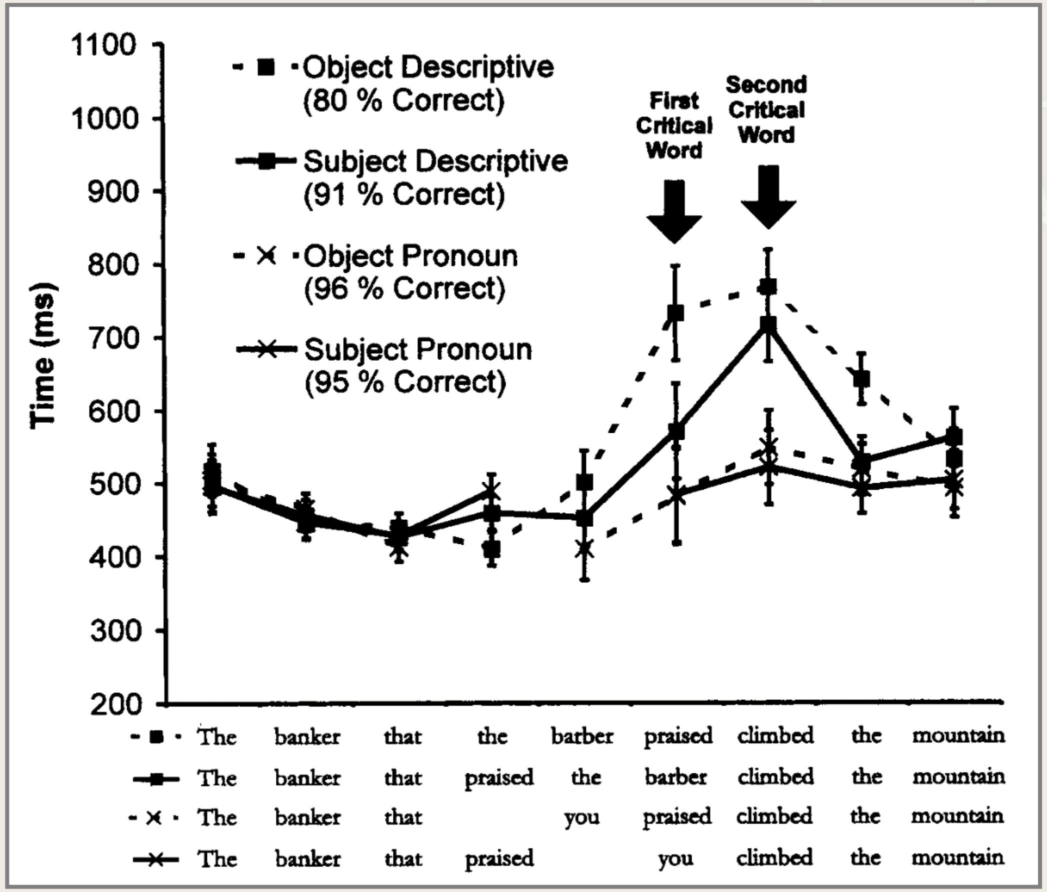
Processing Object Relatives (ORs)

- Gordon et al. (2001) - **Experiment 2**
  complexity can be mitigated by varying the RC Subject typology (reading time (**RT**) and comprehension accuracy in self-paced reading experiments are tested, as before):

- **Experiment 2** (materials): DP (a) vs. Pro (b)
  - The banker [that the barber praised _] climbed the mountain
  - The banker [that you praised _] climbed the mountain

Processing Object Relatives (ORs)

- Gordon et al. (2001)
  Experiment 2 (results)

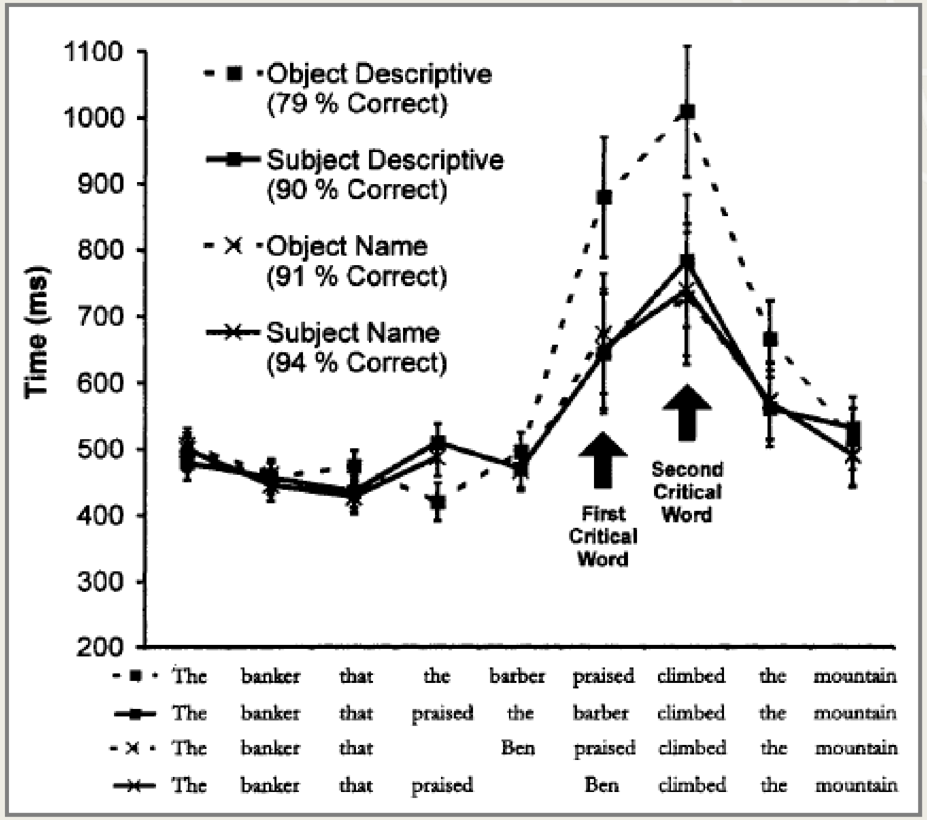


Processing Object Relatives (ORs)

- Gordon et al. (2001) - **Experiment 3** (materials): DP (a) vs. proper names (b)
  - The banker [that the barber praised _] climbed the mountain
  - The banker [that Ben praised _] climbed the mountain

Processing Object Relatives (ORs)

- Gordon et al. (2001)
  Experiment 3 (results)

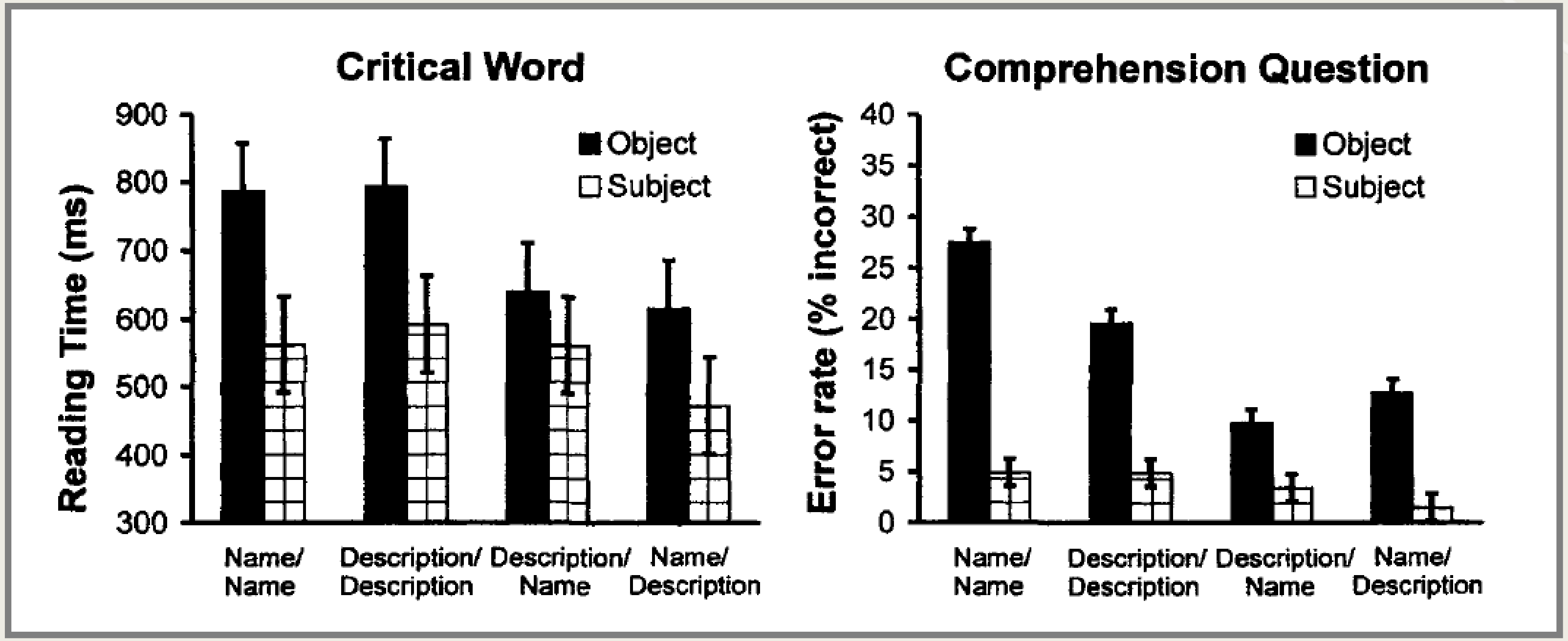


Processing Object Clefts

- Gordon et al. (2001) **- Experiment 4** (materials):
  Subject vs. Object Clefts X DP vs. proper names
  - It was the banker that the lawyer saw _ in the parking lot
  - It was the banker that Bill saw _ in the parking lot
  - It was John that the lawyer saw _ in the parking lot
  - It was John that Bill saw _ in the parking lot

Processing Object Clefts

- Gordon et al. (2001) - **Experiment 4** (results):



Explaining complexity

- **Role-determinant** accounts (MacWhinney & Pleh 1988)
  - Double role for the RC head: **subject** in the matrix sentence, **object** in the RC:
    The banker [that the barber praised _] climbed the mountain (OR)

- **Memory-load** accounts (Ford 1983, MacWhinney 1987, Wanner & Maratsos 1978)
  - The RC head must be **kept in memory longer** in OR before being integrated:

    The banker [**that praised** the barber] climbed … (SR)
    The banker [**that the barber praised** _] climbed … (OR)

Explaining complexity

- **Linguistic Integration Cost** (Gibson 1998:12-13)
  - Processing difficulty is proportional to the **distance** expressed in terms of number of **intervening discourse referents**, following a “**referentiality hierarchy**”:
    descriptions > (short) names > referential pronouns > indexical pronouns

- **Similarity based accounts** (Gordon et al. 2001)
  - Having **two DPs of the same kind** stored in **memory** makes the OR more complex than SR. This models memory interference during encoding, storage and retrieval (Crowder 1976)

Explaining complexity

- More on **Similarity based accounts** (Gordon et al. 2001)
  - It might be able to explain why SR vs. OR asymmetry disappears with RC subject pro/proper names (those DPs are legal heads only for clefts)

- **Intervention effects**
  (Grillo 2008, Friedmann et al. 2009, Rizzi 1990)
  - Processing difficulty is proportional to the number and kind of relevant features shared between the moved item and any possible intervener:

Explaining complexity

- More on **Intervention effects** (Friedmann et al. 2009)
  - **Identity** (bad for adults, bad for children)

    +A ***+A*** ***(+A)***

  - **Inclusion** (ok for adults, bad for children)

    +A +B ***+A*** ***(+A +B)***

  - **Disjunction** (ok for adults, ok for children)

    +A ***+B*** ***(+A)***

Kinds of non-local dependencies

Long distance *Wh*- dependencies

intervener *intervener*

$[_{CP}$ ***What*** do **you** think $[_{CP}$ _ **Mary** will $[_{VP}$ ***buy*** _]]] ?

criterial *intermediate* *argument*

Kinds of non-local dependencies
Object Clefts

- In **Object Clefts** (**OCs**), the **copula** selects a truncated CP (Belletti 2008):

It is [$_{FocP}$ ***an ice cream*** that [$_{TP}$ ***Mary*** will ***buy*** _]]

… BE [$_{CP}$ ~~Force~~ [$_{FocP}$ … [$_{FinP}$ that [$_{TP}$ ***Subject*** … ***Object***]]]]

Comparing Object Clefts

- Warren & Gibson (2005) - **Experiment** (materials):
  definite descriptions vs. **proper names** vs. **pronouns**

a.	It was **the banker**	that **the lawyer**	**avoided _** at the party
b.	It was **the banker**	that **Dan**	**avoided _** at the party
c.	It was **the banker**	that **we**	**avoided _** at the party
d.	It was **Patricia**	that **the lawyer**	**avoided _** at the party
e.	It was **Patricia**	that **Dan**	**avoided _** at the party
f.	It was **Patricia**	that **we**	**avoided _** at the party
g.	It was **you**	that **the lawyer**	**avoided _** at the party
h.	It was **you**	that **Dan**	**avoided _** at the party
i.	It was **you**	that **we**	**avoided _** at the party

Comparing Object Clefts

- Warren & Gibson (2005) - results (Tessa Warren P.C.)
  D = definite description (e.g. **the banker**)
  N = proper names (e.g. **Dan**)
  P = pronouns (e.g. **you**)

condition	**D-D**	**D-N**	**D-P**	**N-D**	**N-N**	**N-P**	**P-D**	**P-N**	**P-P**
Read. time (SE) ms	**365** (19)	**319** (12)	**306** (14)	**348** (18)	**347** (21)	**291** (14)	**348** (18)	**311** (15)	**291** (13)

Predicting reading times (rt) with intervention-based accounts

- Assuming that **Definite Description** = {+NP, N}, **Proper Names** = {+NP, NProper}, **pro** = {} (Belletti & Rizzi 2013), Intervention effects are predicted to be stronger in matching **D-D** and **N-N** condition (against memory-load accounts), while **P-P** is expected not to be critical (because of the +NP absence):

condition	D-D	D-N	D-P	N-D	N-N	N-P	P-D	P-N	P-P
Read. time (SE) ms	**365** (19)	**319** (12)	**306** (14)	**348** (18)	**347** (21)	**291** (14)	**348** (18)	**311** (15)	**291** (13)
prediction	hard	?	easy	?	hard	easy	easy	easy	easy

Some problems with the intervention-based account

- Features triggering movement are those relevant for intervention (Friedmann et al. 2009:82), but:
  - "**+R**" feature causing Object movement in ORs (or "**+Foc**" in OCs) is not present on Subject;
  - Neither the "**lexical restriction**" nor **phi-features** trigger any movement in **ORs** or **OCs**
  - The "**lexical restriction**" should be not accessible at the **edge of the DP**, where features triggering movement should be located (but see Belletti & Rizzi 2013, next slide)
  - Why slow-down is observed at **verb segment**?

