What Mechanisms Underlie Linguistic Generalization In Large Language Models?

A Study On Noun-noun Compounds

Giulia Rambelli

21.02.2024

VU, Amsterdam

Funded by the European Union (GRANT AGREEMENT: ERC-2021-STG-101039777). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

RESEARCH INTEREST

- What is the role of analogy in linguistic productivity?
- What mechanisms underlie linguistic generalization in large language models (LLMs)?
- Are linguistic generalizations in LLMs the result of analogical processes or compositional ones?

CONCEPTUAL COMBINATION

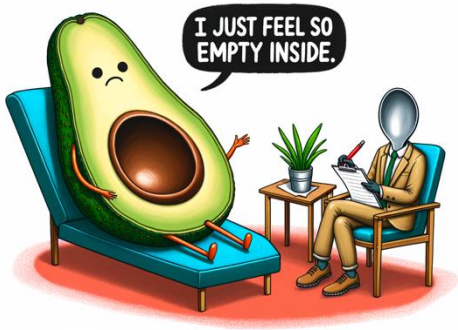
Two lexical concepts are often used together as phrases to represent a combined concept of greater specificity.

apple pie

LEXICALIZED COMPOUNDS

Combining words is a hallmark of language generativity, or productivity.

NOVEL COMPOUNDS



avocado chair

Dunbar & Myers (1988). Concept combination and the characterization of lexical concepts. In Hüllen, Werner; Schulze, Rainer (eds.). Understanding the lexicon: meaning, sense and world knowledge in lexical semantics, 292–302.

THE ROLE OF ANALOGY

Interpreting a novel compound involves:

- accessing the **concepts** denoted by the words
- selecting a **relation** to form a unified conceptual representation

Hypothesis from Gagné and colleagues

The on-line interpretative processing of novel nominal compounds is affected by **analogous lexicalized compounds**

mud man

milk man
'A man who delivers mud'

garbage man
'A man who collects mud'

Gagné & Shoben (2002). Priming Relations in Ambiguous Noun-noun Combinations. *Memory & Cognition*.
Gagneé & Spalding (2006). Conceptual Combination: Implications for the Men- tal Lexicon. *The Representation and Processing of Compound Words*.

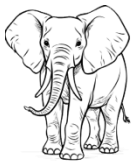
INVESTIGATE CONCEPTUAL COMBINATIONS in LLMS

RQ1: Do LLMs Grasp Semantic Relations in Lexicalized Noun Compounds?

RQ2: Are LLMs able to generalize semantic relations over novel compounds?

Can Large Language Models Interpret Noun-Noun Compouds?
A linguistically motivated study on Lexicalized and Novel Compounds

Rambelli, Collacciani, Chersoni, Bolognesi



ACL 2024
Bangkok, Thailand

METHODOLOGY

Models Llama-2, Mistral, Falcon (7B)(base + instruct)

Task Compound interpretation as multiple-choice task

- LLM has to choose the correct interpretation among 9 paraphrases (to avoid "parroting").

- Surprisal of sentences

  S_{good} = "olive oil is an oil composed of olives"

  S_{bad} = "olive oil is an oil intended for olives"

  $S(S_{good}) < S(S_{bad})$

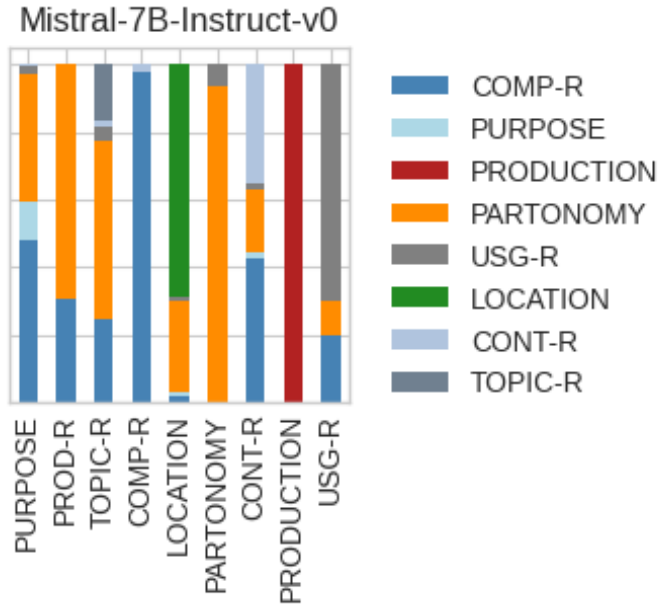
- Metalinguistic prompting

```
Which is the most likely description
of "olive oil"?
1. an oil that uses olives;
2. an oil that is part of olives;
...
9. an oil that is composed of olives
```

