

# Probing the Dynamics of Syntactic Ability Acquisition Throughout LLM Pretraining

Hiroshi Matsuda, Masayuki Asahara

Megagon Labs, Tokyo, Recruit Co., Ltd., National Institute for Japanese Language and Linguistics  
Marunouchi 1-9-2, Chiyoda, Tokyo, Japan, Midorichou 10-2, Tachikawa, Tokyo, Japan  
hiroshi\_matsuda@megagon.ai, masayu-a@ninjal.ac.jp

## Abstract

In this research, we introduce LoRA probing, a lightweight approach for observing how core syntactic abilities emerge during LLM pretraining. Leveraging OLMo-2’s public intermediate checkpoints, we trace learning curves across 24 pretraining stages on 33 Universal Dependencies languages by fine-tuning LoRA with step-by-step parsing instructions and a simple tabular output. To fit the relatively short context length of the OLMo-2, we design a compact `2-step-no-form` prompt template and this matches the baseline in average accuracy while halving the context length and substantially increasing throughput, enabling efficient large-scale evaluation. Token Recall surpasses 0.9 within the first 1–2K pretraining steps, indicating that stable output formatting emerges early. Despite OLMo-2-7B’s English-centric pretraining, LAS exceeds 80 points in 29 of 33 languages; however, relations such as *ibj* and *csubj* show delayed onset and instability across many languages. LoRA probing thus provides a practical, reproducible lens on the cross-lingual dynamics of syntactic acquisition during LLM pretraining.

**Keywords:** syntactic probing, universal dependencies, LoRA fine-tuning, step-by-step prompting

## 1. Introduction

Large language models (LLMs) have achieved striking gains across a wide range of NLP tasks, yet they continue to display gaps in linguistic competence, especially in settings that stress compositionality, morphology, and long-distance dependencies. A plausible explanation for the steady progress of new model releases is the gradual mitigation of such gaps during pretraining. If so, a natural scientific question arises: *when* and *how* do specific syntactic abilities emerge as pretraining unfolds, and which aspects remain fragile even near the end of training?

To answer this question, we need instruments that can observe the development of syntax without excessively altering the model under study. Prior approaches offer complementary perspectives: *structural probing* attempts to reconstruct syntax from internal representations, while *behavioral evaluation* tests task performance via prompting or fine-tuning. Structural probes illuminate geometry but can drift from actual behavior; behavioral tests reveal competence but may blur whether success stems from base models or from heavy task-specific adaptations. We aim to bridge this gap with a probe that is both *lightweight* and *behaviorally faithful*.

We therefore introduce **LoRA probing**, a simple protocol that fine-tunes only the parameters of Low-Rank Adapter (Hu et al., 2022) while keeping the base model fixed, and then evaluates the base model’s latent syntactic competence via the adapter’s outputs. Concretely, we cast dependency parsing as a step-by-step generation task and read out parses in a tabular format introduced in Mat-

suda et al. (2025). To fit models with relatively short context windows, we design a compact `2-step-no-form` parsing prompt: the model first predicts UPOS tags and then predicts HEAD and DEPREL without reproducing word’s FORM. This yields behaviorally interpretable signal while minimizing prompt length and training cost.

Our empirical setting leverages OLMo-2-7B (Team OLMo et al., 2025), which publicly releases intermediate checkpoints during pretraining. These snapshots allow us to trace learning curves over time: we train LoRA at multiple checkpoints and evaluate on Universal Dependencies (UD) treebanks (Nivre et al., 2020) spanning 33 languages. This multilingual canvas exposes both universal and language-specific dynamics, enabling us to ask whether, for instance, tagging stabilizes earlier than relation labeling, or whether particular relations (e.g., *ibj*, *csubj*) consistently lag across typologically diverse languages. We structure the study around four research questions:

**RQ1: Prompt efficiency.** Can a compact `2-step-no-form` prompt match or surpass a conventional `3-step` prompt in accuracy while improving throughput, making large-scale evaluation feasible?

**RQ2: Onset and stabilization.** At what stages of pretraining do token-level formatting discipline emerge (as measured by TOKEN RECALL), and how early do core syntactic signals become reliable?

**RQ3: Cross-lingual trajectory.** How do

learning curves differ across languages for UPOS and for major dependency relations, and which patterns appear universal versus language-specific?

**RQ4: Persistent weakness.** Which relations remain difficult or unstable late in pretraining (e.g., *iobj*, *csubj*), and what does this imply about data imbalance, annotation conventions, or inductive biases?

Beyond reporting aggregate scores, we emphasize *dynamics*: the evolution of competence as pretraining progresses. This perspective complements single-snapshot evaluations and helps explain why certain linguistic contexts remain challenging for LLMs even as overall performance improves. By clarifying *which* syntactic abilities appear early, *which* plateau, and *which* lag or fluctuate, LoRA probing offers actionable guidance for model developers (e.g., curriculum design, data selection) and for the broader community seeking the ways to improve linguistic competence in LLMs.

Finally, the remainder of the paper is organized as follows. Section 2 situates our work relative to structural probes and behavioral evaluations. Section 3 details the LoRA probing protocol, the `2-step-no-form` prompt, datasets, checkpoints, and evaluation criteria. Section 4 presents cross-lingual learning curves and ablations. Section 5 discusses implications and limitations, and Section 6 concludes.

## 2. Related Work

A growing body of work investigates how syntactic abilities emerge in language models during *pretraining*, including studies that track the evolution of linguistic competence over time (Liu et al., 2021; Pérez-Mayos et al., 2021; Blevins et al., 2022). In the multilingual setting, Wang et al. (2024) examine the emergence of cross-lingual alignment in BLOOM (BigScience Workshop et al., 2023), highlighting that syntactic behavior can develop unevenly across languages even within the same pretraining regime. These efforts suggest that probing the *trajectory* of syntactic acquisition—rather than only evaluating end points—yields insight into what models learn and when they learn it.

Following the taxonomy of Liu et al. (2021), probing methods for syntactic competence can be grouped into *structural* and *behavioral* approaches. Structural probes attach a lightweight classifier to representations from a pretrained model and test whether syntactic structure can be recovered. They illuminate the geometry of internal states but may overestimate competence if the readout is powerful relative to the underlying signal.

Behavioral evaluations, by contrast, assess a model’s own inference behavior through prompt-

ing or task-specific fine-tuning; they reflect observable competence but can conflate improvements in the base model with heavy adaptation or prompt engineering, and they often require long prompts that limit scalability. Our work builds on the behavioral perspective while addressing these limitations. We introduce a minimally invasive protocol—LoRA probing—which reduces the risk that the probe itself *creates* the competence it measures.

Although a zero- or few-shot inference with an LLM would entirely remove the effect of adding a probe, in dependency parsing it is difficult to assess the parsing potential of intermediate pretraining checkpoints in such settings for the following reasons:

- Head identification accuracy is quite low in zero- or few-shot settings (Tian et al., 2024).
- The definitions of UPOS tags and DEPREL labels are not learned during pretraining and are therefore almost impossible to reproduce in a zero-shot setting.
- Similarly, given the limited context length, learning the definitions of tags and labels in a few-shot setting is also difficult
- For checkpoints from the early stages of pretraining, it is difficult to make the model adhere to the required output format using few-shot instructions alone.

For these reasons, this study adopts a LoRA-based probing approach.

## 3. Method and Setup

### 3.1. LoRA probing

We propose LoRA probing, a minimally invasive approach to probe the syntactic abilities of LLMs that combines step-by-step parsing prompts with LoRA-based fine-tuning (Matsuda et al., 2025). Conceptually, LoRA probing sits between structural and behavioral paradigms: like structural probes, it adds only lightweight adapter layers to a pretrained model; like behavioral probes, it evaluates the model through its own auto-regressive decoding, yielding behaviorally meaningful signals. Crucially, because we evaluate multiple intermediate checkpoints, LoRA probing enables a dynamic view of syntactic acquisition, complementing structural analyses and snapshot behavioral tests with a temporally resolved account of how specific abilities emerge during pretraining.

Our protocol keeps the base model fixed and fine-tunes only low-rank adapter parameters, then casts dependency parsing as a step-by-step generation task whose outputs are read as tabular TSV parses.