Some problems with the intervention-based account

- Belletti & Rizzi 2013:
  - Evidence that lexically restricted wh-items occupy different positions in the left periphery (Munaro 1999):

a. Con **che tosat** à-tu parlà?
with which boy did you speak?

b. Avé-o parlà de **chi**?
Have you spoken of whom?

Feature Retrieval Cost (FRC)
Why do we need it? (a summary)

- An “integration cost” (cf. Gibson 1998) is **not enough**

È **il bambino**	che	***il signore***	ha salutato …
È **Luigi**	che	***Gianni***	ha salutato …
It is {the boy/L.}	that	{the man/G.}	greeted …

- **Intervention-based** accounts are **not “gradable”** (no quantitative, precise, measurements)

- **Bottom-Up** standard theories **do not make clear predictions on processing**: they predict **what** creates complexity, but not **when**, **why** and **how** exactly in **parsing** and **generation**?

The notion of “expectation”

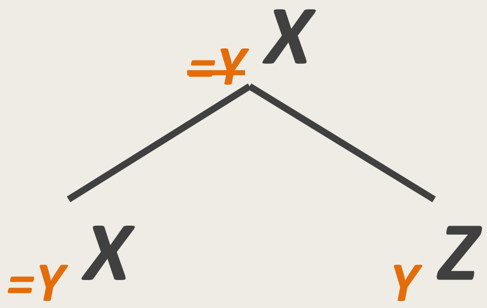
- Robust statistical approaches (GPT-like):
  - Roger Levy’s **relative-entropy**-based approach (Levy 2008)
  - John Hale’s **surprisal**-based approach (Hale 2011)

- Our modest goal:
  - how far we can go if we assume that structure building is only driven by categorial, lexically encoded, expectations?
  - The proposal should then be precise enough to allow one scholar to compare specific assumptions (“parameters”, Chesi 2023: doi.org/10.4000/ijcol.1135)
  - https://github.com/cristianochesi/e-MGs

Processing-friendly Minimalist Grammars

Phase and Expectation-based MGs (PMGs and e-MGs)

- Common restriction on **Merge**:
  - Given two lexical items [$_{=Y}$ X] and [$_{Y}$ Z] such that X selects Z, then:



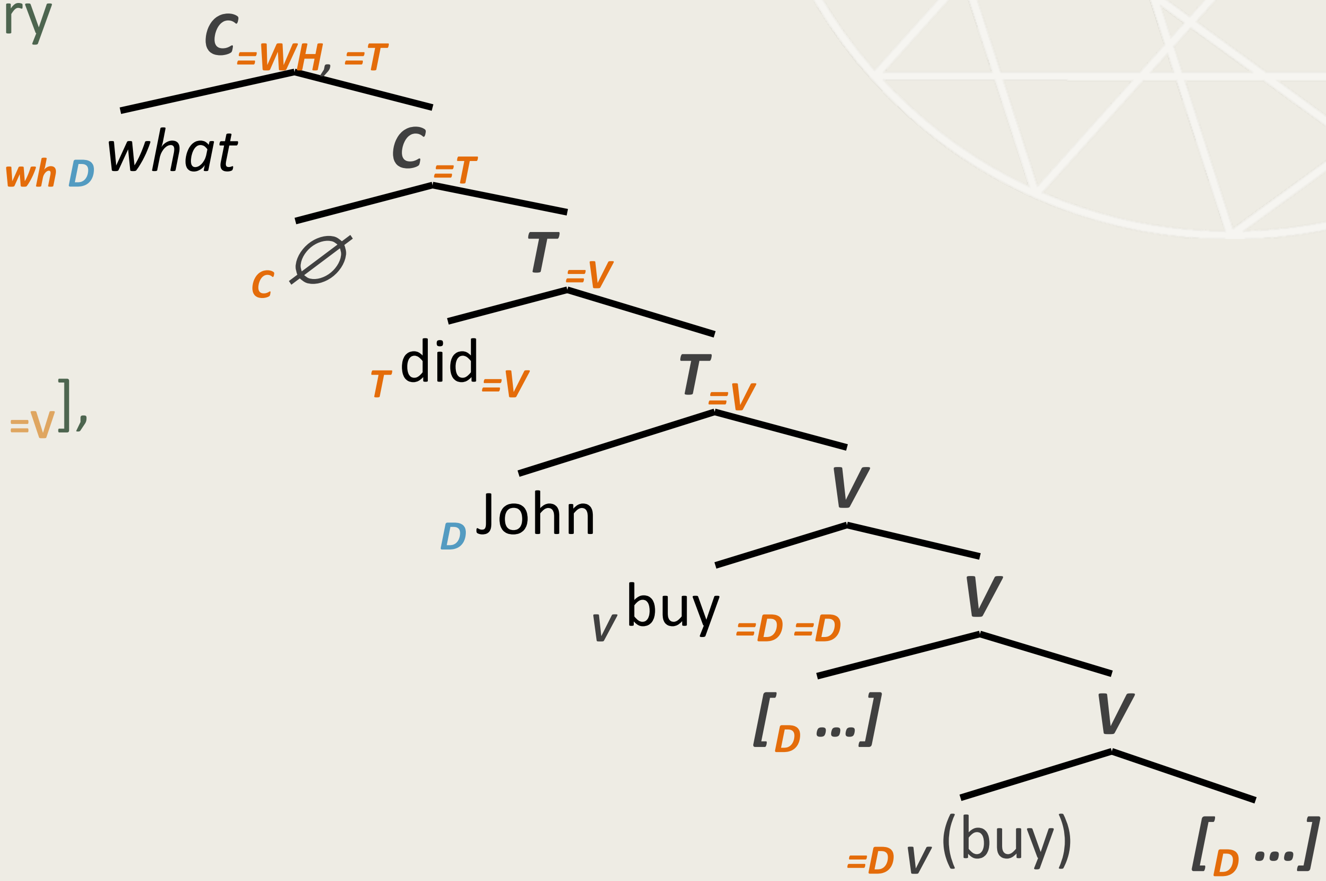
  - [$_{=Y}$ X] is processed before Y
  - When [$_{=Y}$ X] is processed, an expectation for [$_{Y}$ ...] is created

Processing-friendly Minimalist Grammars

Expectation-based MGs (e-MGs)

- A **phase head** is a lexical category (**N**, **V**, **A**)

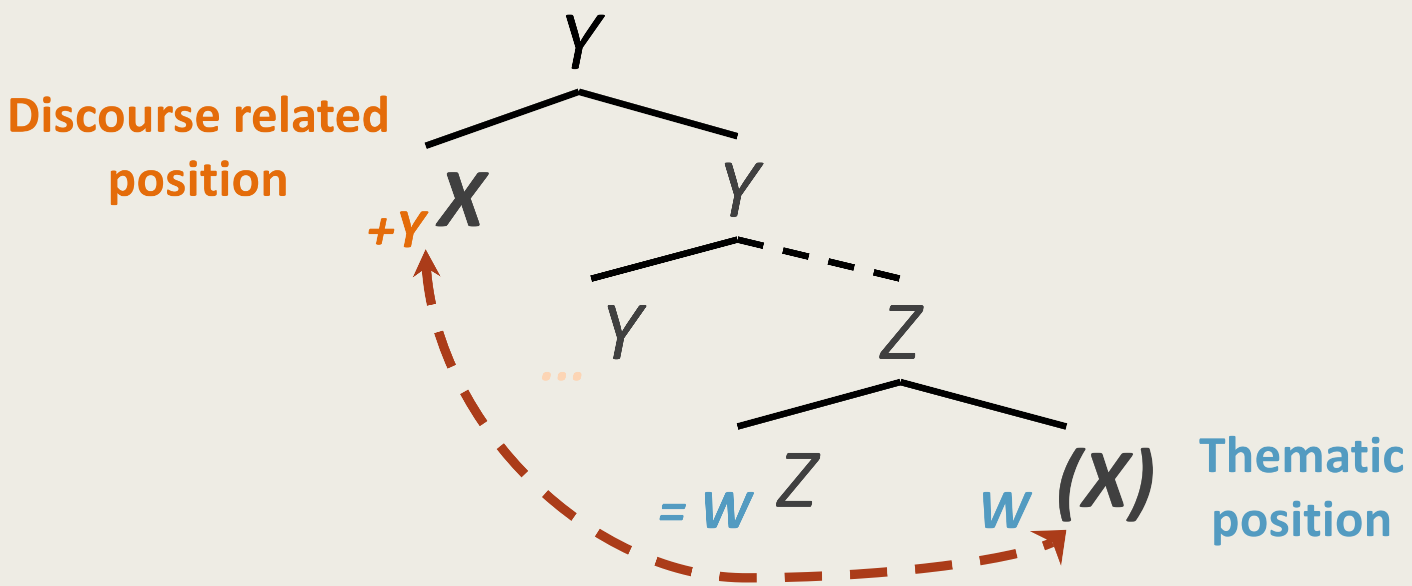
- $_{root}[_{C}\ \varnothing_{=wh\ =T}]$, $[_{wh\ D}$ what$]$, $[_{T}$ did $_{=V}]$, $[_{D}$ John$]$, $[_{V}$ buy$_{=DP\ =DP}]$



Processing-friendly

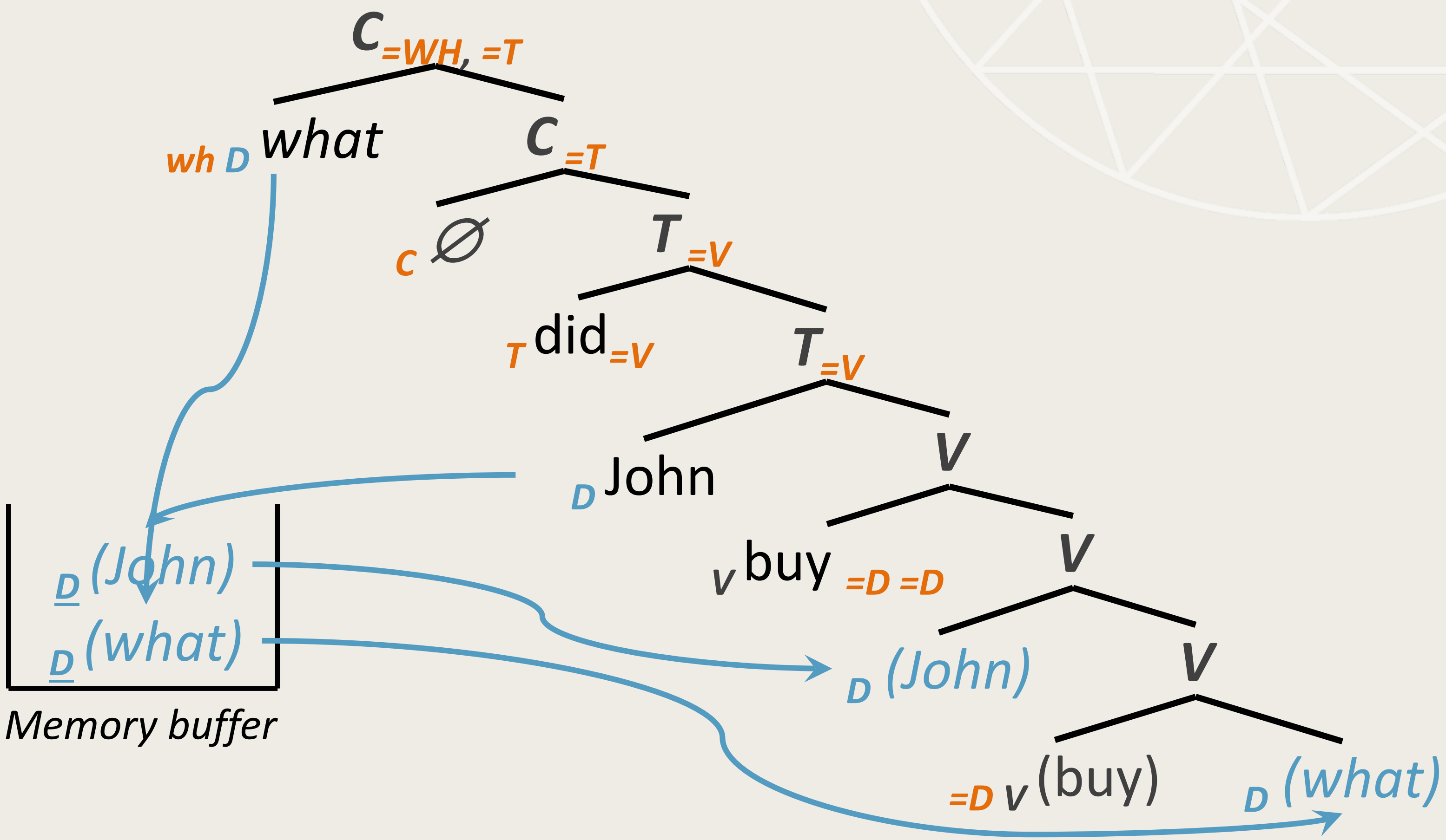
Phase and Expectation-based MGs (PMGs and e-MGs)

- Common trigger for Move:
  - An item [$_{+Y \ldots W}$ X], in a given structure, must be moved if it can not be fully interpreted in its insertion position:



Processing-friendly
Expectation-based MGs (e-MGs)

- $_{root}[_{C}\ \varnothing\ _{=wh\ =T}]$, $[_{T}$ did $_{=V}]$, $[_{V}$ buy$_{=DP\ =DP}]$, $[_{D}$ John$]$, $[_{wh\ D}$ what$]$



