Shared Units, Divergent Trajectories: The Acquisition of Shared and Specialised Syntactic Representations in Bilingual and Multilingual Language Models

Elena Polyakova[1] Suchir Salhan[1]

[1]Department of Computer Science and Technology, University of Cambridge

Motivation

- **Prior work:** Syntactic phenomena recruit shared or language-specific components, but mainly in fully trained models [3].
- **Gap:** When and how these representations emerge during training remains unclear.
- **Challenge:** Heterogeneous training mixtures obscure the effects of language exposure, order, and data proportions.
- **Approach:** Controlled bilingual models + dense checkpoints enable fine-grained analysis across training.
- **Aim:** Test how typological proximity and training curriculum shape syntactic representation development.

Methodology

- **Method:** Builds on the functional unit-localisation methodology of kryvosheieva-etal-2026-different.
  - Top 1% of syntax-responsive units retained per phenomenon
  - Stability assessed using two-fold cross-validation
  - Cross-phenomenon overlap measures shared syntactic components
- **Languages with corresponding BLiMP-style benchmarks:**
  - Russian [9]
  - Dutch [8]
  - Chinese [7]
  - Turkish [2]
  - Japanese [6]
  - Irish [4]
  - Hindi [1]
- **Framework:**
  - Pretrain 125M-parameter checkpointed bilingual models
  - Five BEETLE curricula: balanced, simultaneous, sequential, classroom, and late [5]
  - Track syntax-responsive units across training checkpoints
  - Analyse how typology and exposure schedule affect syntactic representation development

Why This Methodology?

- **Functional localisation:** Syntax-responsive units are identified from activation differences between grammatical and minimally different ungrammatical sentences using a Welch t-test.
- **Localisation stability:** Two-fold cross-validation measures how consistently the same units are recovered from independent data subsets.
- **Cross-phenomenon sharing:** Overlap between localised unit sets tests whether different syntactic phenomena recruit shared components.
- **Training dynamics:** Dense checkpoints allow us to examine when syntax-responsive units emerge and how localisation changes throughout training.
- **Typological comparison:** BLiMP-style benchmarks across typologically diverse languages allow us to test whether linguistic similarity predicts greater sharing or language-specific specialisation.

A Case Study: L1 Dutch L2 English

- **Pilot setting:** Dutch (L1) → English (L2), before extending to broader typological and curriculum comparisons.
- **Models:** GPT-2-style bilingual models under two training regimes:
  - **Simultaneous:** continued mixed Dutch–English exposure.
  - **Sequential:** switches entirely from Dutch to English halfway through training.
- **Training dynamics:** Analyse four checkpoints from English introduction to the end of training.
- **Goal:** Track changes in model behaviour and syntax-responsive representations across training.

Why Dutch–English?

- **Controlled comparison:** Dutch and English are closely related Germanic languages, providing a useful first test of cross-lingual syntactic sharing.
- **Curriculum contrast:** Simultaneous and sequential training isolate the effect of continued bilingual exposure versus an L1→L2 transition.
- **Checkpoint analysis:** Tracking models after English introduction reveals when syntax-responsive units emerge, stabilise, or reorganise.
- **Pilot role:** Establishes the analysis pipeline before extending to more typologically distant language pairs and additional curricula.

Pilot Results

Phenomenon-Specific Effects

- **Different syntactic phenomena follow distinct trajectories** under simultaneous and sequential training.
- **Determiner–noun agreement and ellipsis:** sequential training shows stronger overlap at later checkpoints.
- **Filler–gap dependencies:** overlap remains comparatively stronger under simultaneous training.
- **Key result:** curriculum effects are phenomenon-specific rather than reflecting a uniform shift in syntactic representations.

Training Dynamics Across Checkpoints

- **Early training:** simultaneous models show stronger within-category overlap.
- **Later training:** sequential models exceed simultaneous models for several constructions.
- **Representational development changes over time:** the relative amount of unit sharing depends on both curriculum and training stage.
- **Localisation is reproducible:** two-fold cross-validation shows consistently high overlap between independently localised unit sets across checkpoints.

Language-Specificity and L1 Retention

- **Sequential training:** substantially lower final Dutch BLiMP-NL accuracy and weaker Dutch unit overlap.
- **L1 forgetting:** reduced Dutch performance after Dutch exposure ceases is consistent with catastrophic forgetting.
- **Simultaneous training:** continued Dutch–English exposure better preserves Dutch grammatical knowledge and representational overlap.
- **Implication:** training curriculum affects not only English syntactic representations, but also retention of L1-specific grammatical knowledge.