We leverage OLMo-2’s public intermediate pre-training checkpoints to study *when* and *which* syntactic abilities emerge. Using Universal Dependencies (UD) treebanks across 33 languages, we analyze which features are acquired—and which remain fragile—over the course of pretraining.

### 3.2. Parsing prompt

The 3-step parsing prompt proposed by (Matsuda et al., 2025) assumes underlying LLM with a maximum context length of 8,192 tokens. However, OLMo-2’s maximum context length is only 4,096. To accommodate this limitation, we propose a 2-step-no-form parsing prompt: instead of predicting UPOS, HEAD and DEPREL in three separate steps, we first predict UPOS and then jointly predict HEAD and DEPREL. We also remove the FORM field from the output records. This design halves the token budget and the associated computational cost during both training and inference, while preserving the interpretability. An example instance of 2-step-no-form prompt is shown in Figure 1.

As a preliminary study, we compared the parsing accuracy and throughput of the 3-step prompt with the 2-step-no-form prompt in UD English (en\_ewt) (Silveira et al., 2014) using gemma-2-9b (Gemma Team et al., 2024). The means and standard deviations of four independent runs are shown in Table 1. The 2-step-no-form matches the 3-step baseline in average accuracy while substantially increasing throughput (training +45%; inference +129%), enabling efficient large-scale evaluation.

In addition, the single-run evaluation results for all 33 languages are presented in Appendix B. The macro-averaged values indicate a clear tendency for the 2-step-no-form prompt to perform comparably to, or better than, the 3-step prompt in all models. Accordingly, we use the 2-step-no-form prompt for all subsequent experiments.

	3-step	2-step-no-form
Accuracy [%]		
TOKEN RECALL	100.0±0.1	100.0±0.0
UPOS (F1)	98.6±0.1	98.4±0.1
UAS (F1)	95.4±0.1	95.2±0.2
LAS (F1)	93.9±0.1	93.5±0.0
Throughput [sentence/sec]		
Training	3.04±0.01	4.40±0.03
Inference	11.6±0.03	24.1±0.13

Table 1: Comparison of accuracy and throughput of gemma-2-9b trained on en\_ewt over 4 runs. Single-run evaluation results for all 33 languages are shown in Appendix B.

```
{role: system} You are an English linguist and
specialize in English dependency parsing
based on Universal Dependencies.

{role: user} We will now perform dependency
parsing on English sentence. After splitting
the input sentence into words with indexes as
shown below, execute following two tasks:

- Task 1
Create a TSV with two fields: word index
from 1 to 6 + part of speech.

- Task 2
Add two fields to each row of the output of
Task 1: the dependent word index + the
Universal Dependencies relation label.
However, for the word that is the main
predicate of the sentence, the dependent word
index should be 0.

input sentence:
I'm just speculating now.

words:
I
'm
just
speculating
now
.

{role: assistant}
- Task 1
1 PRON
2 AUX
3 ADV
4 VERB
5 ADV
6 PUNCT
- Task 2
1 PRON 4 nsubj
2 AUX 4 aux
3 ADV 4 advmod
4 VERB 0 root
5 ADV 4 advmod
6 PUNCT 4 punct
<eos>
```

Figure 1: An example of the 2-step-no-form prompt instance. The parts that change from instance to instance are shown in **BOLD**. The **SHADED REGION** in the assistant-role corresponds to the range over which the loss gradient is computed during training, and to the decoded text during inference. At inference time, the span from the system-role up to the assistant-role header is provided as input, and decoding of the subsequent segment continues until <eos> is generated.

### 3.3. Multilingual capability of OLMo-2-7B

Our primary analysis targets OLMo-2-7B and its publicly released intermediate pretraining checkpoints, which enable stage-wise probing of syntactic acquisition in LLMs. OLMo-2-7B is pretrained on 3.9T English tokens drawn chiefly from the DCLM-baseline corpus (Li et al., 2025). DCLM-baseline retains only documents with a language-identifier estimate of ≥ 99% English, so non-English content is vanishingly small. Given this near-monolingual pretraining mixture, it is important to

	en_ewt	fr_gsd	cs_pdt	ja_gsd	lt_alksnis	33-AVG
gemma-2-9b	93.5	94.6	94.5	93.9	84.9	91.1
Llama-3.1-8B	93.4	94.4	94.2	93.5	81.0	90.3
Qwen2.5-7B	93.1	94.4	94.1	92.9	76.6	89.5
OLMo-2-7B	92.4	92.8	92.1	91.5	64.9	86.4

Table 2: Comparison of LAS (F1; %) among multilingual LLMs and English-centric OLMo-2-7B using 2-step-no-form prompt. The overall results for all 33 languages are shown in Appendix B.

test whether OLMo-2-7B can still be meaningfully evaluated on multilingual syntactic tasks.

As a preliminary evaluation, we verify that OLMo-2-7B achieves reasonable performance on multilingual syntactic evaluation, comparing it with strong multilingual LLM baselines. Table 2 presents the comparison of accuracy metrics<sup>1</sup> of three multilingual LLMs—gemma-2-9b, Llama-3.1-8B (Grattafiori et al., 2024), and Qwen2.5-7B (Qwen et al., 2025)—as well as OLMo-2-7B on the datasets described in Table 3. OLMo-2-7B achieved performance comparable to the multilingual LLMs in English and exhibited unexpectedly strong results in French, Czech, and Japanese, even though these languages are almost absent from its pretraining corpus. In contrast, for Lithuanian, OLMo-2-7B scored 20 LAS points lower than gemma-2-9b<sup>2</sup>. The entire evaluation results for all 33 languages are shown in Appendix B. The macro-average LAS of OLMo-2-7B across 33 languages was 86.4%, a level considered sufficient for comparing dependency parsing accuracies. These findings suggest that OLMo-2-7B has sufficient syntactic ability to support a meaningful evaluation of its multilingual performance. We then use LoRA probing to obtain temporally resolved trajectories of syntactic acquisition across checkpoints.

### 3.4. Datasets and language families

We evaluate our method on the Universal Dependencies (UD) v2.15 release.<sup>3</sup> We selected 33 treebanks listed in Table 3, each representing a distinct language and containing more than 40,000 words in the *training* split. To balance the fine-tuning data volume across languages, we set the number of training epochs so that the product of the training set size (in words) and the number of epochs falls within the range of 300K–600K. In addition, more

<sup>1</sup>The metrics definition is described in Section 4.

<sup>2</sup>The languages in which OLMo-2-7B achieved LAS below 80% were Irish, Lithuanian, Simplified Chinese, and Turkish.

<sup>3</sup>At the time of submission, UD v2.17 was the latest release; however, all results in this paper are evaluated on UD v2.15 to allow for a fair comparison with baseline.

detailed statistics including all of the UPOS tags and the DEPREL labels for all 33 languages are presented in the Appendix C.

For typological analysis, we group languages by family following the WALS Online (Dryer and Haspelmath, 2013) adopted on the UD website.

### 3.5. Fine-tuning method

We adopt Low-Rank Adapters (LoRA) for fine-tuning the LLMs, largely following the publicly available implementation and experimental setup<sup>4</sup> of Matsuda et al. (2025). We modify only two aspects: (i) the number of training epochs, adjusted to balance the training volume across treebanks, and (ii) the learning rate, reduced from  $3 \times 10^{-4}$  to  $1 \times 10^{-4}$  to mitigate overfitting when fine-tuning the intermediate checkpoints of OLMo-2-7B.

### 3.6. Tokens per step in OLMo-2 pretraining

During the pretraining of OLMo-2, the batch size is set to 1,024 and the context length to 4,096, resulting in approximately 4 million tokens being processed per training step. The X-axis of the learning curves in the experimental results shown in the following sections represents the number of steps; therefore, to convert this to the number of pretrained tokens, multiply by 4 million.

