19th Workshop on Multiword Expressions (MWE 2023)
Colocated with: EACL 2023 (Dubrovnik, Croatia)

Are Frequent Phrases Directly Retrieved like Idioms?

An Investigation with Self-paced Reading and Language Models

Giulia Rambelli, Emmanuele Chersoni, Marco S.G. Senaldi, Philippe Blache, Alessandro Lenci

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

McGill
UNIVERSITY

Aix*Marseille
université
Socialement engagée

UNIVERSITÀ
DI PISA

Processing Expressions with Different Degrees of Compositionality

Idioms

Andy stole the thunder.

Frequent expressions

Andy stole the wallet.

Andy stole the trolley.

- Ease in processing
  - facilitation effects in reading (Conklin & Schmitt, 2008; Titone et al., 2019)
  - more positive electric signal in brain activity (Vespignani et al., 2010)
- How are idioms represented in the lexicon?

non-compositional view

(Swinney and Cutler, 1979; Cacciari and Tabossi, 1988, i.a.)

hybrid models

(Libben & Titone, 2008; Titone et al., 2019)

- Ease in processing
  - Lexical boundles(Tremblay et al.,2011)
  - 4-word expressions (Bannard and Matthews, 2008; Arnon and E. V. Clark, 2011)
- How frequent a sequence should be to be stored in the lexicon?

Research Question

Question Do IDIOMS and FREQUENT expression have the same *facilitation effect* in processing?

Experimental Conditions		
1.	idiomatic expressions (ID)	*kick the habit*
2.	compositional and highly frequent expressions (HF)	*kick the ball*
3.	compositional and low frequent expressions (LF)	*kick the sister*

Experiments For direct objects in the 3 conditions, we compare

1. **Reading times** (RTs) collected by Self-Paced Reading (SPR) experiment
2. **Surprisal values** of Neural Language Models (NLMs)

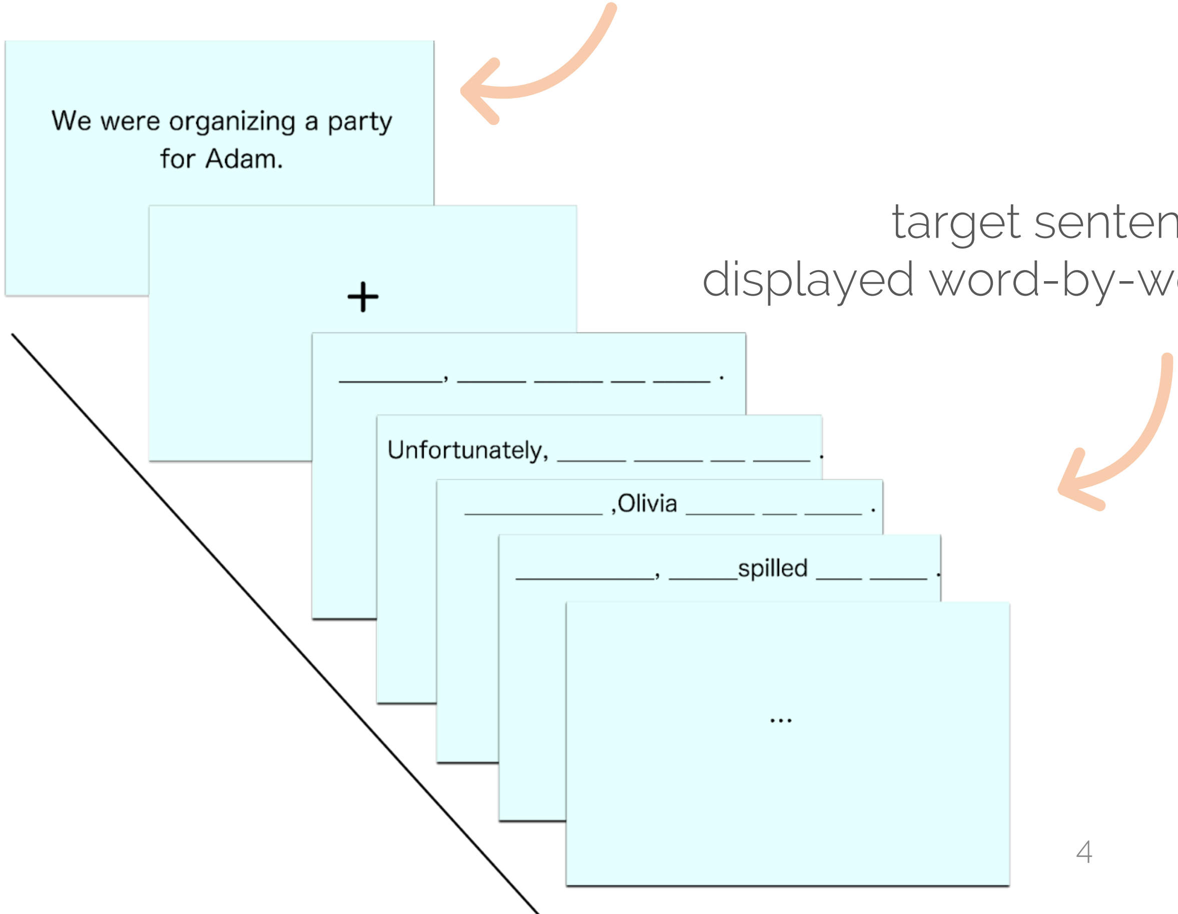
Exp 1: Self-paced Reading (SPR)