EXP.1: INTERPRETING LEXICALIZED NC

Data 668 compositional and lexicalized compounds

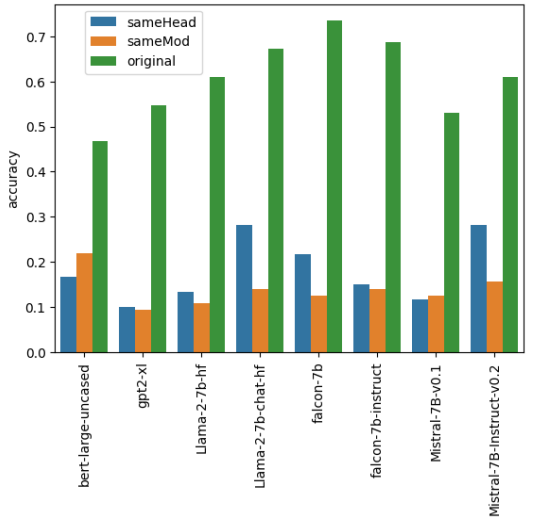
compound	coarse-grained (Tratz, 2011)	fine-grained (Tratz, 2011)	Hatcher-Bourque (Pepper, 2022)	paraphrase (Pepper, 2021)
plastic bag	containment	SUBSTANCE -MATERIAL- INGREDIENT	COMP(OSITION)-R	a bag that is composed of plastic
trash bag	containment	CONTAIN	CONT(AINMENT)-R	a bag that contains trash
supermarket shelf	loc_part_whole	LOCATION	LOCATION	a shelf that is located in a supermarket
car door	loc_part_whole	WHOLE+ PART_OR _MEMBER_OF	PARTONOMY	a door that is part of a car
food company	purpose	CREATE- PROVIDE- GENERATE- SELL	PRODUCTION	a company that produces food
bank loan	causal	CREATOR- PROVIDER- CAUSE_OF	PROD(UCTION)-R	a loan that a bank produces
research group	purpose	PERFORM& ENGAGE_IN	PURPOSE	a group intended for research
art class	topical	TOPIC	TOPIC-R	a class that is about art
wind turbine	topical	MEAN	US(A)G(E)-R	a turbine that uses wind



	Surprisal	Prompt
Mistral	0.403	0.59
Llama-2-7B	0.448	0.41

- COMP(OSITION)-R and PRODUCTION are almost perfect
  - PURPOSE, PROD-R, and TOPIC-R are mostly mistaken
- Compounds characterized by higher **concreteness** are interpreted more accurately

EXP. 2: INTERPRETING NOVEL NC



Data 64 novel compounds

- Head/modifier substituted with a hypernym from WordNet

EQUIPMENT BOX

GLOVE BOX

GLOVE **COMPARTMENT**

	sameHead	sameMod
Mistral	0.578	0.469
Llama-2-7B	0.156	0.141

- Changing the modifier is less problematic than changing the head
- Suboptimal solution: choose PURPOSE relation
  - *equipment box* -> "a box that contains equipment"
  - *glove container* -> "a container intended for gloves"

NEW: L1 & L2 ENGLISH SPEAKERS IN INTERPRETING NOVEL NC (preliminar results)

Differences between L1 and L2 English speakers (Italian high school students, B2 level)

- Stimuli: 126 compounds (63 lexicalized + 63 novel), balanced across 7 semantic relations
- Task: select the appropriate paraphrase

For novel compounds, English L1 speakers reach higher accuracy than Italians

- Greater familiarity with lexicalized compounds and ability of native speakers to handle the complexity of semantic categories → analogy
- Influence of the native language
  - Italian has fewer nominal compounds than English and often prefers paraphrased or prepositional expressions, which could influence how learners semantically categorize compounds in English
- They prefer *intended to* similar to LLMs
  - *boat trip* → a trip that uses a boat BUT
  - *conveyance trip* → a trip intended for conveyance

Results collected for a Master thesis by Marta Mulazzani
"Interpretazione dei composti nominali in inglese: confronto tra apprendenti italiani L2 e nativi inglesi"

FUTURE WORKS

- There are still questions unanswered regarding how people and LLMs interpret compounding.
  - **When analogies take place in language comprehension?**
- Future works:
  - We are collecting several norms of lexicalized compounds for more than 2000 compounds following similar for single words compounds
  - We are investingating compound interpretation with L2 learners of English (Italian students)