## 4. Experiments

In this section, we address the question of learning dynamics—specifically, how the syntactic abilities of OLMo-2-7B emerge over the course of pretraining when evaluated at intermediate checkpoints. We use the datasets and protocol described in Section 3 (UD v2.15, 33 languages, 2-step-no-form parsing prompt with LoRA).<sup>5</sup>

**Scope.** Evaluation covers 33 UD languages and use the official train/dev/test splits without modification. We probe intermediate pretraining checkpoints of OLMo-2-7B to obtain temporally resolved trajectories of syntactic acquisition. At every checkpoint, we perform SFT on LoRA from its initial state.

**Tasks and metrics.** Following our setup, we cast dependency parsing as step-by-step generation and report UPOS, UAS, and LAS as *F1 score*<sup>6</sup>

<sup>4</sup><https://github.com/megagonlabs/llmpp/tree/20250612>

<sup>5</sup>All experiments were run on Google Cloud; the aggregate compute for training and inference was approximately 1,600 GPU-hours, measured in H100-80GB.

<sup>6</sup>UPOS: pos tagging accuracy. UAS (Unlabeled Attachment Score): HEAD accuracy. LAS (Labeled Attachment Score: simultaneous accuracy of HEAD and DEPREL. Following standard evaluation tool of UD, we

Family	Language	Treebank	Words in Training-set Total ( <i>obj</i> , <i>csbj</i> )	Training Epochs	Training Words	Words in Test-set Total ( <i>obj</i> , <i>csbj</i> )
Germanic	Danish	da_ddt (Johannsen et al., 2015)	80,378 (120, 0)	4	321,512	10,023 (15, 0)
	Dutch	nl_alpino (Bouma and van Noord, 2017)	186,027 (509, 400)	2	372,054	11,046 (13, 21)
	English	en_ewt (Silveira et al., 2014)	204,579 (649, 293)	2	409,158	25,094 (71, 25)
	German	de_gsd (McDonald et al., 2013)	263,791 (0, 214)	2	527,582	16,498 (0, 16)
	Norwegian	no_bokmaal*	243,886 (471, 794)	2	487,772	29,966 (51, 103)
	Swedish	sv_talbanken (Nivre and Megyesi, 2007)	66,646 (105, 249)	6	399,876	20,377 (37, 93)
Romance	Catalan	ca_ancora (Taulé et al., 2008)	429,578 (0, 736)	1	429,578	59,610 (0, 101)
	French	fr_gsd (Guillaume et al., 2019)	354,652 (813, 195)	1	354,652	10,018 (34, 12)
	Italian	it_isdt (Bosco et al., 2013)	276,014 (648, 300)	2	552,028	10,417 (20, 3)
	Portuguese	pt_gsd*	255,116 (377, 490)	2	510,232	31,477 (45, 61)
	Romanian	ro_rrt (Mititelu, 2018)	185,125 (1,366, 755)	2	370,250	16,324 (139, 41)
	Spanish	es_ancora (Taulé et al., 2008)	453,039 (0, 988)	1	453,039	53,622 (0, 94)
Slavic	Belarusian	be_hse (Shishkina and Lyashevskaya, 2021)	273,181 (2,269, 828)	2	546,362	15,997 (82, 31)
	Bulgarian	bg_btb (Osenova and Simov, 2004)	124,336 (2,818, 358)	3	373,008	15,724 (336, 43)
	Croatian	hr_set (Agić and Ljubešić, 2015)	152,857 (538, 271)	2	305,714	24,260 (77, 40)
	Czech	cs_pdt (Hajič et al., 2024)	1,173,285 (455, 4,870)	0.35	410,650	173,918 (66, 680)
	Polish	pl_pdb (Wróblewska, 2018)	281,685 (5,311, 185)	2	563,370	33,616 (673, 16)
	Russian	ru_syntagrus (Droganova et al., 2018)	1,204,640 (10,424, 6,743)	0.35	421,624	157,718 (1,425, 897)
	Serbian	sr_set (Batanović et al., 2018)	74,259 (2, 160)	5	371,295	11,421 (0, 24)
	Slovak	sk_snk (Zeman, 2017)	80,628 (80, 132)	4	322,512	12,736 (1, 18)
	Slovenian	sl_ssaj (Dobrovoljc et al., 2017)	215,155 (1,412, 722)	2	430,310	25,442 (145, 78)
	Ukrainian	uk_iu*	92,927 (298, 426)	4	371,708	17,217 (33, 57)
Baltic	Latvian	lv_lvtb (Pretkalniundefineda et al., 2018)	255,954 (6,141, 853)	2	511,908	37,162 (874, 115)
	Lithuanian	lt_alksnis (Bielinskienė et al., 2016)	47,641 (2, 322)	8	381,128	10,846 (0, 64)
Celtic	Irish	ga_idt (Lynn and Foster, 2016)	95,881 (0, 746)	4	383,524	10,109 (0, 95)
Finnic	Estonian	et_edt (Muischnek et al., 2016)	344,581 (0, 1,546)	1	344,581	48,465 (0, 243)
	Finnish	fi_ftb*	127,602 (0, 402)	3	382,806	16,286 (0, 58)
Greek	Greek	el_gdt (Prokopidis and Papageorgiou, 2017)	42,326 (59, 144)	9	380,934	10,672 (9, 40)
Turkic	Turkish	tr_boun (Marşan et al., 2022)	100,713 (184, 467)	4	402,852	12,210 (24, 72)
Afro-Asiatic	Arabic	ar_padt (Hajič et al., 2009)	223,881 (89, 442)	2	447,762	28,264 (6, 42)
Japanese	Japanese	ja_gsd (Asahara et al., 2018)	168,333 (0, 138)	2	336,666	13,034 (0, 12)
Korean	Korean	ko_gsd (Chun et al., 2018)	56,687 (77, 12)	7	396,809	11,677 (15, 5)
Sino-Tibetan	Chinese	zh_gsdsimp*	98,616 (63, 296)	4	394,464	12,012 (6, 37)

Table 3: Treebanks for 33 languages and their language families used in the experiments. These treebanks are distributed under licenses that permit research use; see the LICENSE file in the repository for details. Treebanks marked with an asterisk (\*) lack a clear bibliographic reference in their repositories. All statistics are based on Universal Dependencies v2.15 and more detailed statistics including the occurrence count of each UPOS tag and DEPREL label for all 33 languages are presented in the Appendix C.

on test sets, alongside TOKEN RECALL<sup>7</sup> to quantify output formatting stability.

**Reporting.** The accuracy figures reflect a *single* fine-tuning and decoding run. When multiple runs are conducted, we report the mean and standard deviation over the stated number of seeds.

#### 4.1. Emergence in output formatting

Figure 2 shows the learning curves of TOKEN RECALL for 2-step-no-form prompt. Across all 33 languages, TOKEN RECALL exceeded 90 points within the first 1,000 to 2,000 steps, confirming that evaluation of the core syntactic parsing capability becomes reliable at an early stage of pretraining.

use F1 score to consider word segmentation correctness.

<sup>7</sup>Word segmentation errors may appear in the output even when the correct segmentation is provided as input, due to the hallucination-like behavior of LLMs. Therefore, we adopt Recall as the evaluation metric, measuring the proportion of words that are correctly generated.

#### 4.2. Emergence in syntactic ability

The results are presented by grouping the languages into four categories: Germanic, Romance, Slavic, and Others.

Since the pretraining data of OLMo-2-7B is predominantly English, the UPOS, UAS, and LAS curves for non-English languages remain below the envelope formed by English throughout the pretraining process. No substantial differences are observed among the Germanic, Romance, and Slavic language groups, whereas several languages in the Others group exhibit a slower initial rise.