Material 48 VERB+det+NOUN idioms and corresponding HF and LF bigrams -> 144 stimuli

Method Moving-window SPR paradigm

Participants 90 L1 English speakers from North America (M=29.6 ± 7.55). Delivered remotely.

Hypothesis RT(ID) < RT(HF) < RT(LF)



context sentence

We were organizing a party for Adam.

target sentence displayed word-by-word

+
______, ____ ____ __ ____ .
Unfortunately, ____ ____ __ ____ .
________ ,Olivia ____ __ ____ .
________, ____spilled ___ ____ .
...

Exp 1: SPR Results

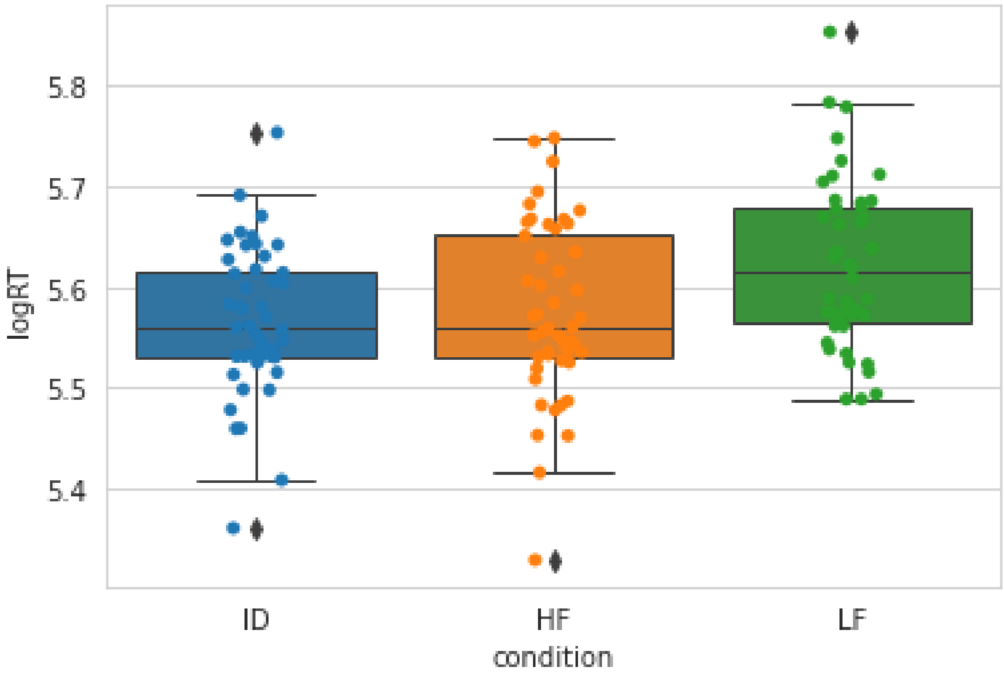
Result Participants responded similarly to idioms and frequent phrases but more slowly to the unfrequent expressions.

