# Memory limitations and compositionality (phd thesis)

Large language models and transformer language models in general have revolutionized the field of natural language processing.

Although impressive, these models are granted an unrealistic amount of training data and an unrealistic amount of processing ressources that may actually be detrimental to generalization and increase overfitting and rote memorization of training data.

The kind of generalization we care about are structural and compositional generalizations (Kim and Linzen, 2020; Li et al., 2023). We seek to identify whether these models can generalize for structural grammatical rules, that is when being trained on sentences such as:

- Ava saw the ball in the bottle
- Emma saw the dog / the cat ran

and generalize to patterns like:

- Ava saw the ball in the bottle on the table
- Emma saw the cat

**Observations (synthetic generalization)** It has been observed that transformer language models (1) struggle to generalize on the COGS benchmark on the so called structural cases (Kim and Linzen, 2020; Li et al., 2023) and (2) that parsers or models with a strong structural bias have significantly better properties for generalizing (Yao and Koller, 2022).

**Observations (human behavior)** It has also been observed that transformer language models, as they are getting larger and larger, are becoming bad predictors of human reading times in the sense that they are trained on so much data that they underestimate human reading difficulties (Oh and Schuler, 2023).

**The hypothesis** The broad underlying hypothesis under study states that human memory limitations shape linguistic competence to some extent. The literature on sentence processing emphasizes on the importance of memory limitations to explain processing difficulties (Gibson et al., 2019; Futrell, Gibson, and Levy, 2020; Hahn et al., 2022) and the current neural language models do echo some of the traditional models of memory processing (Ryu and Richard L. Lewis, 2021; Timkey and Linzen, 2023) but without their limitations.

We will test (and possibly design in partnership with the ANR COMPO consortium) several models of memory limitations that are known in the literature. These are either the RNN family including their most recent evolutions (Beck et al., 2024). The most iconic memory limited models are certainly traditional and neural markovian language models and there are also some forms of syntactic parsers such as those of the incremental shift reduce family. Among the interesting perspectives are memory limited language models with theoretically inspired limitations functions. These may be, among others, models enforcing decaying activations or interferences (Richard L Lewis, Vasishth, and Van Dyke, 2006; Timkey and Linzen, 2023) or lossy memories inspired by the noisy channel model (Hahn et al., 2022).

It remains to explain why memory limitations should entail better generalization ? For language modeling, we can get insights on this issue from the machine learning theory (Peyrard et al., 2022): the idea is to view next word prediction as a causal task and relying on irrelevant causes, or irrelevant cues in the memory, leads to model overfitting. On the other hand, being able to only select the relevant causes should lead to better generalizations. The task of selecting relevant items in memory is extra statistical and we intend to rely on the above mentioned theoretically motivated inductive biases to drive these choices.

**Method** The overall methodology of the thesis will rely on computational experiments and possibly model design inspired by the psycholinguistic literature. We intend to take advantage of existing datasets of two kinds.

Well established synthetic datasets such as COGS/SLOG (Kim and Linzen, 2020; Li et al., 2023) to measure compositional generalization. On the one hand these datasets offer a framework to carry controlled experiments on language modeling generalization: in particular it is possible to train a language model on a controlled subset of the dataset before testing its generalization properties. On the other hand they suffer various methodological issues (Wu, Manning, and Potts, 2023; Sun, Williams, and Hupkes, 2023). In our case it is not known to us to which extent they will be sensitive enough to models with different memory properties.

Behavioral data sets might show up better sensitivity for testing memory properties: at least psycholinguistic memory models have been essentially designed with reading time data. There exists a set of well established datasets for reading times such as the Dundee corpus (Alan Kennedy and Pynte, 2003), natural stories (Futrell, Gibson, Tily, et al., 2021) or MECO (Kuperman, Schroeder, and Gnetov, 2024). The first key property of natural reading times datasets is that reading times tend to be largely explained by lexical frequency effects: measuring structural effects amounts to seek a needle in a haystack. Measuring structural effects require experimental designs that neutralize frequency effects to a large extent and datasets relying on these designs start to emerge (Huang et al., 2024). By contrast with synthetic datasets, the second key property of reading time datasets is that they require the model to be pretrained on a larger and uncontrolled corpus.

The thesis will take place within the ANR funded COMPO project and it is expected to be realized in close interaction with the partners in the project.

# Bibliography

Alan Kennedy, Robin Hill and Joel Pynte (2003). "The dundee corpus". In: *Proceedings of the 12th European conference on eye movement.*

Beck, Maximilian et al. (2024). *xLSTM: Extended Long Short-Term Memory.*

Futrell, Richard, Edward Gibson, and Roger P Levy (2020). "Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing". In: *Cognitive science* 44.3, e12814.

Futrell, Richard, Edward Gibson, Harry J Tily, et al. (2021). "The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions". In: *Language Resources and Evaluation* 55, pp. 63–77.

Gibson, Edward et al. (2019). "How efficiency shapes human language". In: *Trends in cognitive sciences* 23.5.

Hahn, Michael et al. (2022). "A resource-rational model of human processing of recursive linguistic structure". In: *Proceedings of the National Academy of Sciences* 119.43, e2122602119.

Huang, Kuan-Jung et al. (2024). "Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty". In: *Journal of Memory and Language* 137, p. 104510.

Kim, Najoung and Tal Linzen (2020). "COGS: A Compositional Generalization Challenge Based on Semantic Interpretation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020.* Association for Computational Linguistics.

Kuperman, Victor, Sascha Schroeder, and Daniil Gnetov (2024). "Word length and frequency effects on text reading are highly similar in 12 alphabetic languages". In: *Journal of Memory and Language* 135, p. 104497.

Lewis, Richard L, Shravan Vasishth, and Julie A Van Dyke (2006). "Computational principles of working memory in sentence comprehension". In: *Trends in cognitive sciences* 10.10, pp. 447–454.

Li, Bingzhi et al. (2023). "SLOG: A Structural Generalization Benchmark for Semantic Parsing". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023.* Association for Computational Linguistics.

Oh, Byung-Doh and William Schuler (2023). "Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?" In: *Transactions of the Association for Computational Linguistics* 11.

Peyrard, Maxime et al. (2022). "Invariant Language Modeling". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.* Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Ryu, Soo-Hyun and Richard L. Lewis (2021). “Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention”. In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, CMCL*. Association for Computational Linguistics.

Sun, Kaiser, Adina Williams, and Dieuwke Hupkes (Dec. 2023). “The Validity of Evaluation Results: Assessing Concurrence Across Compositionality Benchmarks”. In: *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. Ed. by Jing Jiang, David Reitter, and Shumin Deng. Singapore: Association for Computational Linguistics.

Timkey, William and Tal Linzen (2023). “A Language Model with Limited Memory Capacity Captures Interference in Human Sentence Processing”. In: *Findings of the Association for Computational Linguistics: EMNLP*.

Wu, Zhengxuan, Christopher D. Manning, and Christopher Potts (2023). “ReCOGS: How Incidental Details of a Logical Form Overshadow an Evaluation of Semantic Interpretation”. In: *Trans. Assoc. Comput. Linguistics* 11, pp. 1719–1733.

Yao, Yuekun and Alexander Koller (2022). “Structural generalization is hard for sequence-to-sequence models”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.