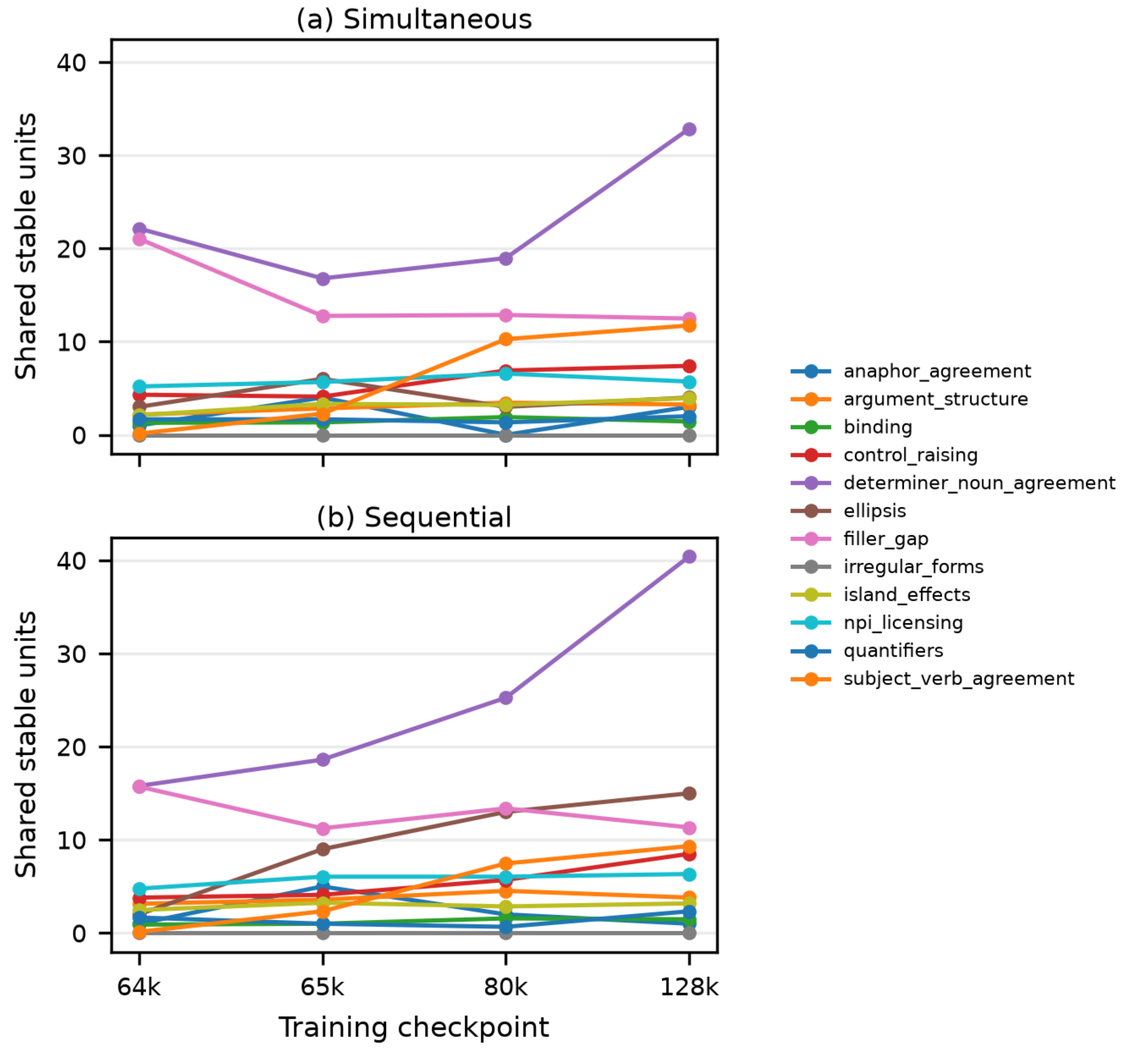


Figure 1. BLiMP within-category cross-overlap of syntax-responsive units for (a) simultaneous and (b) sequential models across checkpoints 64k, 65k, 80k, and 128k.

Discussion & Implications

- **Training curriculum shapes syntactic representations:** simultaneous and sequential exposure produce distinct representational trajectories rather than a uniform shift across syntax.
- **Effects are phenomenon-specific:** different syntactic constructions show different patterns of unit sharing, suggesting that syntactic representation development is not homogeneous.
- **Representation sharing is dynamic:** overlap between syntax-responsive units changes across checkpoints, showing that conclusions drawn only from final models can miss important developmental differences.
- **L1 knowledge depends on continued exposure:** weaker Dutch performance and overlap under sequential training are consistent with greater forgetting once L1 exposure ceases.
- **Controlled bilingual models are informative:** they allow training order, exposure, and language pairing to be manipulated directly, enabling cleaner analysis of the factors shaping multilingual representations.
- **Next step:** extend beyond Dutch–English to typologically diverse language pairs and additional curricula to test whether these developmental patterns generalise.

References

[1] Farah Adeeba, Brian Dillon, Hassan Sajjad, and Rajesh Bhatt.
UrBLiMP: A benchmark for evaluating the linguistic competence of large language models in Urdu.
In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Findings of the Association for Computational Linguistics: ACL 2026*, pages 602–617, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[2] Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza.
Turblimp: A turkish benchmark of linguistic minimal pairs.
In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 16506–16521. Association for Computational Linguistics, 2025.

[3] Daria Kryvosheieva, Andrea Gregor de Varda, Evelina Fedorenko, and Greta Tuckute.
Different types of syntactic agreement recruit the same units within large language models.
In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–227, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[4] Josh McGiff, Khanh-Tung Tran, William Mulcahy, Dáibhidh Ó Luinín, Jake Dalzell, Róisín Ní Bhroin, Adam Burke, Barry O'Sullivan, Hoang D. Nguyen, and Nikola S. Nikolov.
Irish-blimp: A linguistic benchmark for evaluating human and language model performance in a low-resource setting, 2025.

[5] Suchir Salhan, Catherine Arnett, James Michaelov, and Paula Buttery.
Beetle: Structured exposure pretraining in bilingual language models for modelling L2 language processing.
Under review, 2026.

[6] Taiga Someya and Yohei Oseki.
JBLiMP: Japanese benchmark of linguistic minimal pairs.
In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[7] Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer.
Sling: Sino linguistic evaluation of large language models, 2022.

[8] Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank.
BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation.
Computational Linguistics, 51(4):1267–1301, December 2025.

[9] Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov.
RuBLiMP: Russian benchmark of linguistic minimal pairs.