#### 4.3. Detailed analysis in each language

We analyze learning curves over the intermediate pretraining steps for 11 major UPOS tags and 13 key DEPREL labels. In the following subsections, we first analyze English as a representative of the Germanic family; we then analyze French as a representative of the Romance family and Czech as a representative of the Slavic family. The learn-

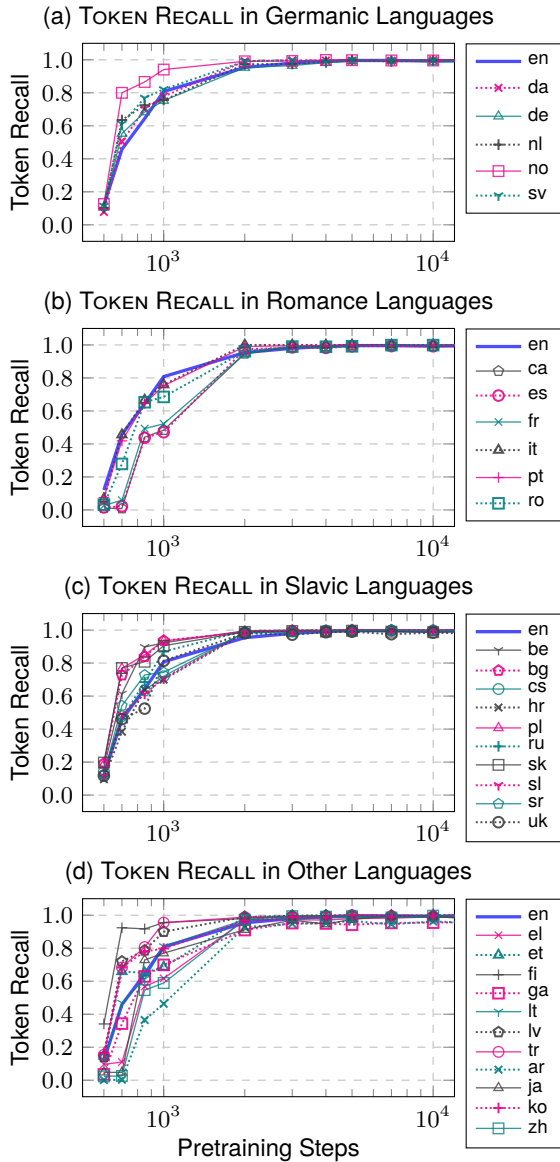


Figure 2: TOKEN RECALL acquisition through pre-training of OLMo-2-7B in 33 languages.

ing curves for all 33 languages are included in Appendix C.

#### 4.3.1. Germanic: English (en\_ewt)

The learning curves of UPOS and LAS of en\_ewt are shown in Figure 6. For UPOS, learning first progresses on the closed-class categories *AUX*, *CCONJ*, *DET*, and *PRON*, followed by *VERB* and *PART*. Subsequently, the open-class categories *NOUN*, *ADJ*, and *ADV* are acquired, with *PROPN* and *SCONJ* learned in the later stages.

For dependency relations, the model initially learns *root*, *nsubj*, *advmod*, and *amod*, followed by *obj*, *nmod*, *obl*, and *xcomp*. Learning of the clausal relations *ccomp*, *acl*, and *advcl* then proceeds smoothly, whereas *iobj* and *csubj* show de-

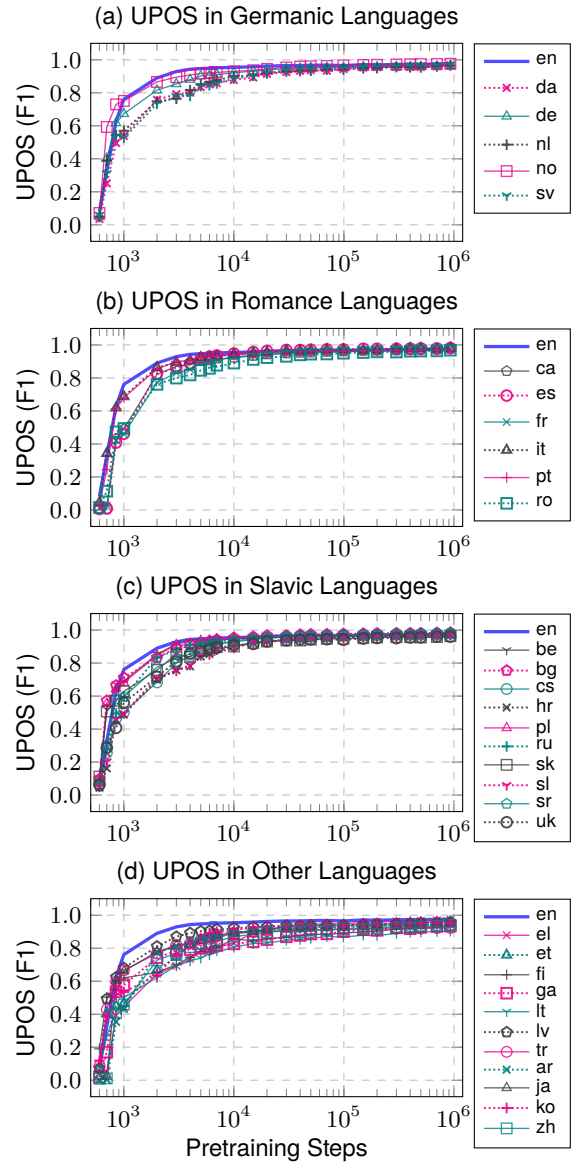


Figure 3: UPOS acquisition through pre-training of OLMo-2-7B in 33 languages.

layed onset and exhibit large fluctuations in their learning curves, indicating under-trained or annotation inconsistency. The unstable tendency on *iobj* is common in 30 out of 33 languages but exceptions are Bulgarian, Latvian, and Russian. In the same way, the unstable tendency on *csubj* is found in 29 out of 33 languages but exceptions are Irish, Norwegian, Russian, and Swedish.

#### 4.3.2. Romance: French (fr\_gsd)

The learning curves of UPOS and LAS of fr\_gsd are shown in Figure 7. For UPOS, the model first acquires closed-class categories such as *DET* and *CCONJ*, followed by *AUX*, *PRON*, and *PROPN* together with *NOUN*. Subsequently, *VERB* and *ADV* emerge almost simultaneously, followed by *ADJ*.

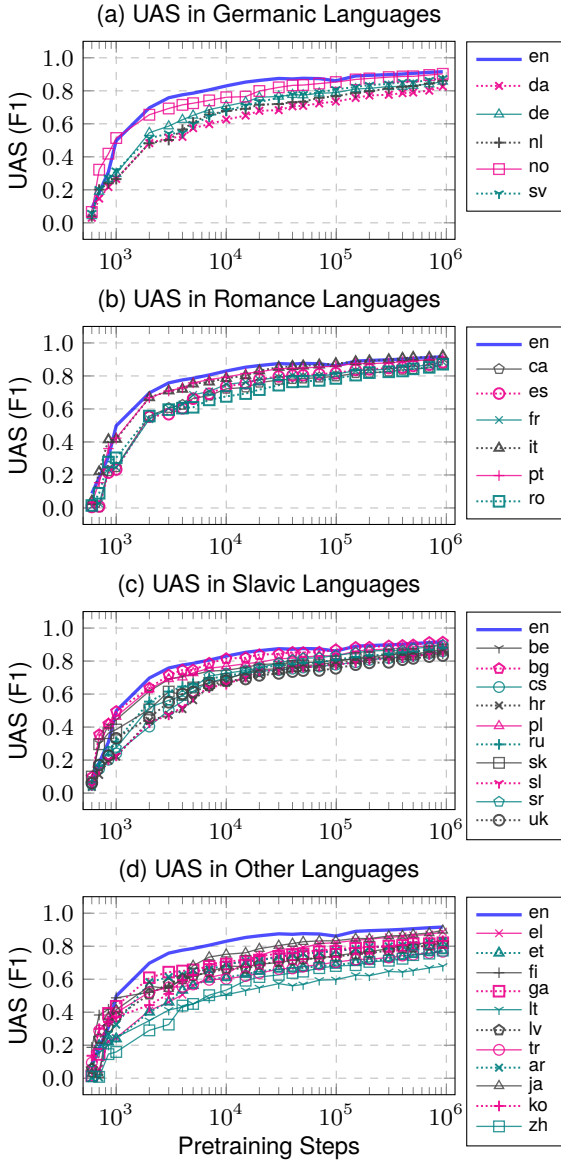


Figure 4: Unlabeled Attachment acquisition through pretraining of `OLMo-2-7B` in 33 languages.

The learning of *SCONJ* initially rises to around 0.7, after which it remains somewhat unstable but continues to show a gradual upward trend.