There are facilitation effects in the comprehension of both figurative meaning of idioms and the compositional one of HF.

Explanations

1. same mechanism
2. facilitation effects are similar but depend on different mechanisms

M: 262.17 ms M: 262.69 ms M: 270.42 ms



Exp 2: Modeling RTs with NLMs

Material The same 144 stimuli sentences

Architectures

- autoregressive models -> GPT2 (small, medium, large, xl)
- bidirectional models -> BERT-base-case and T5-base
- recurrent neural networks (RNN) -> tinyLSTM(Stephen et al., 2017); GRNN (Gulordava et al., 2018)

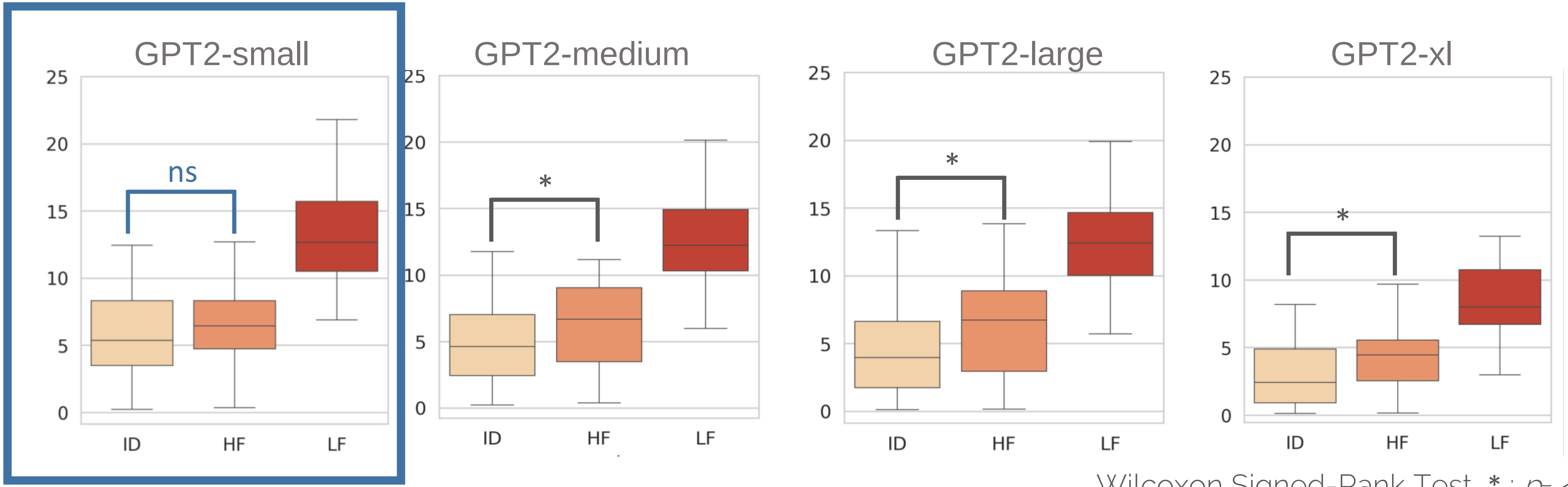
Method Measure the Surprisal(Hale, 2001; Levy, 2008) of a word

$$\boxed{Surprisal(w_i) = -l\ og\ P(w_i|context)}$$

$$context \begin{cases} w_{0,1..i-1} \text{ for unidirectional LM} \\ w_{0,1,..i-1,i+1,..n} \text{ for bidirectional LM} \end{cases}$$

Hypothesis The Surprisal values are distributed in the same way of huma reading times (RTs)