Processing-friendly PMGs / e-MGs

- The derivation unfolds **Top-Down** and (as a consequence) **Left-Right**
- **Unexpected features** trigger **movement**
- **Phases** restrict the domain in which a **non-local dependency** must be satisfied
- **Last-In-First-Out memory** buffer, as a first approximation, is used to store and retrieve items for **non-local dependencies** (memory buffer must be empty at the end of the derivation)
- The order in which phases are expanded makes a difference: the last selected phase has a special status (**sequential phase**) while phases that are not the last selected ones (e.g. phases that results from expansion of functional features) qualifies as **nested phases** (Bianchi & Chesi 2006)

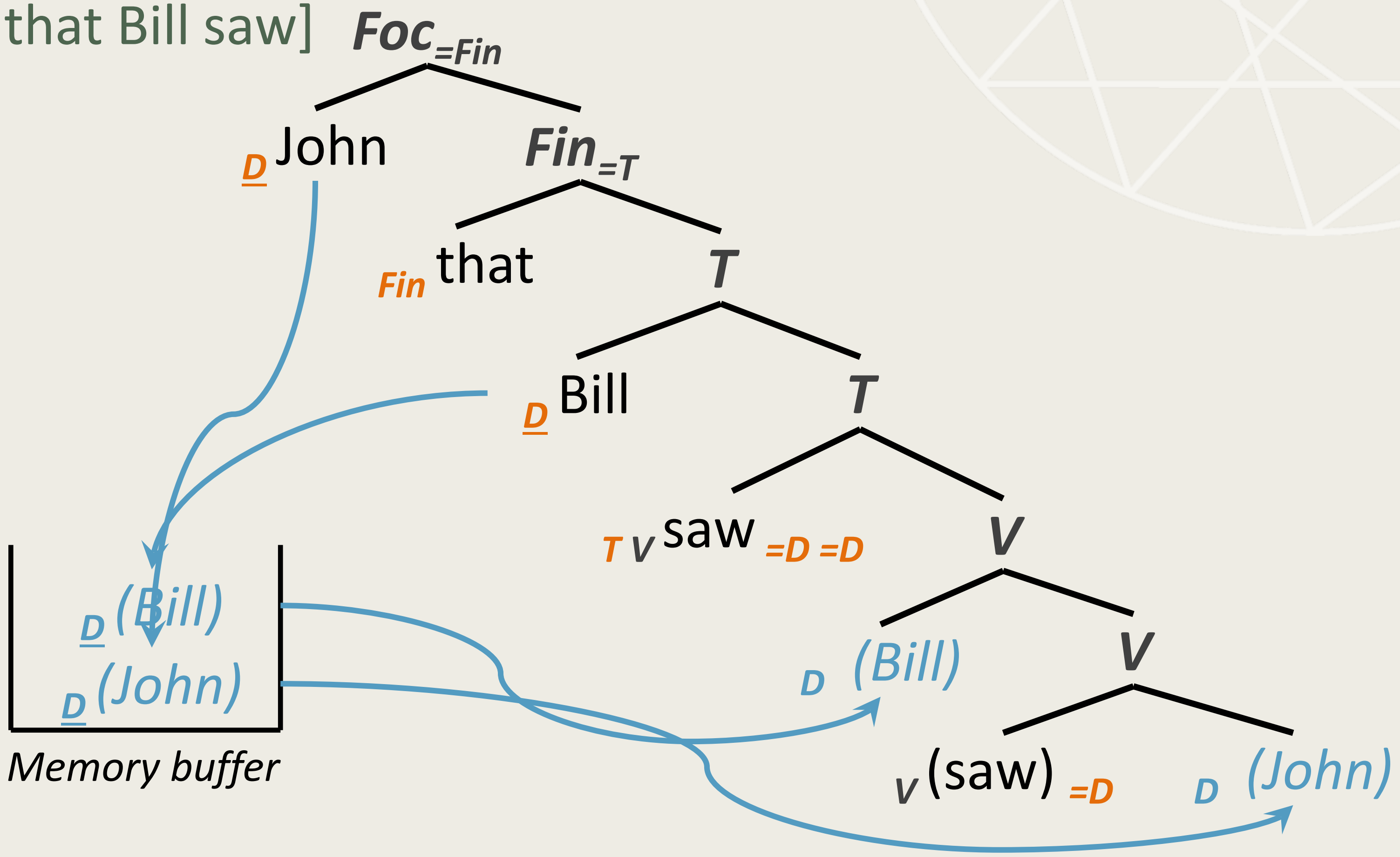
Deriving OCs Top-Down

- In Object Clefts (OCs), the copula selects a truncated CP (Belletti 2008):

  … BE [$_{CP}$ Force [$_{FocP}$ … [$_{FinP}$ che [$_{TP}$ Subject … Object]]]]

Deriving OCs Top-Down

- It [… $_{=CPr}$ … was] [$_{CPr}$ John that Bill saw]



Cue-based retrieval and intervention

- **interference** is the major constraint on accessing information in memory (Anderson & Neely 1996; Crowder 1976; see Nairne 2002 for a review).
- the locus of the interference effect is at **retrieval**, with little or no effect on memory encoding or storage (Dillon & Bittner 1975; Gardiner et al. 1972; Tehan & Humphreys 1996)
- **Content-adressable memory** (e.g. memory load paradigm, Van Dyke & McElree 2006), no exhaustive search, no delay
- Search of **Associative Memory** (**SAM**) model (Gillund & Shiffrin 1984)

$$P(I_i | Q_1, \ldots, Q_n) = \frac{\prod_{j=1}^{m} S(Q_j, I_i)^{w_j}}{\sum_{k=1}^{N} \prod_{j=1}^{m} S(Q_j, I_k)^{w_j}}$$

On DP features (and structure)

- Elbourne (2005)
  [[THE *i*] NP]

- Zamparelli (1995-2000)
  [$_{SDP}$ Strong QP [$_{PDP}$ Week QP [$_{KIP}$ (Restrictive Adj) [$_{NP}$ Noun]]]]

- Longobardi (1994-2005), a rough summary:
  - **Definite Descriptions** [$_D$ the [$_N$ man]]
  - **Proper Names** [$_D$ John$_i$ [$_N$ t_i]]
  - **Pronouns** [$_D$ you [$_N$ $\varnothing$]]

Relevant DP features
Definite Descriptions & Proper Names

- Both proper names and common nouns have category N

***N in situ* (common nouns)**	***N-to-D raising***
Il mio Gianni (Il mio amico)	*mio Gianni
La sola Maria (la sola amica)	Maria sola (*l'amica sola)

- Two different kinds of N: $\mathbf{N_{proper}}$, $\mathbf{N_{(common)}}$

Relevant DP features
On D and Pronouns

- Both **determiners** and **personal pronouns** introduce a "**referential pointer**" to an individual constant or variable in the domain of discourse

- **Pro** are **NP-ellipsis licensors** (they can be used as determiners «we italians»):
  $[_D$ noi $[_N$ ~~italiani~~$]]$
  (**D** introduces an *index*, that bounds a variable predicated in N)

- (More) features on **pro**:
  - **1st** and **2nd** person (highly accessible referents) vs. **3rd** person (**default person**, context-determined referent)
  - **case**

Relevant DP features

- **Definite descriptions:** {D, N}
- **Proper names:** {D, N_{prop}}
- **Pronouns:** {D, case, pers}

Feature Retrieval Cost (FRC) metrics at work

- Cost function (at ***X*** given $\boldsymbol{m_x}$ items to be retrieved from memory)

- FRC(x) = $\prod_{i=1}^{m_x} \frac{(1+nF_i)^{m_i}}{(1+dF_i)}$

  - ***m*** = number of items stored in memory at retrieval
  - ***nF*** = new features to be retrieved from memory
  - ***dF*** = number of distinct cued features (e.g. agreement and case features probed by the verb)

Feature Retrieval Cost (FRC) metrics at work

$$\text{FRC}(x) = \prod_{i=1}^{m_x} \frac{(1+nF_i)^{m_i}}{(1+dF_i)}$$

- ***D-D*** matching

  it was **the lawyer**$_{\{D,\ N\}}$ who **the businessman**$_{\{D,\ N\}}$ *avoided…*

  FRC (*avoided*) = 27

  that is **9 · 3**:

  9 for retrieving **the businessman**,

  since ***nF*=2** (***D*** and ***N*** count as one), ***m*=2** because two DPs are in memory at this time, and ***dF*=0** because no feature is cued by the verb distinguishing one DP from the other;

  3 for retrieving **the lawyer**,

  since ***nF*=2** (D and N are new now), ***m*=1** and ***dF*=0**

Feature Retrieval Cost (FRC) metrics at work

FRC(x) = $\prod_{i=1}^{m_x} \frac{(1+nF_i)^{m_i}}{(1+dF_i)}$

- ***N-N*** matching

  it was **Dan**$_{\{D,\ N\_prop\}}$ who **Patricia**$_{\{D,\ N\_prop\}}$ *avoided…*

  FRC (*avoided*) = **18**

  that is **9 · 2**:
  9 for retrieving **Dan**,
  nF*=2** (even though ***D should be contextually salient, being two proper names presents, the same ***D***, i.e. a co-referential index, cannot be sufficient to distinguish them, then an extra cost must be paid here as in the ***D-D*** condition), ***m*=2**, ***dF*=0**;
  2 for retrieving **Patricia**,
  since ***nF*=0** (just N is new since the determiner is now contextually salient and unique, *m*=1 and *dF*=0)***m*=1** and ***dF*=0**

Feature Retrieval Cost (FRC) metrics at work

$$\text{FRC}(x) = \prod_{i=1}^{m_x} \frac{(1+nF_i)^{m_i}}{(1+dF_i)}$$

- ***P-P*** matching

  it was **you**$_{\{D,\ pers\_II,\ case\}}$ who **we**$_{\{D,\ pers\_I,\ case\_nom\}}$ *avoided…*

  FRC (*avoided*) = **4**

  that is **2 · 2**:

  2 for the **we**, ***nF*=1**, ***m*=2** and ***dF*=1** (**number**, **person** and **case** mismatches are always present; **case** is cued by the verb),

  2 for retrieving **you**, ***nF*=1**, ***m*=1** and ***dF*=0** for the object pronoun

Feature Retrieval Cost (FRC) metrics at work

FRC(x) = $\prod_{i=1}^{m_x} \frac{(1+nF_i)^{m_i}}{(1+dF_i)}$

- ***D-N*** matching

  it was **the lawyer**$_{\{D,\ N\}}$ who **Patricia**$_{\{D,\ N\_prop\}}$ *avoided…*

  FRC (*avoided*) = 12

  that is **4 · 3**:
  4 for **Patricia**, *nF*=1, that is N, since D is contextually salient, m=2, dF=0,
  3 for retrieving **the lawyer** (nF=2, m=1, nF=0)

Feature Retrieval Cost (FRC) metrics at work

$$\text{FRC}(x) = \prod_{i=1}^{m_x} \frac{(1+nF_i)^{m_i}}{(1+dF_i)}$$

- ***D***-***P*** condition

  it was **the lawyer**$_{\{D, N\}}$ who **we**$_{\{D, pers\_I, case\_nom\}}$ *avoided…*

  FRC (*avoided*) = 6

  that is **2 · 3**:
  2 for retrieving **we** (*nF*=1 even if deictic pronouns are contextually salient, the correct person must be retrieved, *m*=2, *dF*=1 since a distinct case on pronouns is cued by the verb),
  3 for retrieving **the lawyer** (*nF*=2, *m*=1, *nF*=0)

Feature Retrieval Cost (FRC) metrics at work

FRC(x) = $\prod_{i=1}^{m_x} \frac{(1+nF_i)^{m_i}}{(1+dF_i)}$

- ***P*-*D*** condition

  it was **you**$_{\{D,\ pers\_II,\ (case)\}}$ who **the businessman**$_{\{D,\ N\}}$ *avoided…*

  FRC (*avoided*) = 18

  that is **9 · 2**:
  9 for the **the businessman** (*nF*=2, *m*=2, *dF*=0);
  2 for retrieving **you** (*nF*=1, *m*=1, *dF*=0);

Feature Retrieval Cost (FRC) metrics at work

- The complete prediction set:

condition	D-D	D-N	D-P	N-D	N-N	N-P	P-D	P-N	P-P
Read. time (SE) ms	**365** (19)	**319** (12)	**306** (14)	**348** (18)	**347** (21)	**291** (14)	**348** (18)	**311** (15)	**291** (13)
prediction log(FRC)	1,43	1,08	0,78	1,26	1,26	0,60	1,26	0,90	0,69

Feature Encoding Cost (FEC)

- **Feature Encoding Cost** (*FEC*) is a numerical value associated to each new item merged that is proportional to the number of new relevant features integrated in the structure:

$$FEC(x) = \sum_{i=1}^{n} eF_i$$

- eF is the cost of each new relevant feature to be encoded at *x*.
- For simplicity $\boldsymbol{eF}$ = **1** for a **new categorial feature** introduced (e.g. 1 for D and 1 for N), 2 for a **duplication** of the same lexical category requiring structural integration (i.e. 2 for the second N both in D_1-D_2 and N_1-N_2), 0 otherwise.