For DEPREL, *nmod* and *nsubj* emerge first, followed by *amod*, *advmod*, and *root* with a slight delay, then *obj* and *obl*. After further lag, the model begins to learn clausal relations such as *acl* and *xcomp*, followed later by *advcl* and *ccomp*. The learning of *iobj* fluctuates while gradually improving, similar to the case in English, whereas *csubj* shows little progress throughout training process.

#### 4.3.3. Slavic: Czech (`cs_pdt`)

The learning curves of UPOS and LAS of `cs_pdt` are shown in Figure 8. For UPOS, learning first progresses among the closed-class categories ex-

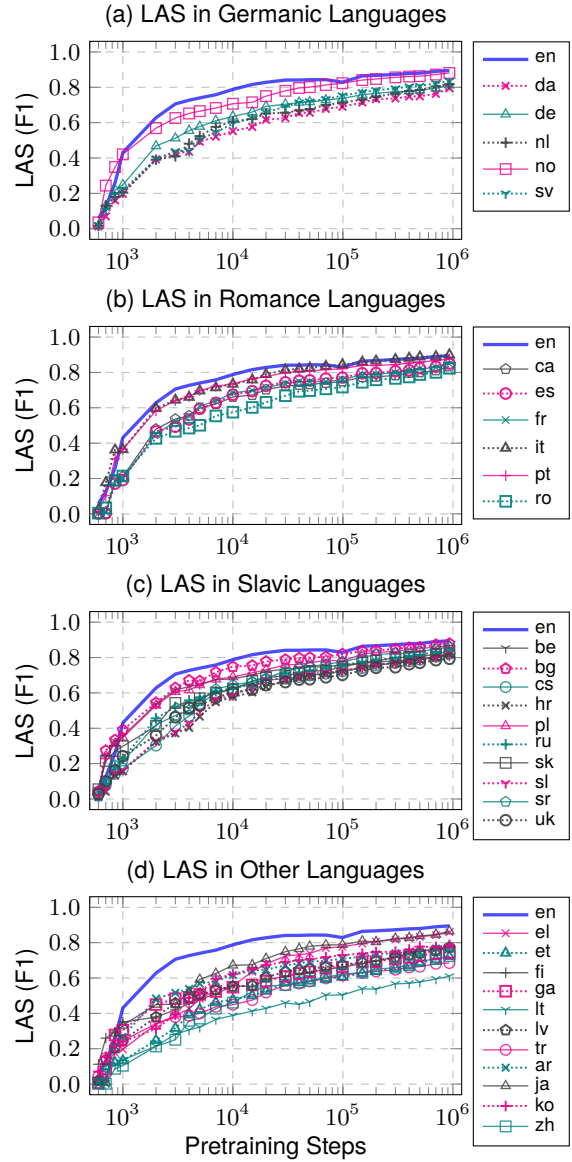


Figure 5: Labeled Attachment acquisition through pretraining of `OLMo-2-7B` in 33 languages.

cluding *PART*. Within the open-class categories, *PROPN* emerges earlier than *NOUN*, but its growth soon plateaus, after which *NOUN*, *ADJ*, and *VERB* overtake it around 2,000–3,000 training steps. *ADV* rises more slowly than the other open-class categories. *PART* shows the slowest onset and reaches only about 80 points accuracy in the final stage, which is lower than other languages.

For DEPREL, *root*, *obl*, *amod*, and *nmod* increase first up to roughly 2,000 steps. Subsequently, *amod* and *advmod* accelerate, while *obl* decelerates and is followed by *nsubj*. With a considerable delay, *obj*, *acl*, and *xcomp* start to rise synchronously, followed by *ccomp* and *advcl*. *csubj* exhibits the slowest onset but increases steadily compared with other languages. The accuracy of *iobj* remains close to zero throughout pretraining.

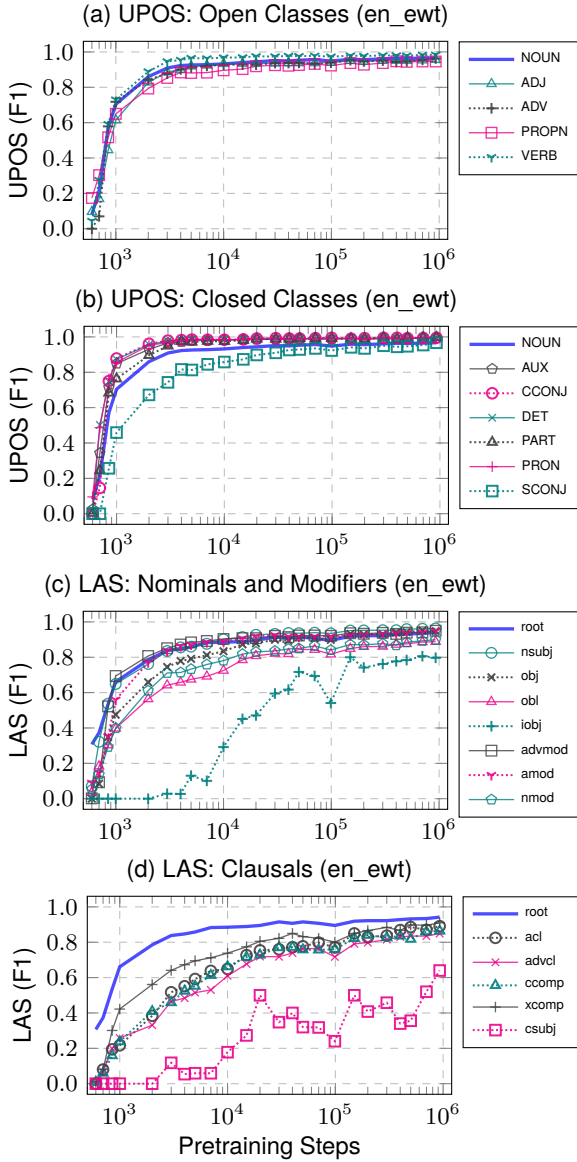


Figure 6: Syntactic acquisition through pretraining of OLMo-2-7B (English: en\_ewt).

## 5. Discussion

The results answer the four research questions posed in Section 1.

For **RQ1**, the compact `2-step-no-form` prompt matches the conventional `3-step` prompt in accuracy while halving the context length and substantially improving throughput, making large-scale multilingual evaluation feasible. This tendency is also supported by the single-run results across 33 languages, where the `2-step-no-form` prompt performs comparably to, or slightly better than, the `3-step` prompt on macro-average.

For **RQ2**, Token Recall exceeds 0.9 within the first 1,000–2,000 pretraining steps in all 33 languages, indicating that output formatting stabilizes early. Core syntactic signals also emerge from an

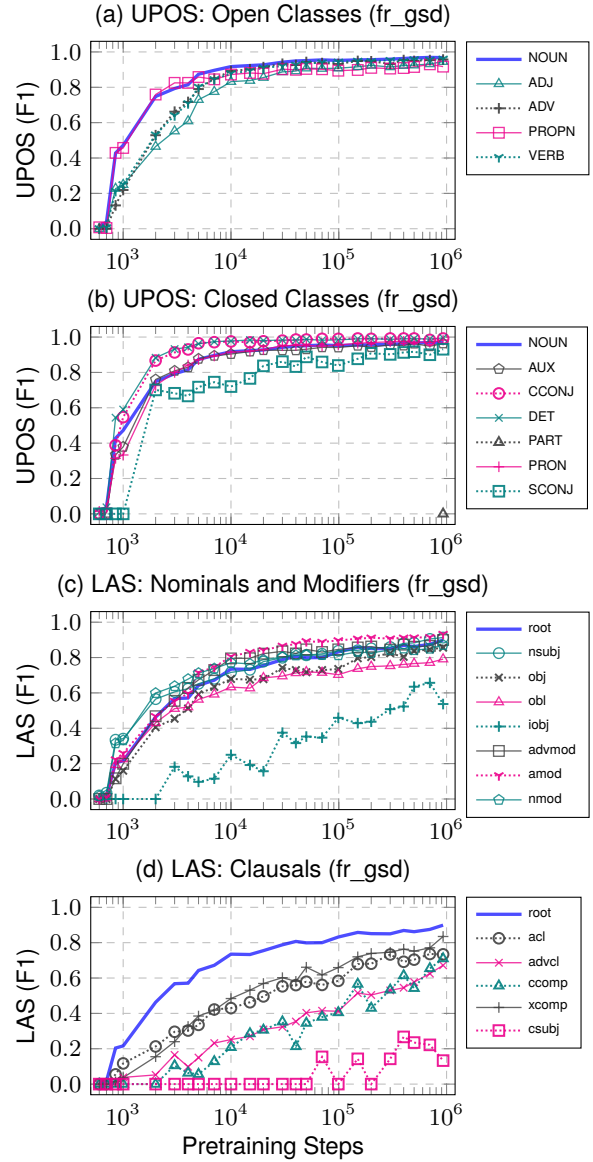


Figure 7: Syntactic acquisition through pretraining of OLMo-2-7B (French: fr\_gsd).

early stage: closed-class UPOS categories generally precede open-class ones, and major nominal relations tend to precede clausal relations.