Exp 2: GPT2 Surprisals



GPT2-small

GPT2-medium

GPT2-large

GPT2-xl

Wilcoxon Signed-Rank Test, * : *p*= < .05

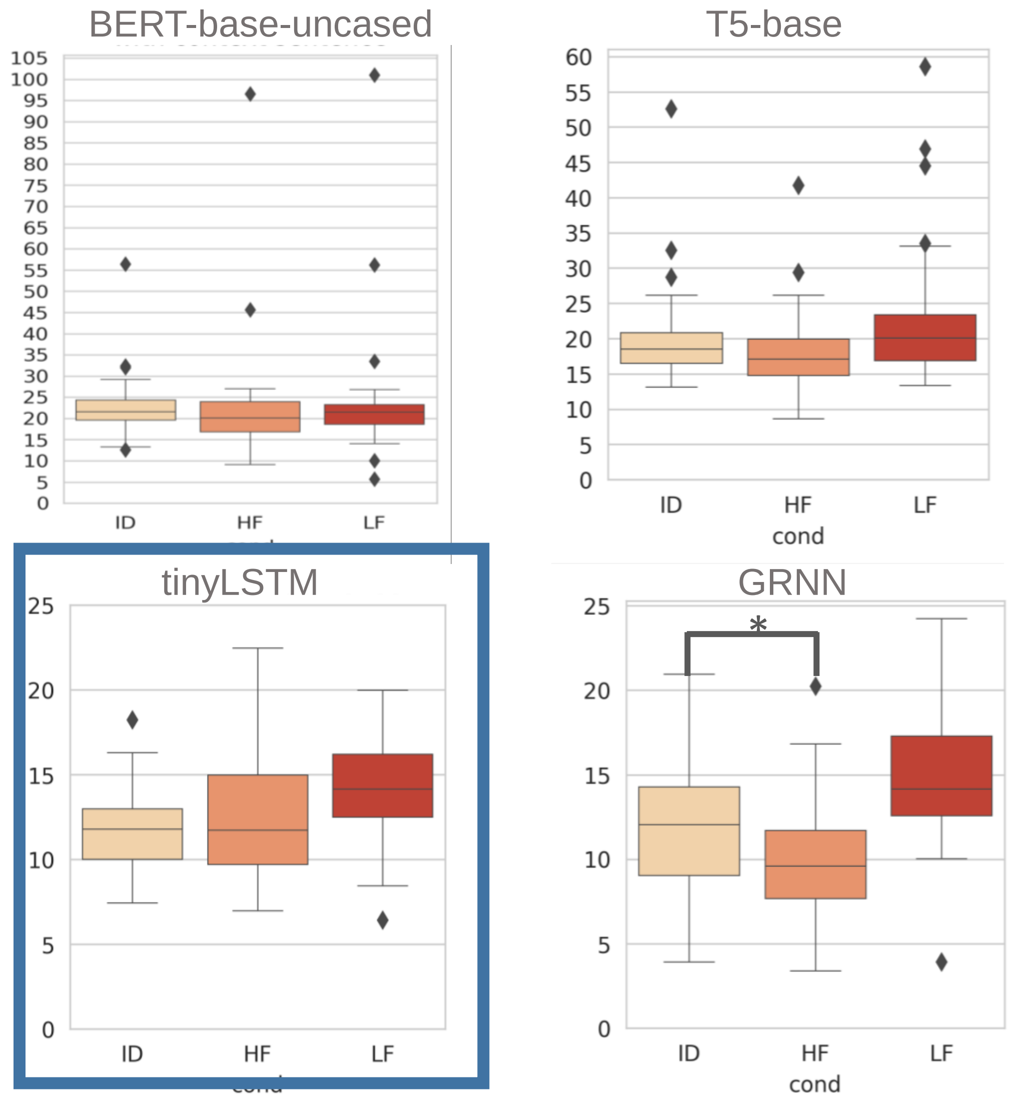
Results

1. All the GPT2 models produce Surprisal(ID) <Surprisal(HF)…
   … with the exception of GPT2-small

Exp 2: BERT, T5 & RNNs Surprisals

Results (continue)