Feature Encoding Cost (FEC)

	object$_{focalized}$	*subject*	*verb*	*spill-over*	*condition*
a.	It was *(1)* **the banker** *(2)*	that *(1)* **the lawyer** *(3)*	**avoided** _	*(2)* at the party *(3)*	*[D_1-D_2]*
b.	It was *(1)* **the banker** *(2)*	that *(1)* **Dan** *(1)*	**avoided** _	*(2)* at the party *(3)*	*[D_1-N_2]*
c.	It was *(1)* **the banker** *(2)*	that *(1)* **we** *(0)*	**avoided** _	*(2)* at the party *(3)*	*[D_1-P_2]*
d.	It was *(1)* **Patricia** *(1)*	that *(1)* **the lawyer** *(2)*	**avoided** _	*(2)* at the party *(3)*	*[N_1-D_2]*
e.	It was *(1)* **Patricia** *(1)*	that *(1)* **Dan** *(2)*	**avoided** _	*(2)* at the party *(3)*	*[N_1-N_2]*
f.	It was *(1)* **Patricia** *(1)*	that *(1)* **we** *(0)*	**avoided** _	*(2)* at the party *(3)*	*[N_1-P_2]*
g.	It was *(1)* **you** *(0)*	that *(1)* **the lawyer** *(2)*	**avoided** _	*(2)* at the party *(3)*	*[P_1-D_2]*
h.	It was *(1)* **you** *(0)*	that *(1)* **Dan** *(1)*	**avoided** _	*(2)* at the party *(3)*	*[P_1-N_2]*
i.	It was *(1)* **you** *(0)*	that *(1)* **we** *(0)*	**avoided** _	*(2)* at the party *(3)*	*[P_1-P_2]*

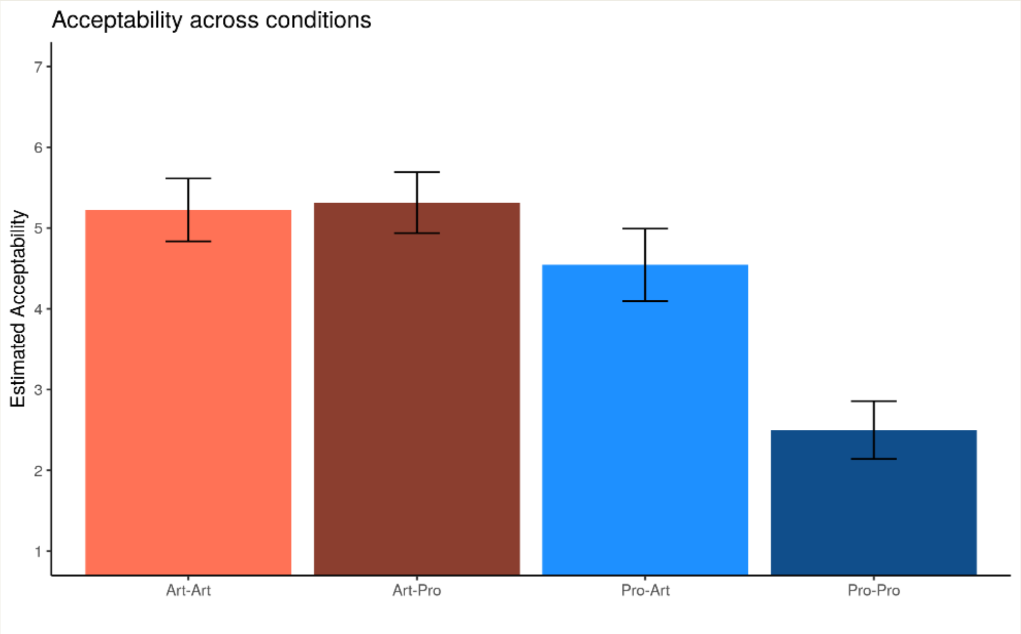
Chesi & Canal (2019)

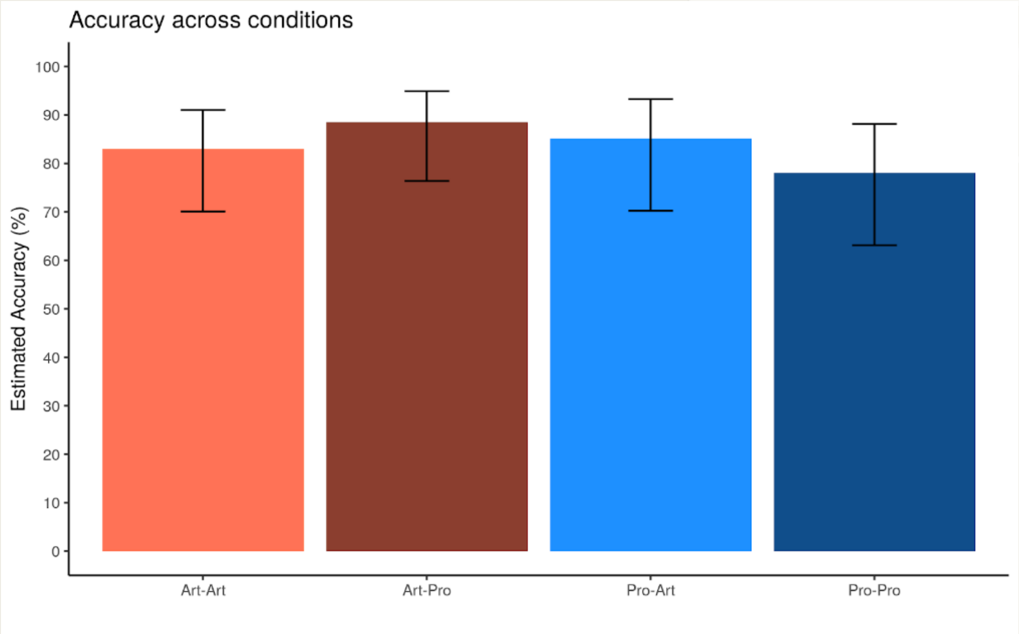
	***object**$_{focalized}$*	***subject***	***verb***	***spill-over***	***condition***
a.	Sono [**gli** architetti]$_i$ che *are$_{3P_PL}$ **the** architects that*	[**gli** ingegneri] ***the** engineers*	**hanno** consultato _$_i$ ***have$_{3P_PL}$** consulted*	prima di iniziare i lavori. *before beginning the works*	**D$_{art}$-D$_{art}$**
b.	Sono [**gli** architetti]$_i$ che *are$_{3P_PL}$ **the** architects that*	[**voi** ingegneri] ***you** engineers*	**avete** consultato _$_i$ ***have$_{2P_PL}$** consulted*	prima di iniziare i lavori. *before beginning the works*	**D$_{art}$-D$_{pro}$**
c.	Siete [**voi** architetti]$_i$ che *are$_{2P_PL}$ **you** architects that*	[**gli** ingegneri] ***the** engineers*	**hanno** consultato _$_i$ ***have$_{3P_PL}$** consulted*	prima di iniziare i lavori. *before beginning the works*	**D$_{pro}$-D$_{art}$**
d.	Siete [**voi** architetti]$_i$ che *are$_{2P_PL}$ **you** architects that*	[**voi** ingegneri] ***you** engineers*	**avete** consultato _$_i$ ***have$_{2P_PL}$** consulted*	prima di iniziare i lavori. *before beginning the works*	**D$_{pro}$-D$_{pro}$**

Chesi & Canal (2019)

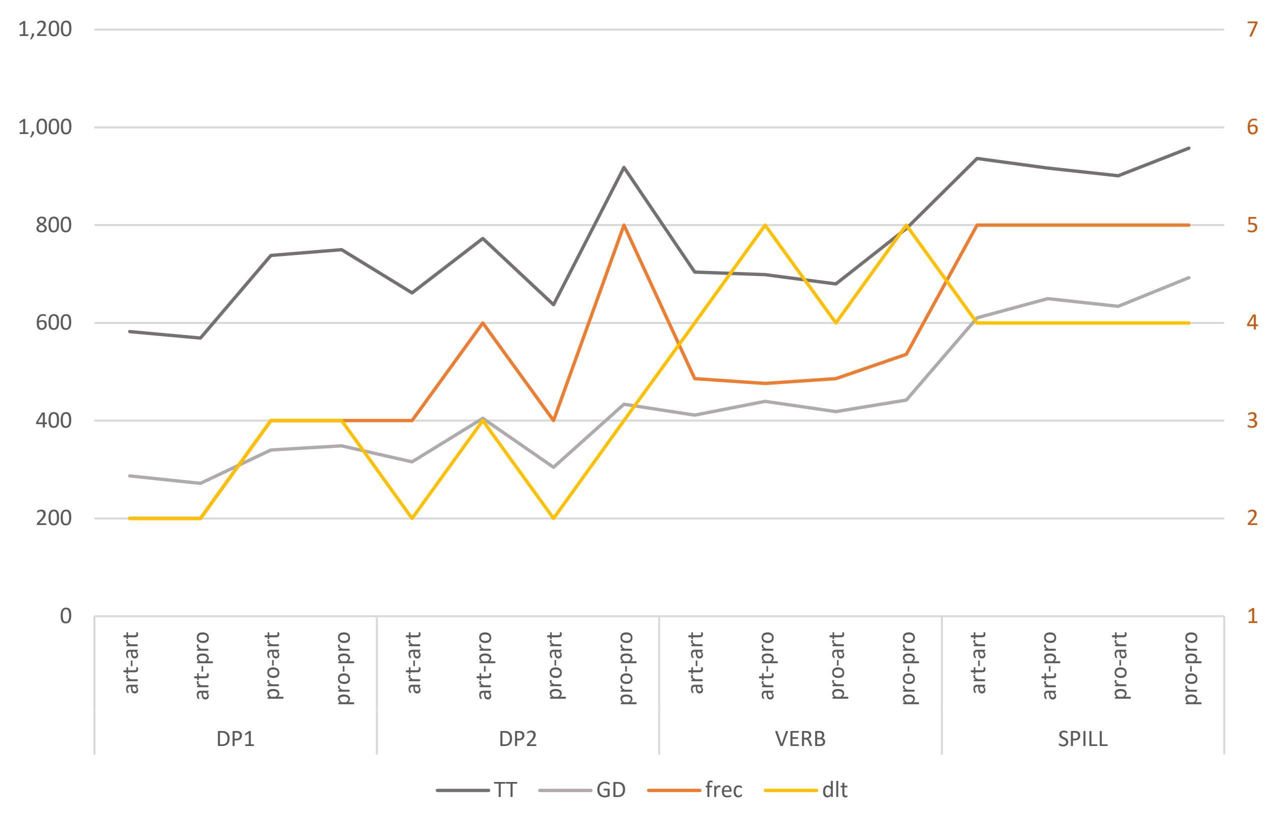
condition	Art_1-Art_2	Pro_1-Pro_2	Art_1-Pro_2	Pro_1-Art_2
Similarity-based prediction	hard	hard	medium	medium
Intervention-based prediction	hard	hard	medium	medium
Top-down prediction (FRC) – H1	hard	hard	medium	medium
Top-down prediction (FRC) – H2	hard	hardest	medium	hard
Memory-load prediction – A1	hard	hard	hard	hard
Memory-load prediction – A2	harder	hard	hard	harder
Memory-load prediction – A3	hard	harder	harder	hard
ACT-R-based prediction	hard	hard	hard	hard

Chesi & Canal (2019)





Chesi & Canal (2019)



Conclusion

- We rephrased the **intervention-based** idea (Friedmann et al. 2009) in **Top-Down** terms, trying to reconcile the formal account of intervention (**what**) with processing evidence (**when** and **how**)

- What permits to express the exact **complexity cost** is a **Top-down** (that in the end produce a **left-right**) derivation (this way the model fitting can be directly compared with other complexity metrics, e.g. SPLT, Gibson 1998)

- The special role of intervention has been expressed in terms of **interference** at **retrieval** (e.g. Van Dyke & McElree 2006)

Further development

- Feature structures (and actual cues) need to be further refined (other features, e.g. animacy, Kidd et al. 2007, and semantic selection, Gordon et al. 2004, should be considered)
- The counterintuitive idea that Subject “is harder” to retrieve than Object in ORs should receive experimental support
- Is it a purely privative system (+/- F) enough?
- Doing away with LIFO structure which is computationally OK, but psycholinguistically odd (cf. content-adressable memory).

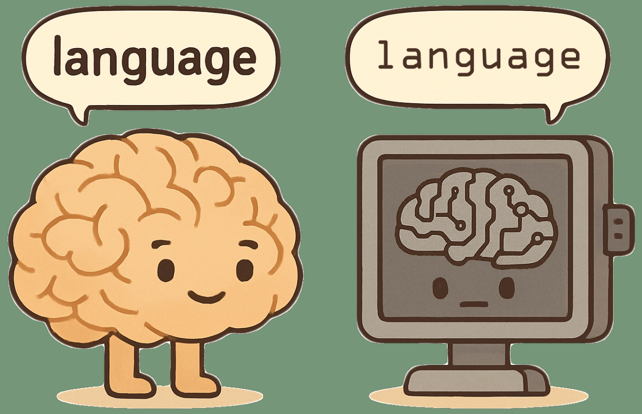
Crucial concepts of this course

- What's a formal grammar and why do we need to specify it
  - Rewriting rules and recursion
  - Restrictions on rule format and generative power (Chomsky's hierarchy)
  - Equivalence between grammars, finite state automata and push-down automata
  - Where natural languages are located in Chomsky's Hierarchy
- What's a computation
  - Problem space and its (algorithmic) exploration
  - Complexity calculus
  - Parsing algorithms (Earley)
- What's a Top-Down derivation
  - A reconciling view of Competence and Performance
  - Reconstruction and islands
  - Predictions and phases
  - Complexity and intervention (possibly in terms of retrieval)

STAGE IV

Large Language Modeling

(our digital twin?)



Outline

- Models for **language acquisition**
  - The poverty of stimulus hypothesis
- From **children** to **machines**
  - Recurrent Neural Networks and incrementality
  - Attention mechanism
  - Training and assessment
- A little experiment on linguistic biases
  - Minimalism
  - BabyLM challenge
  - Some experiment on English & Italian

The Poverty of Stimulus argument

Chomsky (1975), Pullum and Scholz (2002), Lasnik and Lidz (2017)

1. Speakers do **acquire** some aspect of grammatical representation;
2. The data the child is exposed to is consistent with **multiple representations**;
3. There are **“trigger” data** that could be used to distinguish the true representation from the alternatives;
4. That data does **not exist** in the **primary linguistic data**;

- **Conclusion**: the aspect of the grammatical representation acquired in (1) is not determined by experience but by properties internal to the learner

The Poverty of Stimulus argument

2. The data the child is exposed to is consistent with **multiple representations**

- **Yes-No questions** in English
  - *The man [who **is** tall] **is** happy*
  - ***Is**$_i$ the man [who **is** tall] _$_i$ happy?*
- Possible rules
  - Move the **third word** in front of the sentence

    Who the man [_$_i$ **is** tall] **is** happy*
  - Move the **first auxiliary** in front of the sentence

    Is the man [who _$_i$ tall] **is** happy*
  - Swap the **matrix auxiliary** with the matrix subject

2013 animated documentary film by **Michel Gondry**

The Poverty of Stimulus argument

1. Speakers do **acquire** some aspect of grammatical representation

Crain & Nakayama (1987)

- *30 children, **3-** to **5-year-old** (divided in two groups)*
- ***Elicitation task*** (Bellugi 1971): Jabba the Hutt (from Star Wars) was the target of the child question elicited with a prompted picture representing a complex situation
- Experimenter:

  "Ask Jabba if the boy who is watching Mickey Mouse is happy"

The Poverty of Stimulus argument

1. Speakers do **acquire** some aspect of grammatical representation

Crain & Nakayama (1987)

- Type of possible errors:
  - **Type I** ("prefix" error) ***Is** the boy who is watching Mickey Mouse is happy?
  - **Type II** ("restarting error") ***Is** the boy who is watching Mickey Mouse, **is he happy**?
  - **Type III** ("structure independent error") ***Is** the boy who **watching** Mickey Mouse is happy?