For **RQ3**, the overall learning trajectories are broadly similar across languages. English forms the upper envelope throughout pretraining, consistent with the English-centric pretraining mixture. No substantial differences are observed among the Germanic, Romance, and Slavic groups, whereas several languages in the Others group show a slower initial rise. Even so, OLMo-2-7B achieves LAS above 80 in 29 of 33 languages, suggesting that substantial syntactic ability emerges across typologically diverse languages.

For **RQ4**, *iobj* and *csubj* remain the most difficult relations. Stable *iobj* curves are observed in Bulgarian, Latvian, and Russian, while stable *csubj*

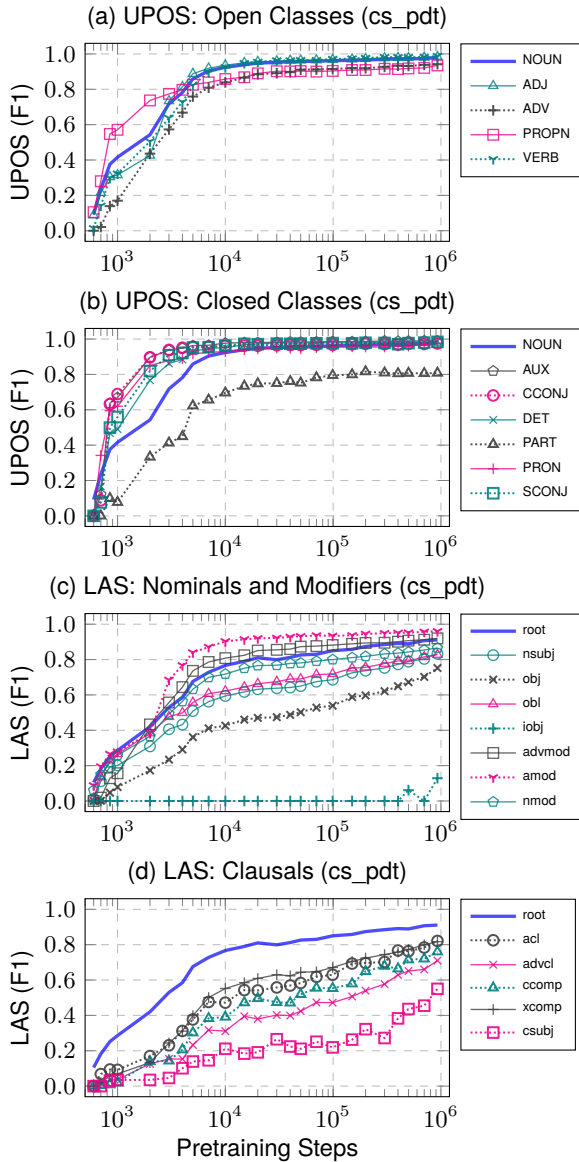


Figure 8: Syntactic acquisition through pretraining of OLMo-2-7B (Czech: cs\_pdt).

curves are found in Irish, Norwegian, Russian, and Swedish. Their instability elsewhere is partly explained by low frequency or train-test mismatch, but cases such as Czech and Greek suggest that data scarcity alone is insufficient. We therefore interpret the persistent weakness of these relations as reflecting a combination of data scarcity, annotation inconsistency, and the inherent difficulty of the phenomena. Beyond *iobj* and *csubj*, several other DEPRELs also show unstable learning and persistently low accuracy, which should be investigated in future work.

## 6. Conclusion

We proposed LoRA probing, a novel approach to investigate how LLMs acquire syntactic abilities dur-

ing pretraining by combining parsing prompts with LoRA fine-tuning. To accommodate OLMo-2’s relatively short context window, we introduced a compact `2-step-no-form` parsing prompt, halving the context length while achieving higher throughput with comparable accuracy.

We further applied LoRA fine-tuning with UD treebanks of 33 languages and the `2-step-no-form` parsing prompt to OLMo-2-7B’s intermediate checkpoints. The proposed method achieved TOKEN RECALL above 90 points within the first 2,000 pretraining steps across all 33 languages, demonstrating that LoRA probing can reveal syntactic structures from the very early stages of LLM pretraining. Through LoRA probing, we analyzed the development of syntactic ability across 33 languages in intermediate checkpoints of OLMo-2-7B.

We found that the learning of *iobj* and *csubj* relations remained unstable, with low accuracy even in the final stages of pretraining in many languages. These findings suggest that specific dependency relations are inherently harder for LLMs to internalize during pretraining, and that LoRA probing provides a practical means to observe and quantify such cross-lingual differences in syntactic acquisition.

## Acknowledgments

This work was conducted as part of a collaborative research project between Recruit Co., Ltd. and the National Institute for Japanese Language and Linguistics. We are grateful to all those involved in the management and support of this project. We would also like to express our sincere gratitude to Yuji Matsumoto of RIKEN AIP for his valuable advice from the early stages of this research. Finally, we thank the anonymous reviewers for their constructive and detailed comments.

## Limitations

### Effects of test set contamination in language model pretraining

The Universal Dependencies (UD) datasets used in our experiments, including the test sets, are freely and publicly available on the internet. Meanwhile, all four large language models evaluated in this study were pretrained on large-scale web corpora. Consequently, when evaluating UD tasks with these LLMs, there is a potential risk that test set contamination may artificially inflate performance.

According to Matsuda et al. (2025), such contamination may contribute to higher initial accuracy in UPOS tagging; however, no significant effects have been observed for UAS or LAS. In our experiments, because all metrics—TOKEN RECALL, UPOS, UAS,

and LAS—were evaluated using the same base model configurations, any contamination effects are expected to cancel out. Therefore, we believe that the overall conclusions of this study remain unaffected.

### Use of an English-centric LLM for multilingual evaluation

In this study, we evaluated multilingual syntactic parsing tasks using `OLMo-2-7B` and its intermediate checkpoints from pretraining. However, it should be noted that the DCLM-baseline dataset, which constitutes the majority of `OLMo-2-7B`'s pretraining data, is composed of texts that passed an English filter, retaining only documents estimated to have a probability of over 99% of being English. Consequently, non-English languages are at an inherent disadvantage in the evaluation setting.

## Bibliographical References

- BigScience Workshop, :, Teven Le Scao, et al. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#).
- Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. [Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.
- Gemma Team, Morgane Riviere, et al. 2024. [Gemma 2: Improving open language models at a practical size](#).
- Aaron Grattafiori et al. 2024. [The llama 3 herd of models](#).
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Jeffrey Li et al. 2025. [Datacomp-1m: In search of the next generation of training sets for language models](#).
- Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. [Probing across time: What does RoBERTa know and when?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hiroshi Matsuda, Chunpeng Ma, and Masayuki Asahara. 2025. [Step-by-step instructions and a simple tabular output format improve the dependency parsing accuracy of LLMs](#). In *Proceedings of the 18th International Conference on Parsing Technologies (IWPT, SyntaxFest 2025)*, pages 11–19, Ljubljana, Slovenia. Association for Computational Linguistics.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043. European Language Resources Association.
- Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. 2021. [How much pretraining data do language models need to learn syntax?](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1582, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Qwen, :, An Yang, et al. 2025. [Qwen2.5 technical report](#).
- Team OLMo, Pete Walsh, et al. 2025. [2 olmo 2 furious](#).
- Yuanhe Tian, Fei Xia, and Yan Song. 2024. [Large language models are no longer shallow parsers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7131–7142, Bangkok, Thailand. Association for Computational Linguistics.
- Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. [Probing the emergence of cross-lingual alignment during LLM training](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

## Language Resource References

- Agić, Željko and Ljubešić, Nikola. 2015. [Universal Dependencies for Croatian \(that work for Serbian, too\)](#). INCOMA Ltd. Shoumen, BULGARIA.