2. BERT and T5 show a Surprisal(HF) < Surprisal(ID)
3. GRNN is similar to T5
4. Only tinyLSTM is comparable to human RTs



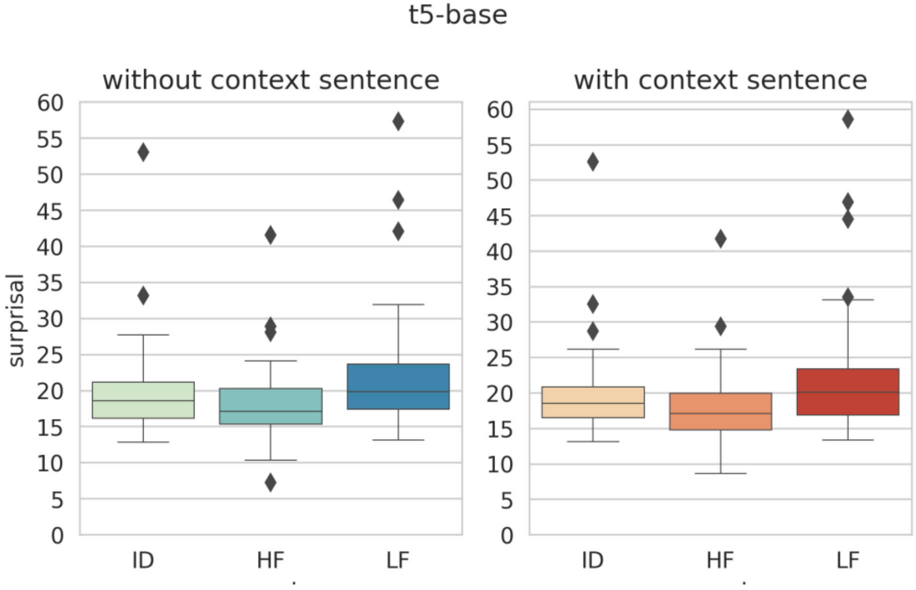
Exp 2: The Role of Context

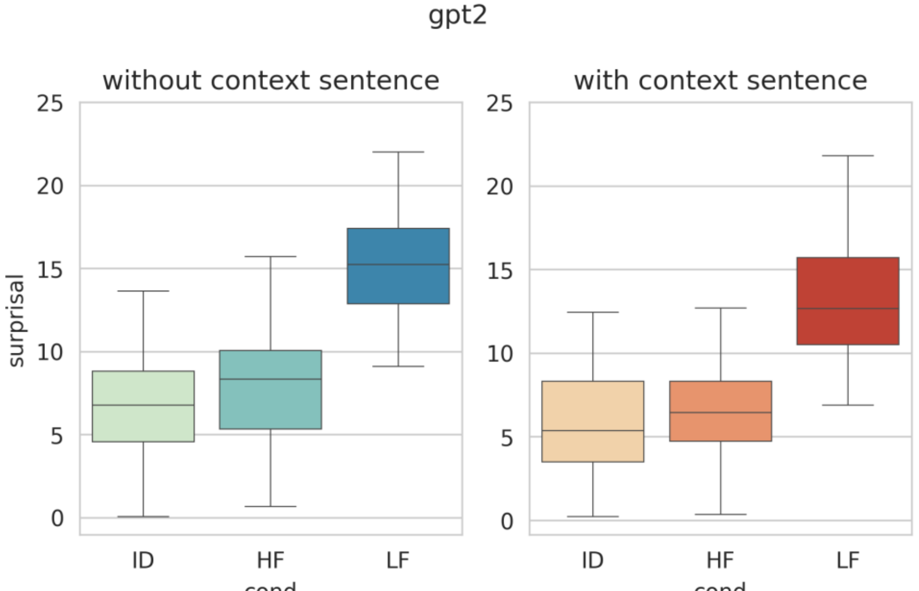
Question Are NLMs sensible to context?

Method: fed NLMs only with the target sentence

Results:

- RNN and bidirectional models produce the same Surprisal with or without the context sentence.
- GPT2 models have lower Surprisal scores giving a context sentence.





Contributions

- People read idioms and frequent compositional units at comparable speed
    - How are represented in the mental lexicon?
- Both idiomatic and frequent expressions are highly expected by GPT2 models, not by bidirectional models
    - GPT2-small has comparable to RTs -> *inverse scaling* effect (Oh and Schuler, 2022)
- Context seems to affect little or not at all the Surprisal scores

✉ giulia.rambelli4@unibo.it

g_rambelli

Find more info in our paper
or get in touch at the MWE 2023
on May 6th!

CODE and MATERIALS

https://qrco.de/bdsxvp