	Type I	Type II	Type III	Total
Group I (81)	30 (60%)	10 (20%)	0	50 (62%)
Group II (87)	9 (53%)	5 (29%)	0	17 (20%)
Total 168	39 (58%)	15 (22%)	0	67 (40%)

The Poverty of Stimulus argument

3. There are **“trigger” data (?)**

Pullum and Scholz (2002), Legate & Yang (2002)

- 1% of relevant cases in a **typical corpus**.
  (Pullum and Scholz 2002: 45)
- 10 million words of language use -> about 7,500 questions that crucially falsify the structure-independent auxiliary-fronting generalization, before reaching the age of 3.
- Legate & Yang (2002) suggests that even though the primary data are present, this **is not a sufficient condition** to trigger acquisition:
  - For *pro-drop*, they found 1.2% evidence (*there* sentences) in the **primary linguistic input**
  - **No complex** *yes-no question* found in Nina corpus, for instance.

From child to machine learning

Tenenbaum et al (2011)

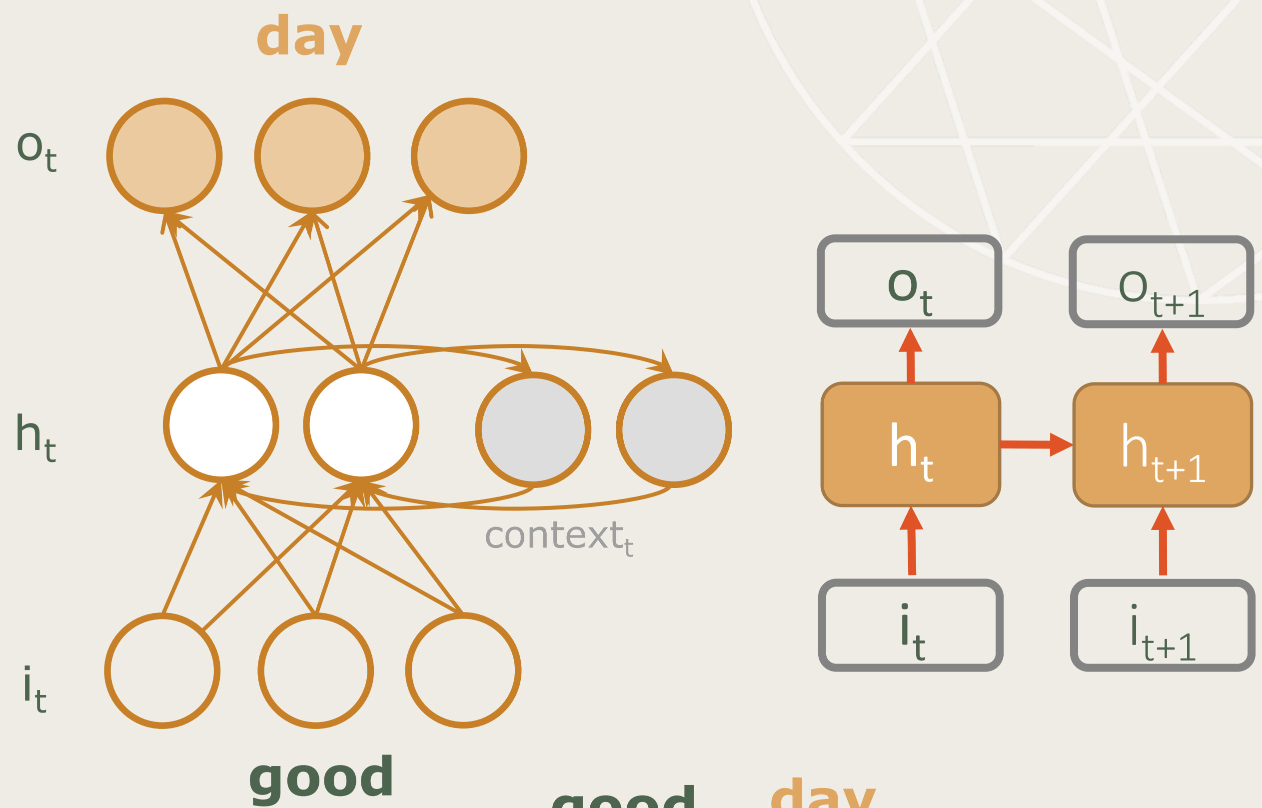
- Constructivism, theory theory… Inductive biases, Bayesian approaches, Hidden Markov models, (multiple) regression, connectionism, error-driven vs. (?) Hebbian learning…
- Let's just consider the "connectionist" metaphor (**Neural Networks**) and the **error-driven**, cross-entropy loss minimization (or simply **loss**)

$$\mathbf{H}(p,q) = -\sum_x p(x) \log q(x)$$

Simple (recurrent) Artificial Neural Networks

Elman (1990)

- Simple Recurrent Neural Networks (RNN)
  - This is a **good** … **day**

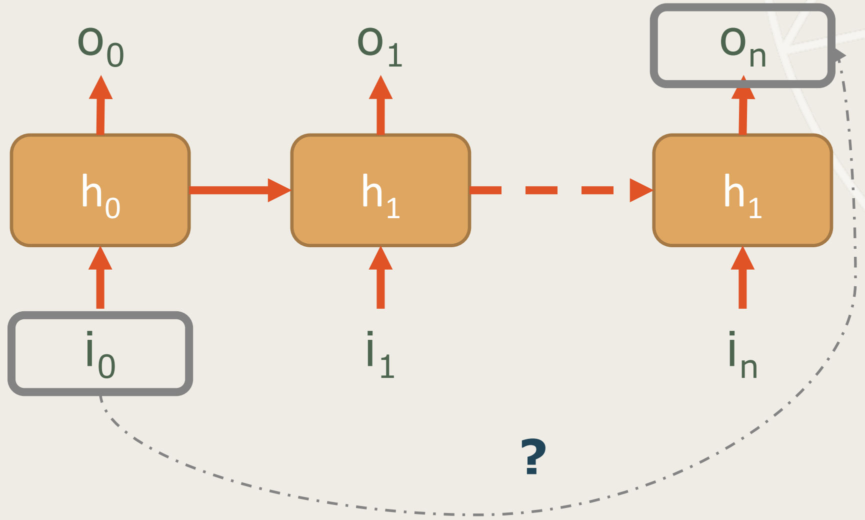


  - [00001, 00010, 00100, **01000**, **10000**]

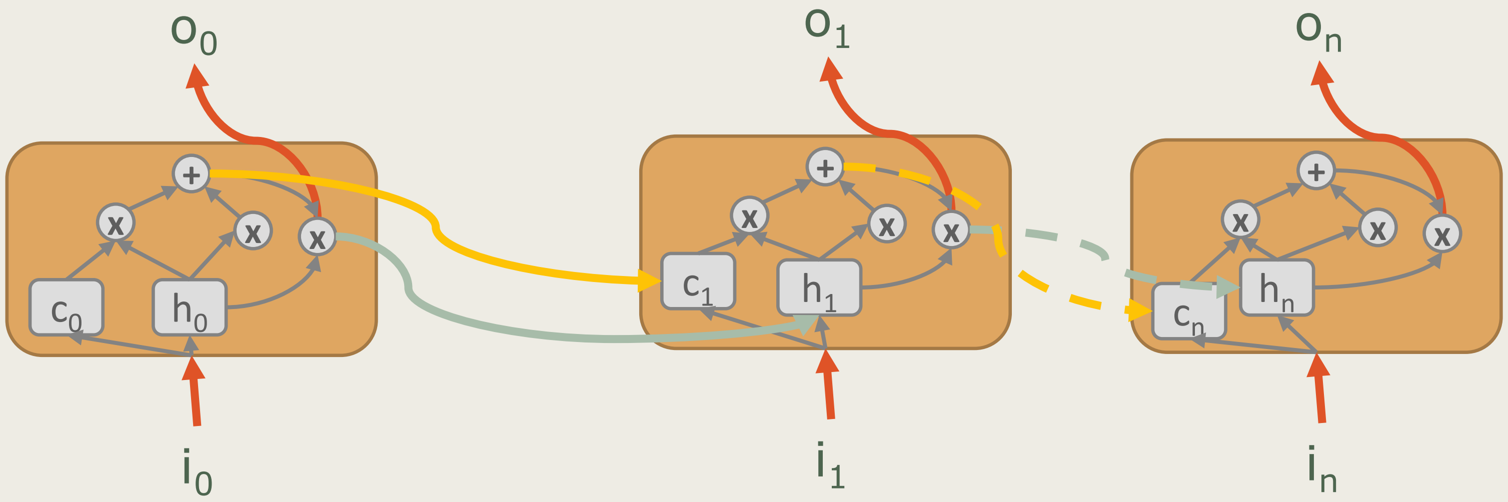
Long Short Term Memory (LSTM) networks

(Hochreiter & Schmidhuber 1997)

- Standard RNN

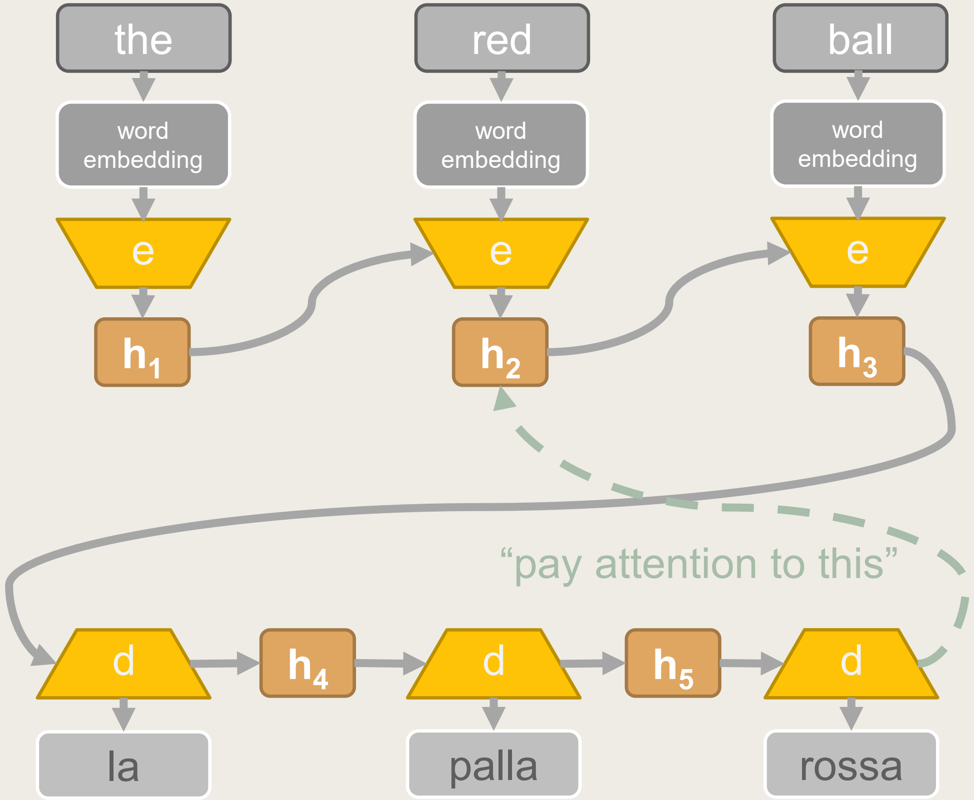


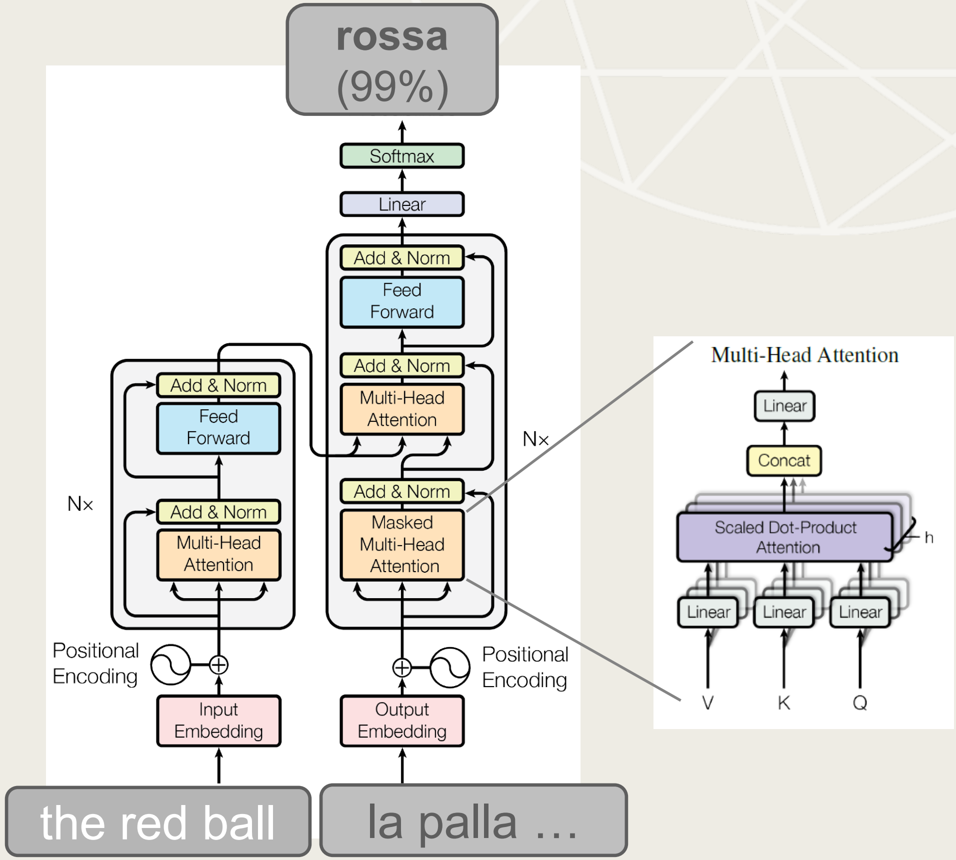
- Revisiting LSTM:



The age of **Transformers**: "Attention is all you need" Vaswani et al. (2017)

Seq to Seq Machine Translation example





The (Self-)Attention Mechanism

Vaswani et al. (2017)

Word embeddings

time	flies	like	an	arrow
0.314	0.187	0.872	0.172	0.873
0.971	0.896	0.493	0.498	0.120
0.126	0.061	0.953	0.277	0.187
0.743	0.167	0.815	0.175	0.167
…	…	…	…	…
0.522	0.011	0.487	0.470	0.778

d

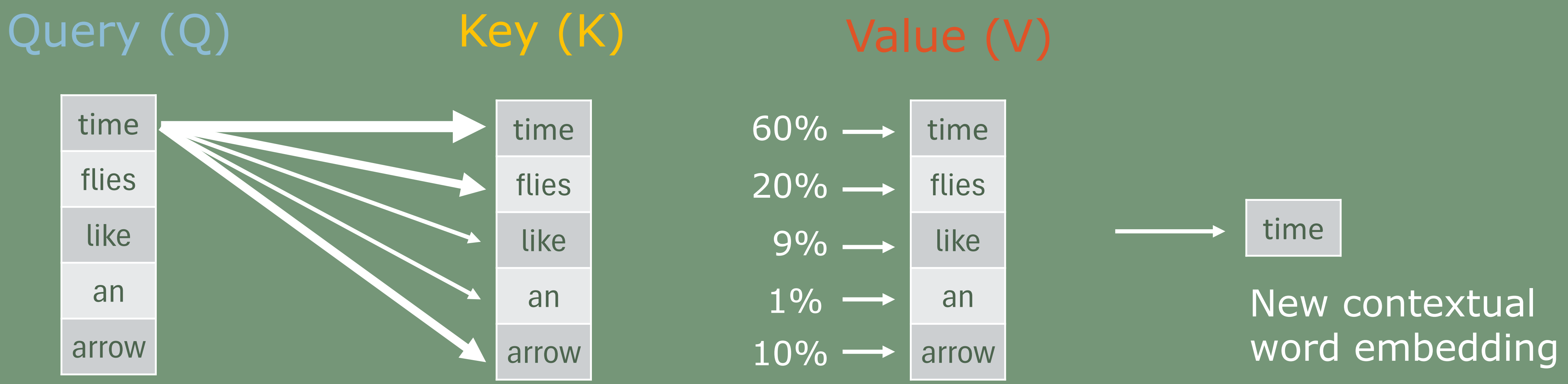
Self-attention

	time	flies	like	an	arrow
time	1	0.876	0.123	0	0.571
flies	0.876	1	0.493	0.1	0.011
like	0.123	0.493	1	0.1	0.487
an	0	0.1	0.1	1	0.230
arrow	0.571	0.011	0.487	0.230	1

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The (Self-)Attention Mechanism

Vaswani et al. (2017)



$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The (Self-)Attention Mechanism

Vaswani et al. (2017)

time

$\text{Query}_{\text{time}}$ Key_{time} $\text{Value}_{\text{time}}$

flies

$\text{Query}_{\text{flies}}$ $\text{Key}_{\text{flies}}$ $\text{Value}_{\text{flies}}$

$$\left(\text{Query}_{\text{time}} \bullet \text{Key}_{\text{flies}} \right) \longrightarrow Q_{\text{time}} K_{\text{flies}}$$

Scalar Value
(Score)

$$Q_1K_1 + Q_2K_2 + \ldots + Q_nK_n$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The (Self-)Attention Mechanism

Vaswani et al. (2017)

time

$\text{Query}_{\text{time}}$ Key_{time} $\text{Value}_{\text{time}}$

flies

$\text{Query}_{\text{flies}}$ $\text{Key}_{\text{flies}}$ $\text{Value}_{\text{flies}}$

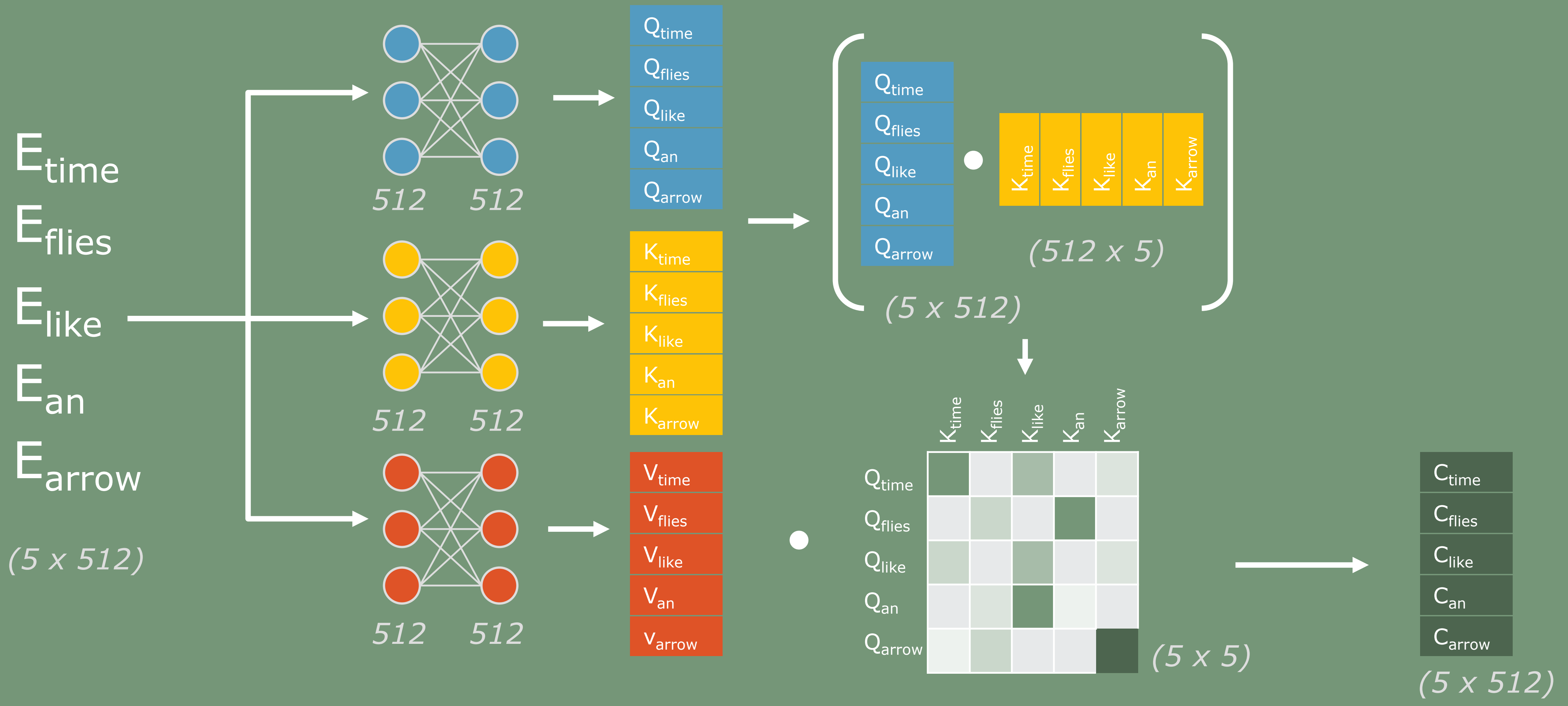
$$\left(\begin{matrix} \text{Query}_{\text{time}} \\ \text{Query}_{\text{flies}} \end{matrix} \bullet \begin{matrix} \text{Key}_{\text{time}} & \text{Key}_{\text{flies}} \end{matrix} \right) \rightarrow \begin{matrix} Q_{\text{time}} K_{\text{time}} & Q_{\text{time}} K_{\text{flies}} \\ Q_{\text{flies}} K_{\text{time}} & Q_{\text{flies}} K_{\text{flies}} \end{matrix}$$

Attention Score

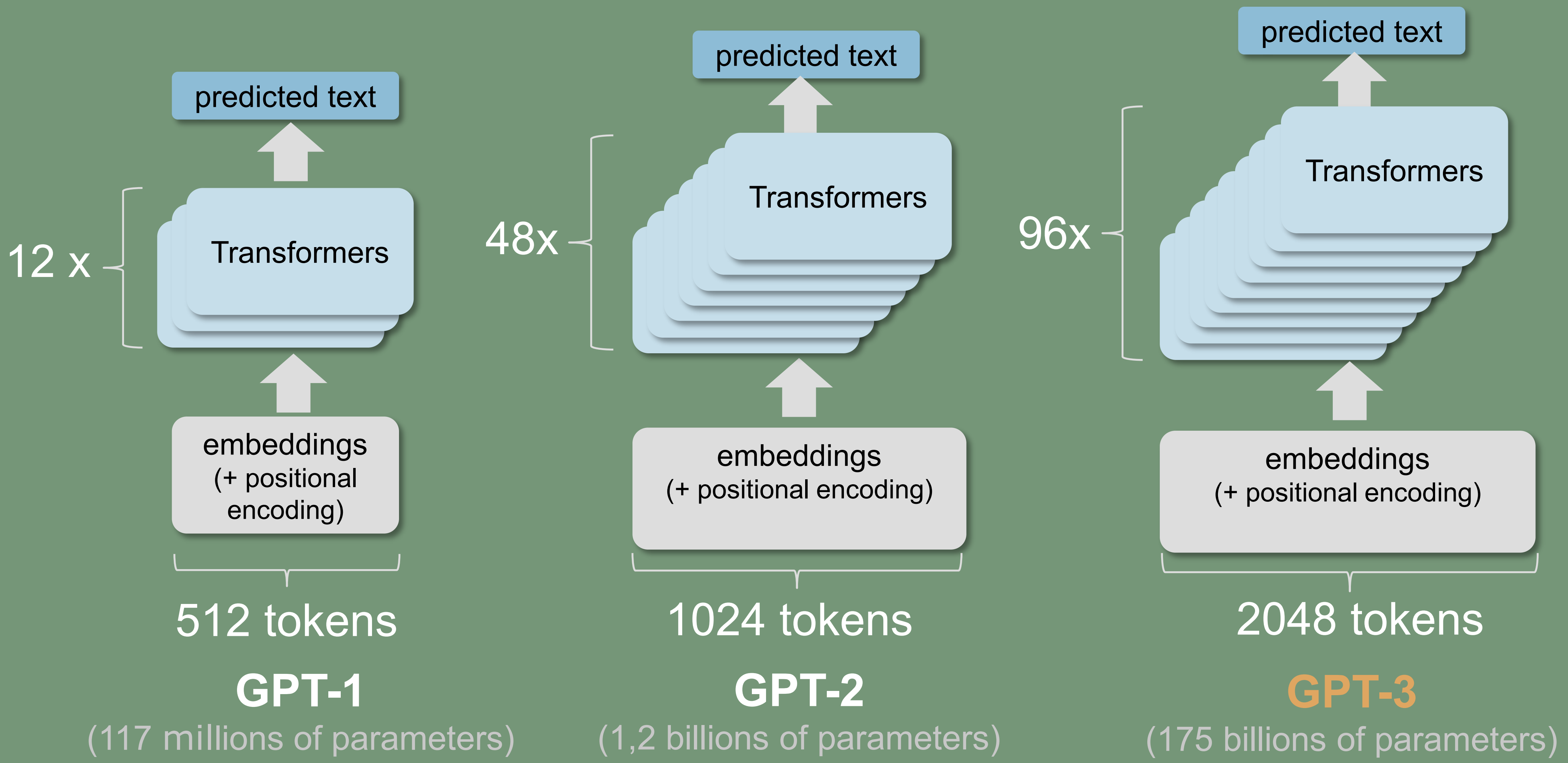
$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The (Self-)Attention Mechanism

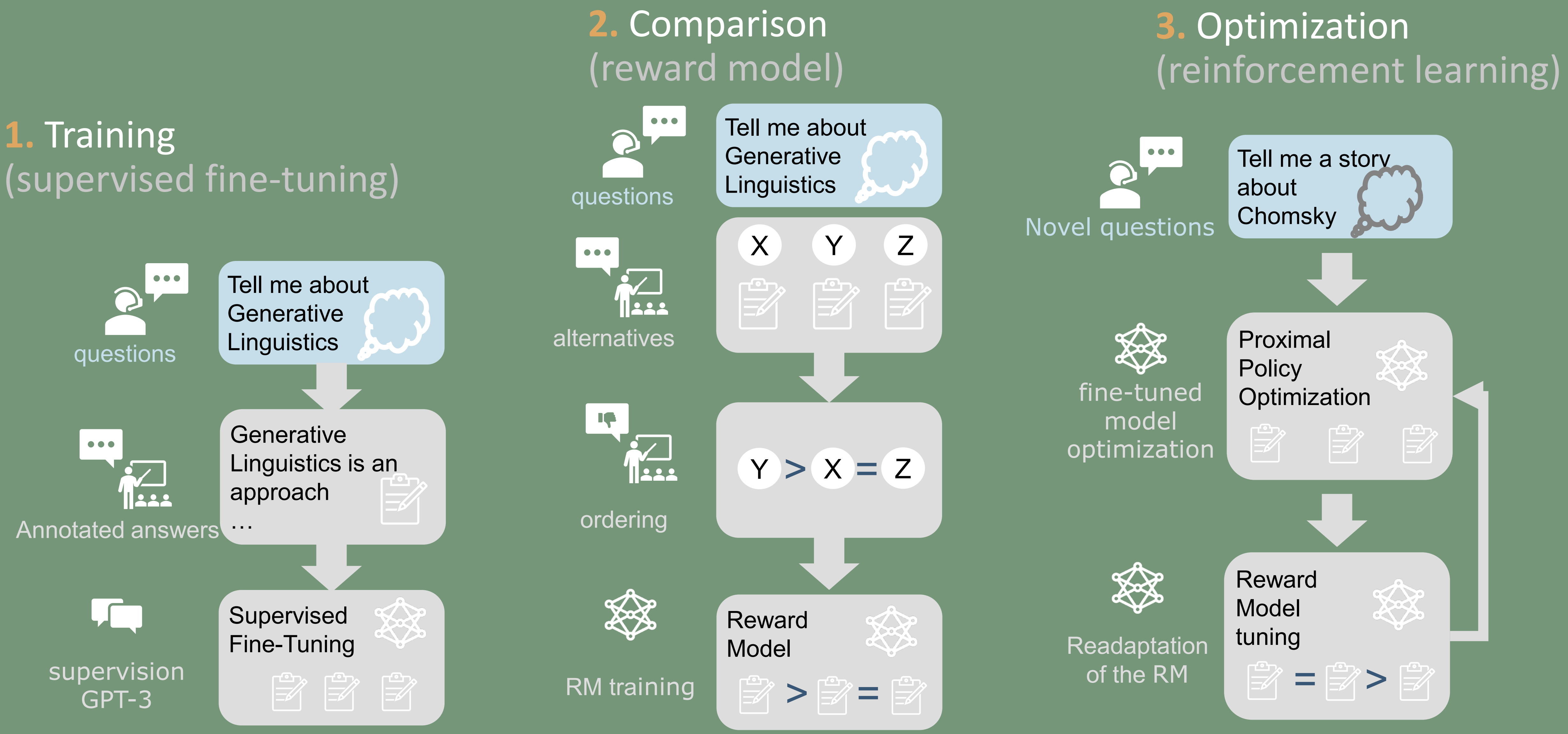
Vaswani et al. (2017)