- Asahara, Masayuki and Kanayama, Hiroshi and Tanaka, Takaaki and Miyao, Yusuke and Uematsu, Sumire and Mori, Shinsuke and Matsumoto, Yuji and Omura, Mai and Murawaki, Yugo. 2018. *Universal Dependencies Version 2 for Japanese*. European Language Resources Association (ELRA).
- Batanović, Vuk and Ljubešić, Nikola and Samardžić, Tanja and Erjavec, Tomaž. 2018. *Training corpus SETimes.SR 1.0*. Slovenian language resource repository CLARIN.SI.
- Agnė Bielinškienė and Loïc Boizou and Jolanta Kovalevskaitė and Erika Rimkutė. 2016. *Lithuanian Dependency Treebank ALKSNIS*. In: *I. Skadiņa and R. Rozis (Eds.): Human Language Technologies – The Baltic Perspective*. Amsterdam: IOS Press.
- Bosco, Cristina and Montemagni, Simonetta and Simi, Maria. 2013. *Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank*. Association for Computational Linguistics.
- Bouma, Gosse and van Noord, Gertjan. 2017. *Increasing Return on Annotation Investment: The Automatic Construction of a Universal Dependency Treebank for Dutch*. Association for Computational Linguistics.
- Chun, Jayeol and Han, Na-Rae and Hwang, Jena D. and Choi, Jinho D. 2018. *Building Universal Dependency Treebanks in Korean*. European Language Resources Association (ELRA).
- Dobrovoljc, Kaja and Erjavec, Tomaž and Krek, Simon. 2017. *The Universal Dependencies Treebank for Slovenian*. Association for Computational Linguistics.
- Kira Droganova and Olga Lyashevskaya and Daniel Zeman. 2018. *Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks*. Linköping University Electronic Press.
- Guillaume, Bruno and de Marneffe, Marie-Catherine and Perrier, Guy. 2019. *Conversion et améliorations de corpus du français annotés en Universal Dependencies [Conversion and Improvement of Universal Dependencies French corpora]*. ATALA (Association pour le Traitement Automatique des Langues).
- Hajič, Jan and Smrž, Otakar and Zemánek, Petr and Pajas, Petr and Šnaidauf, Jan and Beška, Emanuel and Kracmar, Jakub and Hassanová, Kamila. 2009. *Prague Arabic Dependency Treebank 1.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Hajič, Jan and others. 2024. *Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Anders Johannsen and Héctor Martínez Alonso and Barbara Plank. 2015. *Universal Dependencies for Danish*.
- Lynn, Teresa and Foster, Jennifer. 2016. *Universal Dependencies for Irish*. Association pour le Traitement Automatique des Langues. Dépendances universelles de l'irlandais.
- Marşan, Büşra and Akkurt, Salih Furkan and Şen, Muhammet and Gürbüz, Merve and Güngör, Onur and Özateş, Şaziye Betül and Üsküdarlı, Suzan and Özgür, Arzucan and Güngör, Tunga and Öztürk, Balkız. 2022. *Enhancements to the BOUN Treebank Reflecting the Agglutinative Nature of Turkish*.
- McDonald, Ryan and Nivre, Joakim and Quirmbach-Brundage, Yvonne and Goldberg, Yoav and Das, Dipanjan and Ganchev, Kuzman and Hall, Keith and Petrov, Slav and Zhang, Hao and Täckström, Oscar and Bedini, Claudia and Bertomeu Castelló, Núria and Lee, Jungmee. 2013. *Universal Dependency Annotation for Multilingual Parsing*. Association for Computational Linguistics.
- Verginica Barbu Mititelu. 2018. *Modern Syntactic Analysis of Romanian*. Publishing House of "Alexandru Ioan Cuza" University.
- Muischnek, Kadri and Müürisepp, Kaili and Puolakainen, Tiina. 2016. *Estonian Dependency Treebank: from Constraint Grammar tagset to Universal Dependencies*. European Language Resources Association (ELRA).
- Nivre, Joakim and Megyesi, Beáta. 2007. *Bootstrapping a Swedish Treebank Using Cross-Corpus Harmonization and Annotation Projection*.
- Petya Osenova and Kiril Simov. 2004. *BTB-TR05: BulTreeBank Stylebook 4 05*.
- Pretkalniundefineda, Lauma and Rituma, Laura and Saulundefinedte, Baiba. 2018. *Deriving Enhanced Universal Dependencies from a Hybrid Dependency-Constituency Treebank*. Springer-Verlag.
- Prokopidis, Prokopis and Papageorgiou, Haris. 2017. *Universal Dependencies for Greek*. Association for Computational Linguistics.
- Shishkina, Yana and Lyashevskaya, Olga. 2021. *Sculpting Enhanced Dependencies for Belarusian*. Springer-Verlag.

Natalia Silveira and Timothy Dozat and Marie-Catherine de Marneffe and Samuel Bowman and Miriam Connor and John Bauer and Christopher D. Manning. 2014. *A Gold Standard Dependency Corpus for English*.

Taulé, Mariona and Martí, M. Antònia and Recasens, Marta. 2008. *AnCorà: Multilevel Annotated Corpora for Catalan and Spanish*. European Language Resources Association (ELRA).

Wróblewska, Alina. 2018. *Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format*. Association for Computational Linguistics.

Zeman, Daniel. 2017. *Slovak Dependency Treebank in Universal Dependencies*.

## A. Settings

### A.1. Environment and Hyper-parameters

The environment and the hyper-parameters used in the experiments are shown in Table 4 and 5, respectively.

Hardware	
GPU	8 × NVIDIA H100 80 GB
CPU	Intel Xeon 208 core
RAM	1,360 GB
SSD	10 TB
Software	
OS	Ubuntu 22.04
CUDA	12.6
Python	3.12.11
PyTorch	2.7.0
Transformers	4.51.3
TRL	0.18.1
PEFT	0.15.2
vLLM	0.9.1

Table 4: Experimental environment.

Model Comparison Experiments	
Learning Rate	3.0e-4
Epochs	3
Intermediate Checkpoint Experiments	
Learning Rate	1.0e-4
The training epochs are adjusted based on the number of words in the training set to ensure that the amount of training for each language is equivalent.	

Table 5: Hyper-parameters used in LoRA fine-tuning.

### A.2. Model Specifications

The specifications of LLMs used in the experiments are shown in Table 6.

### B. Comparison between 3-step and 2-step-no-form prompts

The detailed results of the experiment for the comparison between 3-step and 2-step-no-form prompts are shown in Table 7 and 8.

	Model Parameters	Tokenizer Vocab Size	Context Length	Training Tokens	LoRA Trainable Parameters	Supported Languages
gemma-2-9b	9.27B	256k	8,192	8T	27.0M	(multilingual)
Llama-3.1-8B	8.05B	128k	131,072	15T	21.0M	en, de, fr, it, pt, hi, es, th
Qwen2.5-7B	7.64B	152k	131,072	18T	20.2M	zh, en, fr, es, pt, de, it, ru, ja, ko, vi, th, ar, and more
OLMo-2-7B	7.32B	100k	4,096	4T	20.0M	en

Table 6: Model specs.

