L95 Introduction to Natural Language Syntax and Parsing

Prof Paula Buttery, Dr Fermín Moscoso del Prado Martín, David Strohmaier & Suchir Salhan

Michaelmas 2025

Motivation

The *Introduction to Natural Language Syntax & Parsing* course explores the intersections between linguistic theory and computational modelling. It motivates why linguistics remains central to the study of language technology. [1]

[1] Handout prepared by Suchir Salhan

LLMs are not a silver bullet. Linguistics provides theoretical structure, methodological discipline, and typological diversity that we can use to understand the capabilities and limitations of LLMs, and how we can improve the linguistic capabilities of models. Despite the sophistication of large language models (LLMs), linguistic theory continues to offer indispensable insights for interpretability, evaluation, and understanding of cross-linguistic structure.

Motivating Paper:
Ferrer-i-Cancho, R. & Solé, R. (2024). *Natural Language Processing relies on Linguistics.*
arXiv:2405.05966.

Areas of Study

- **Interpretability:** How do LLMs encode syntactic, morphological structure?
- **Resources:** Linguistically informed datasets and benchmarks to evaluate linguistic capabilities of models.
- **Evaluation:** Designing controlled experiments and minimal pairs in English and beyond.
- **Low-resource languages:** Evaluating cross-linguistic generalisation and typological gaps.
- **Study of language:** Grammar formalisms, morphological systems. Applications to Parsing Tasks in Computational Linguistics and Language Models tokenization.

Session Plan

Mon 12 Oct (Fermín)

Introduction: Natural Language Processing relies on Linguistics. arXiv:2405.05966.

Wed 15 Oct (Fermín)

Morphosyntax.

Mon 20 Oct (Paula)

Constituency, Heads, and Phrase Structure Grammar: Arguments, adjuncts, and long-distance dependencies.

Wed 22 Oct

Probing Syntax in LLMs.

- Hewitt & Manning (2019). *A Structural Probe for Finding Syntax in Word Representations.* (NAACL 2019). `https://aclanthology.org/N19-1419/`
- Hall Maudslay & Cotterell (2021). *Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing.* NAACL 2021. `https://aclanthology.org/2021.naacl-main.11/`
- Finlayson et al. (2021). *Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models.* ACL 2021. `https://aclanthology.org/2021.acl-long.144/`

Mon 27 Oct (Suchir/David)

Introducing Minimal Pair Datasets.

Wed 29 Oct (Suchir/David)

Minimal Pair Evaluation.

- Xiang et al. (2021). *CLiMP: A Benchmark for Chinese Language Model Evaluation.* EACL 2021. `https://aclanthology.org/2021.eacl-main.242/`
- Song et al. (2022). *SLING: Sino Linguistic Evaluation of Large Language Models.* EMNLP 2022. `https://aclanthology.org/2022.emnlp-main.305/`

Mon 3 Nov (Paula)

Grammars I: PCFGs and Dependency Grammar.

Wed 5 Nov

Parsing and Benchmarking.

- *Dynamic Head Selection for Neural Lexicalized Constituency Parsing.* ACL 2025. `https://aclanthology.org/2025.acl-long.786.pdf`

- *Better Benchmarking LLMs for Zero-Shot Dependency Parsing*. NoDaLiDa 2025. https://aclanthology.org/2025.nodalida-1.13/

Mon 10 Nov (Paula)

Grammars II: Combinatory Categorial Grammar (CCG) and Head-Driven Phrase-Structure Grammar (HPSG).

Wed 12 Nov

Cognitive Dependency and Supertagging.

- Gómez-Rodríguez (2024). *Revisiting Supertagging for Faster HPSG Parsing.* EMNLP 2024. https://aclanthology.org/2024.emnlp-main.635.pdf
- *Additional Paper to be suggested by Fermín on dependencies and language processing*

Mon 17 Nov (Fermín)

Morphology.

Wed 19 Nov (Fermín)

- Hofmann,V. et al. Derivational morphology reveals analogical generalization in large language models https://ora.ox.ac.uk/objects/uuid:076bf87c-1dda-4e6c-b461-ee0989482567/files/rbc386m018. Proceedings of the National Academy of Sciences (PNAS). NB: This is used as a Morphological Compositional Generalization Benchmark in the BabyLM 2025 Evaluation Pipeline (https://babylm.github.io/)
- Ismayilzada et al. (2025). *Evaluating Morphological Compositional Generalization in Large Language Models.* NAACL 2025. https://aclanthology.org/2025.naacl-long.59.pdf

Mon 24 Nov (Suchir)

Tokenization.

Wed 26 Nov

Tokenization and Morphological Complexity.

- Beinborn & Pinter (2023). *Analyzing Cognitive Plausibility of Subword Tokenization.* EMNLP 2023. https://aclanthology.org/2023.emnlp-main.272/
- Arnett & Bergen (2025). *Why Do Language Models Perform Worse for Morphologically Complex Languages?* COLING 2025. https://aclanthology.org/2025.coling-main.441/

Mon 1 Dec

Introducing the Holiday Assignment.

Wed 3 Dec

Q & A Drop-in Session on the Holiday Task.

Course Summary

By the end of the course, you will understand how linguistic theory interfaces with modern computational models and will have experience critically evaluating LLMs through linguistically motivated experiments and datasets.