What’s inside GPT-3



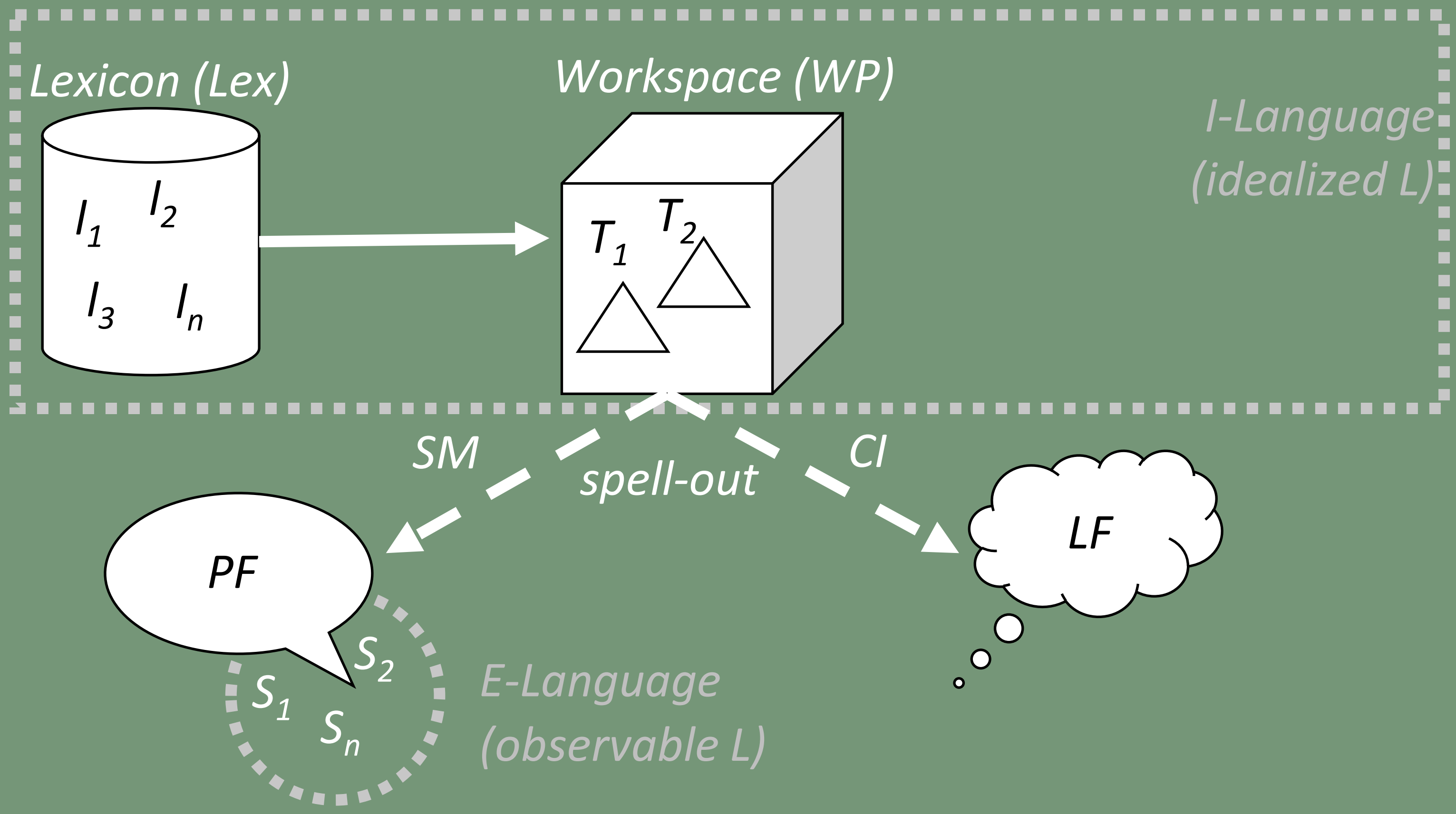
How does it work ChatGPT



Linguistic detour: Minimalist Grammar (MG)

Chomsky et al. 2023 (re-adapted in Chesi 2025)

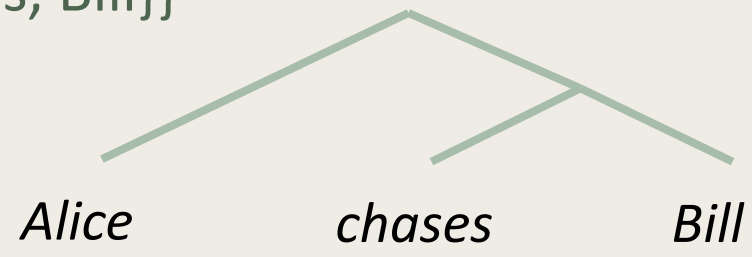
T-Model



Minimalist Grammar (MG)

- A Minimalist Grammar (MG) defines an infinite set of derivations (sequences of steps, D_S) obtained through the applications of essentially one simple structure building operation (Merge) over lexical items (l_i) selected from the language lexicon (Lex_L),

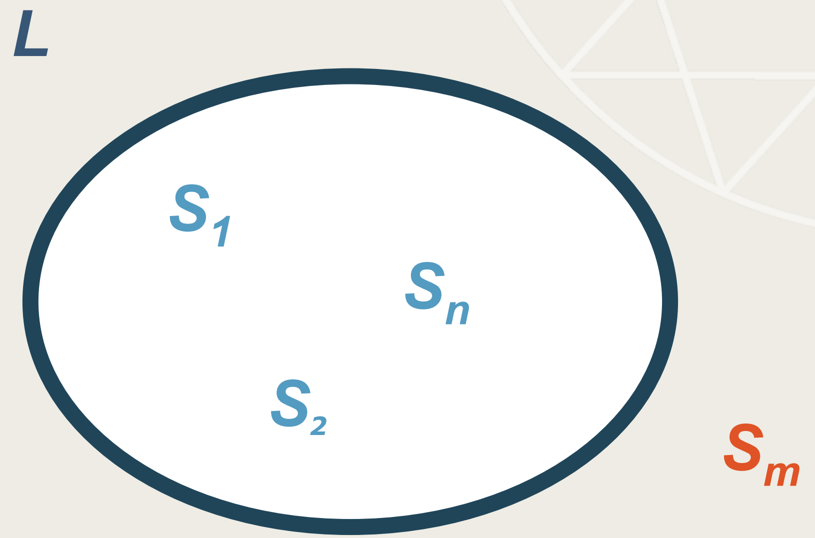
- Derivation (Ds) of the sentence S "Alice chases Bill":
- Select (Alice, Bill, chases) where {Alice, Bill, chases} $\in$ $Lex_{English}$
- Merge (chases, Bill) = {chases, Bill}
- Merge ({chases, Bill}, Alice) = {Alice, {chases, Bill}}



A «genuine» linguistic theory

- **Language Problem**
  (Observational adequacy)

  Is theory X capable of generating and recognizing all and only the sentences Ss belonging to language L?



S_n = John runs

S_m = *John run

Sentences: linear order, hierarchy and grammaticality

(1) a. ***The author*** that the senators hurt ***is*** good
b. ****The author*** that the senators hurt ***are*** good

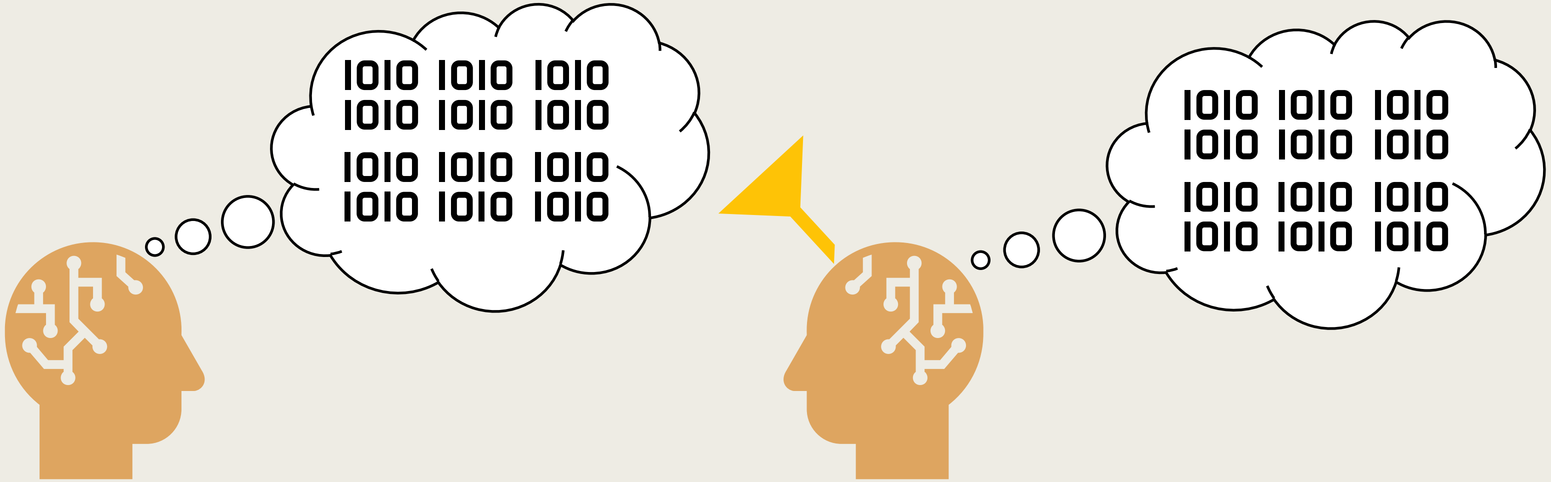
(2) a. ***The author that*** the senators hurt is good
b. ****The author*** the ***that*** senators hurt is good

(3) a. I know ***what***$_i$ the guy broke _$_i$ accidentally and the mechanic fixed _$_i$ skilfully.
b. *I know ***what***$_i$ the guy broke _$_i$ accidentally and the mechanic fixed ***the engine*** skilfully.

The Telepathy Paradox

Chesi 2025, *Linearization (as part of core syntax)*

- If we could use **telepathy**, this would be useless in terms of instantaneous message transmission because of the **finiteness** of our **processing device**



- We hypothesize a **recursive model** because we need to make an infinite use of finite means
- Since we have **finite means**, we can process **finite tokens** at time ***t***. We conclude that "**linearization**" (or **incrementality**) is a **virtual conceptual necessity**.

BabyLM Challenge

Warstadt, Choshen, Mueller, Williams, Wilcox & Zhuang (2023) Call for Papers–The BabyLM Challenge: Sample efficient pretraining on a developmentally plausible corpus. arXiv preprint arXiv:2301.11796.

- **Shared task** intended for participants with an interest in **small scale language modeling**, **human language acquisition**, **low-resource NLP**, and **cognitive modeling**. We provide a platform for approaches to pretraining with a limited-size corpus sourced from data inspired by the input to children.
- Three tracks: two restrict the training data to **pre-released datasets of 10M and 100M words** and are dedicated to explorations of approaches such as **architectural variations**, **self-supervised objectives**, or **curriculum learning**. The final track only restricts the amount of text used, allowing innovation in the **choice of the data**, **its domain**, and even its **modality** (i.e., data from sources other than text is welcome).

The standard (rigid) piepline

Corpus cleaning → Tokenization → Model Training → Evaluation

Tokenization	Model Training	Evaluation
(Fast)**WordPiece** or **BPE**	**Transformers** (ELC-BERT)	**lm_eval**

Corpus

Cleaning

- Italian (~3M tokens) & English (~10M tokens) corpora
  - Child-directed speech in CHILDES Italian section:

*CHI: si. %mor: intj|sì. %gra: 1|1|ROOT 2|1|PUNCT *DON: senti (.) di che colore la vuoi?	senti [PAUSE] di che colore la vuoi ?

  - Songs:

| Title: | Edition 21° Zecchino d'Oro| year 1978 | Salta di qua - rimbalza di là…	salta di qua rimbalza di là

  - Subtitles:

00:02:09,440 --> 00:02:11,440 Mi sono perso! Dov'ÃÂÃÂÃÂÃÂ¨ la fila?	mi sono perso ! dov' è la fila ?

  - Conversations:

A: pronto? B: buonasera potrei parlare con Gianluigi per favore?	pronto ? buonasera potrei parlare con gianluigi per favore ?

  - Fairy Tales:

rispose Babà Mustafà (poiché era proprio lui)	rispose babà mustafà , poiché era proprio lui .

Corpora info

- Italian (~3M words)

Section	Before	After
	Tokens (TTR)	
CHILDES	405,892 (0.05)	346,155 (0.03)
SUBTITLES	959,026 (0.07)	700,729 (0.05)
CONVERSATIONS	80,826 (0.13)	58,039 (0.11)
SONGS	240,309 (0.11)	222,572 (0.08)
FAIRY TALES	1,103,543 (0.10)	1,287,826 (0.05)
ALL	**2,973,879 (0.08)**	**2,431,038 (0.03)**

- Sentences (# of lines): 370,484
- Word per sentence: 7
- 85% of sentences captured with length (min=0 max=1228): 52

- English (~10M words)

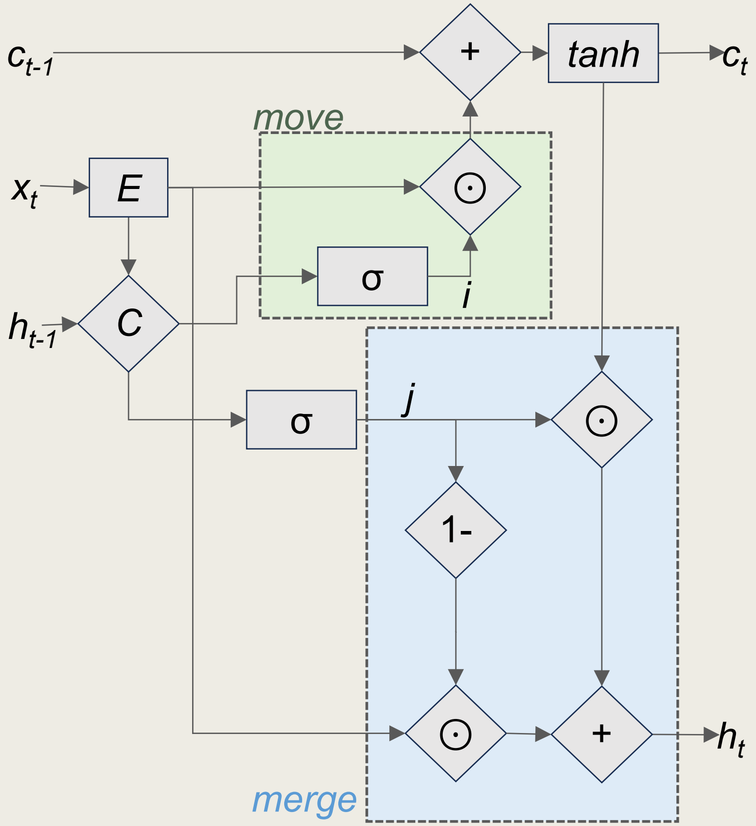
Section	Before	After
	Tokens (TTR)	
CHILDES	1,920,655 (0.02)	1,913,959 (0.01)
SUBTITLES	2,041,868 (0.06)	2,399,780 (0.02)
CONVERSATIONS	1,079,286 (0.04)	1,211,618 (0.02)
GUTENBERG	2,539,489 (0.05)	2,895,199 (0.02)
WIKIPEDIA	1,453,539 (0.09)	1,546,763 (0.05)
ALL	**9,034,837 (0.04)**	**9,967,319 (0.01)**

- Sentences (# of lines): 1,096,918
- Word per sentence: 9
- 85% of sentences captured with length (min=0 max=10,052): 74

Model Architecture

Two ways to forget

- Two pathways, one for non-local dependencies (**move**), the other for embedding (**merge**)



i. … **who** do you think …
retain **who** in memory (c)

ii. … do you think John **appreciate** …
remerge **who** with «**appreciate**» and forget about it

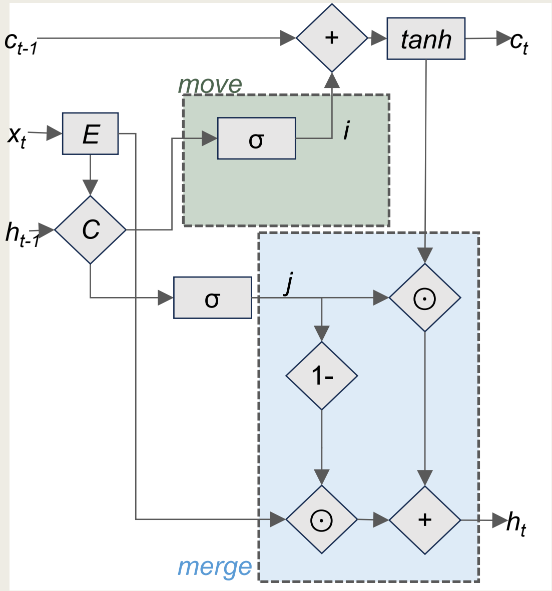
i. … the friend **of** John …
merge **of** with friend as [the friend [**of** …

ii. … the friend of John **is**…
merge **is** with [the friend …] **is**

- 1 layer (650 hidden units),
~60,000 vocab (~ 60M parameters)

Model Architecture: Two experiments

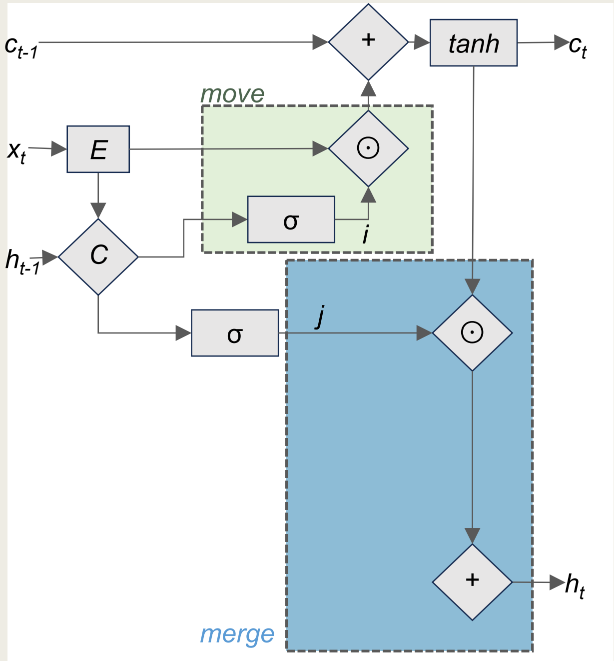
Forget Move



move$_t$= $\sigma(W_{xi}x_t \frown W_{hi}h_{t-1})$

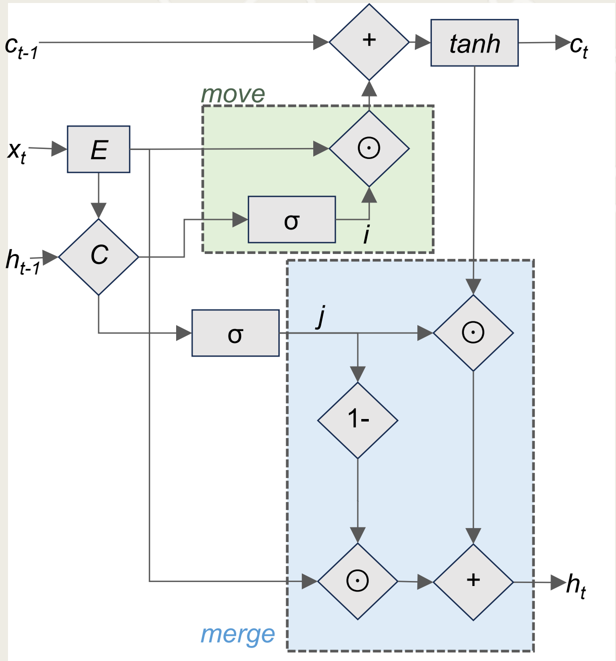
instead of
move$_t$= $\sigma(W_{xi}x_t \frown W_{hi}h_{t-1}) \odot W_{ii}x_t$

Forget Nesting



h$_{t+1}$= $merge_t \odot c_{t+1}$

instead of
h$_{t+1}$= $(1\text{-}merge_t) \odot W_{xi}x_t + merge_t \odot c_{t+1}$



move$_t$= $\sigma(W_{xi}x_t \frown W_{hi}h_{t-1}) \odot W_{ii}x_t$
merge$_t$= $\sigma(W_{xj}x_t \frown W_{hj}h_{t-1})$
c$_{t+1}$= $\tanh(c_t + move_t)$
h$_{t+1}$= $(1\text{-}merge_t) \odot W_{xi}x_t + merge_t \odot c_{t+1}$

Regimen

- Three data batching strategies
  - **Naturalistic** (~10M tokens exposure)
    - *[guarda un po' ?]*
      [ci sono qui le formiche ?]
      [eh !]
      [vieni , vieni a sfogliare qui .] ...
  - **Conversational** (~20M tokens exposure)
    - *[guarda un po' ? ci sono qui le formiche ?]*
      [ci sono qui le formiche ? eh !]
      [eh ! vieni , vieni a sfogliare qui .]
  - **Redundant** (~740M tokens exposure, length=74)
    - *[guarda un po' ? ci sono qui le formiche ? ... basta]*
      [un po' ? ci sono qui le formiche ? ... basta andare]

Evaluation

- LM-eval
  - BLiMP test (English)
    Who is Mary irritating _ after approaching Kenneth? Vs.
    **Who is Mary irritating Kenneth after approaching _ ?*
  - *COnVERSA test (Italian – BLiMP-IT)*
    ***Il** muro della casa è rosso.* Vs.
    ****Il** muro della casa è rossa*
- Probability output:
  - Rough sum: $\sum_{i=0}^{n} -\log p(x_i)$
  - Minimum probability: Max $(-\log p(x_i))$
  - **Normalized**: $\frac{\sum_{i=0}^{n} -\log p(x_i)}{n}$

Results

Training

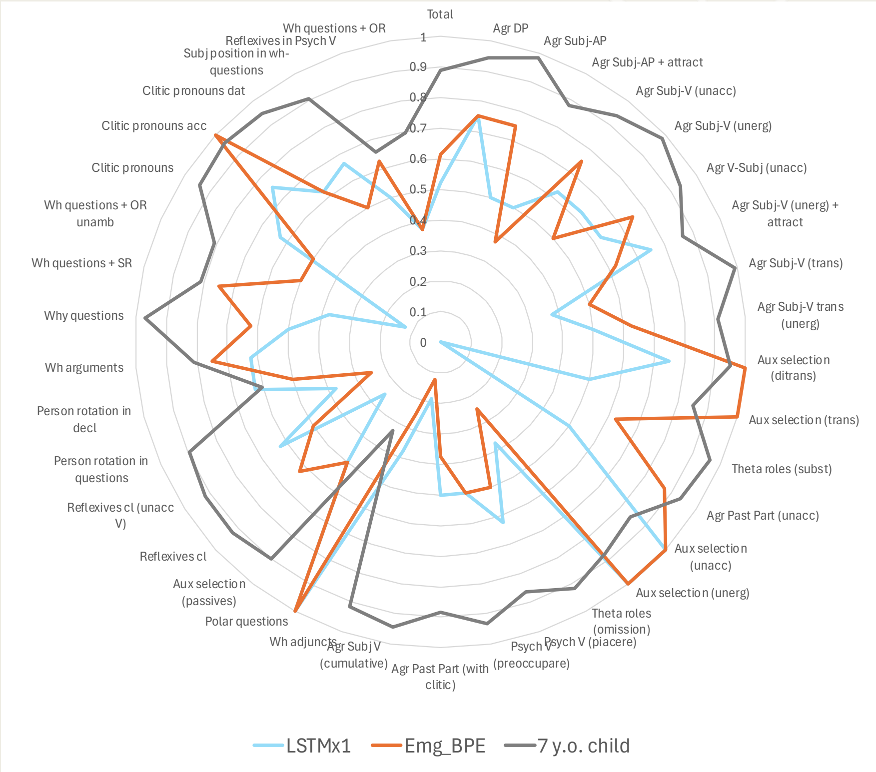
- **Regimen** (*English and Italian are very similar*)
  - Learning plateau around 10-12 epochs for LSTM-eMG (20 for transformers-based architectures)
  - **Naturalistic**: Loss: 2.0211, Accuracy: 0.9064
  - **Conversational**: Loss: 2.5796, Accuracy: 0.8053
  - **Redundant**: Loss: 3.3532, Accuracy: 0.5432

Results - Tests

- BLiMP test (English)
  - Transformers perform randomly
  - Only "redundant" regimen produces non-random tests' performance

	LSTM	eMG-RNN				
	LSTM	1	2	3	F-M	F-N
Ana. agr	0.67	0.82	0.76	0.77	0.88	0.81
Arg. str	0.56	0.65	0.64	0.63	0.64	0.66
Binding	0.54	0.69	0.66	0.63	0.57	0.65
Ctrl. / Rais.	0.59	0.58	0.59	0.60	0.58	0.60
D-N agr	0.57	0.67	0.63	0.67	0.68	0.68
Ellipsis	0.41	0.24	0.30	0.21	0.42	0.39
Filler. gap	0.55	0.64	0.60	0.47	0.48	0.65
Irregular	0.54	0.58	0.69	0.60	0.60	0.58
Island	0.54	0.58	0.54	0.53	0.50	0.62
Npi	0.45	0.33	0.50	0.55	0.32	0.31
Quantifiers	0.57	0.55	0.53	0.53	0.53	0.57
S-V agr	0.50	0.52	0.52	0.52	0.55	0.53
Overall	**0.54**	**0.58**	**0.58**	**0.57**	**0.55**	**0.59**

- COnVERSA test (Italian – BLiMP-IT)
  - Best results after 2-3 epochs



Results - Tests

- *Define a (simple) criterion to interpret these results:*
  - We look at the human performance on BLiMP (~88%, Warstadt et al., 2020)
  - We consider standard deviation (~8%)
  - We assume that the average performance minus 1 or 2 standard deviations (~72-80%) is the threshold for a significant bias
    (**positive**, > 72% or **negative**, < 28%)
- Best **LSTM model:** 4% linguistic bias
- Best **e-MG-RNN**: 44% linguistic bias

In conclusion

- The **poverty of stimulus hypothesis** remains **unchallenged**: none of our trained model equals human performance on none of the tasks (in both languages)
- Linguistically inspired architectural biases significantly improve models' performance in all tasks
- The training regimen significantly impacts on assessment: naturalistic and conversational regimen work well for next-word prediction task, but correlate with random performance at linguistic tasks
- Increasing training time (number of epochs) improves autoregressive training performance, but produces a lower linguistic return

CRISTIANO CHESI
NETS, IUSS LABORATORY
FOR NEUROLINGUISTICS,
COMPUTATIONAL LINGUISTICS
AND THEORETICAL SYNTAX

Thanks

Introduction to Linguistic Computation & Complexity Theory

Ph.D. in Theoretical and Experimental Linguistics (TEL)

(for the "exam": write a **two pages abstract**, including references, discussing a topic of your interest related to what we presented during this mini-course)