	Token Recall								UPOS (F1)							
	3-step			2-step-no-form					3-step			2-step-no-form				
	gemma	Llama	Qwen	gemma	Llama	Qwen	OLMo	gemma	Llama	Qwen	gemma	Llama	Qwen	OLMo		
ar_padt	99.4	99.4	99.8	100.0	100.0	100.0	100.0	97.0	97.1	97.2	97.3	97.1	97.1	96.6		
be_hse	98.4	98.4	98.3	100.0	100.0	100.0	100.0	98.7	98.8	98.6	98.9	98.7	98.7	98.2		
bg_btb	99.8	100.0	99.5	100.0	100.0	100.0	100.0	99.3	99.4	99.0	99.4	99.4	99.3	98.9		
ca_ancora	99.4	99.6	99.5	100.0	100.0	100.0	100.0	98.9	99.0	99.0	99.2	99.2	99.1	98.9		
cs_pdt	99.9	99.6	99.9	100.0	100.0	100.0	100.0	99.3	99.3	99.3	99.4	99.3	99.3	99.1		
da_ddt	100.0	99.3	98.0	100.0	100.0	99.8	100.0	98.6	98.1	97.9	98.8	98.4	97.9	97.1		
de_gsd	100.0	99.9	99.5	100.0	100.0	100.0	99.8	97.0	97.0	97.2	97.3	97.3	97.2	96.8		
el_gdt	99.6	99.3	99.8	100.0	99.7	100.0	100.0	98.2	98.4	97.7	98.3	97.7	97.9	97.2		
en_ewt	100.0	99.9	99.4	100.0	99.9	100.0	100.0	98.6	98.3	98.3	98.4	98.4	98.5	98.0		
es_ancora	100.0	99.8	99.6	100.0	100.0	100.0	100.0	99.1	99.2	99.1	99.2	99.2	99.1	99.0		
et_edt	99.9	99.9	99.8	100.0	100.0	100.0	99.9	98.2	98.1	97.6	98.2	98.0	97.5	96.8		
fi_ftb	99.6	99.7	99.9	100.0	100.0	100.0	100.0	98.1	97.8	96.6	98.2	97.8	96.8	96.4		
fr_gsd	100.0	100.0	98.9	100.0	100.0	100.0	100.0	98.6	98.7	98.6	98.6	98.6	98.5	98.5		
ga_idt	99.0	98.8	99.1	100.0	100.0	100.0	99.8	95.6	96.0	95.8	96.3	96.4	95.8	95.1		
hr_set	99.6	99.3	99.2	100.0	100.0	100.0	100.0	98.7	98.5	98.3	98.7	98.7	98.4	97.9		
it_isdt	100.0	99.7	99.4	100.0	100.0	100.0	100.0	98.6	98.7	98.4	98.7	98.8	98.8	98.3		
ja_gsd	99.2	100.0	99.6	100.0	100.0	99.6	100.0	98.9	98.9	98.8	98.9	98.8	98.4	98.2		
ko_gsd	99.8	99.6	99.8	100.0	100.0	100.0	100.0	96.5	96.9	97.0	97.2	96.9	96.9	96.0		
lt_alksnis	99.5	98.2	97.2	100.0	100.0	100.0	100.0	97.4	96.3	95.1	97.3	96.2	95.4	91.8		
lv_lvtb	100.0	99.7	99.6	100.0	100.0	100.0	100.0	98.3	97.8	97.7	98.3	98.0	97.8	97.3		
nl_alpino	100.0	99.7	98.7	100.0	100.0	100.0	100.0	98.1	98.2	97.8	98.4	98.1	98.3	97.5		
no_bokmaal	99.9	99.9	99.2	100.0	99.9	100.0	100.0	98.7	98.7	98.5	98.8	98.7	98.6	98.2		
pl_pdb	99.9	99.4	99.4	100.0	100.0	100.0	100.0	99.4	99.2	99.2	99.4	99.3	99.3	98.7		
pt_gsd	99.9	99.8	99.9	100.0	100.0	100.0	100.0	98.1	97.9	97.9	98.2	98.2	97.8	97.5		
ro_rrt	99.8	99.7	99.3	100.0	100.0	100.0	100.0	98.1	98.2	97.9	98.3	98.3	98.1	97.7		
ru_syntagrus	100.0	99.8	99.9	100.0	100.0	100.0	100.0	98.8	98.8	99.0	99.0	98.9	98.9	98.7		
sk_snk	99.8	99.1	98.9	100.0	100.0	100.0	100.0	98.2	98.0	97.9	98.3	98.0	97.9	96.8		
sl_ssj	99.8	99.4	99.6	100.0	99.9	100.0	100.0	98.5	98.8	98.5	98.9	98.9	98.7	98.1		
sr_set	99.9	99.7	99.6	100.0	100.0	99.6	100.0	99.1	98.9	98.8	99.0	98.9	98.8	98.4		
sv_talbanken	99.8	99.6	99.1	100.0	100.0	100.0	100.0	98.9	98.6	98.3	98.9	98.8	98.5	97.9		
tr_boun	99.9	99.0	98.3	100.0	100.0	100.0	100.0	93.6	93.3	92.9	93.4	93.0	92.9	92.3		
uk_iu	99.7	98.7	98.1	100.0	99.8	99.9	99.8	98.5	98.2	98.0	98.6	98.0	98.0	96.3		
zh_gsdsimp	100.0	100.0	99.5	99.9	99.9	99.9	99.9	97.3	97.1	96.7	97.1	96.9	97.3	95.6		
Macro Avg.	99.7	99.5	99.3	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	98.1	98.1	97.8	<b>98.3</b>	98.1	97.9	97.3		

Table 7: Comparison of Token Recall and UPOS (F1) between 3-step and 2-step-no-form among the four models; gemma-2-9b, Llama-3.1-8B, Qwen2.5-7B, and OLMo-2-7B.

	Unlabeld Attachment Score (F1)							Labeld Attachment Score (F1)						
	3-step			2-step-no-form				3-step			2-step-no-form			
	gemma	Llama	Qwen	gemma	Llama	Qwen	OLMo	gemma	Llama	Qwen	gemma	Llama	Qwen	OLMo
ar_padt	90.4	90.4	90.1	90.7	90.2	89.9	87.2	85.9	86.0	85.6	86.4	85.8	85.7	82.2
be_hse	92.3	92.1	90.7	92.4	91.8	90.6	87.0	90.5	90.4	88.7	90.8	90.0	88.7	84.4
bg_btb	96.4	96.0	95.4	96.2	96.1	95.4	93.8	94.1	93.7	93.0	94.0	93.7	93.0	90.7
ca_ancora	94.5	95.0	94.6	95.3	95.2	94.9	93.4	92.9	93.6	93.1	94.0	93.8	93.4	91.6
cs_pdt	95.1	95.6	95.4	95.8	95.6	95.4	93.9	93.6	94.2	94.0	94.5	94.2	94.1	92.1
da_ddt	90.3	88.8	87.4	90.5	88.8	87.7	85.7	88.7	86.7	85.2	88.7	86.6	85.4	82.9
de_gsd	89.1	90.3	89.9	89.3	89.9	89.7	89.2	85.3	86.6	86.2	85.7	86.1	86.0	85.2
el_gdt	94.7	94.4	93.1	94.2	94.2	92.5	89.7	93.2	92.5	91.0	92.4	92.2	90.3	86.7
en_ewt	95.4	94.8	94.3	95.2	95.1	94.8	94.2	93.9	93.2	92.7	93.5	93.4	93.1	92.4
es_ancora	93.7	94.3	94.1	94.6	94.4	94.0	93.1	92.0	92.8	92.6	93.2	92.9	92.4	91.2
et_edt	92.5	91.3	89.9	92.1	90.9	89.5	86.4	90.4	89.0	87.2	89.9	88.6	86.8	83.3
fi_ftb	95.2	94.0	92.2	95.1	94.1	92.5	90.6	93.5	91.9	89.1	93.4	91.9	89.6	87.2
fr_gsd	95.6	95.9	95.7	96.0	95.8	95.9	94.5	94.0	94.2	94.2	94.6	94.4	94.4	92.8
ga_idt	88.4	87.7	86.3	88.2	88.2	86.6	85.6	83.2	82.3	80.0	82.7	82.7	80.6	78.8
hr_set	94.0	93.7	92.5	94.1	93.6	92.9	90.8	91.1	90.7	89.4	91.2	90.6	89.6	87.1
it_isdt	95.6	95.9	95.2	95.8	96.1	95.7	94.4	94.0	94.5	93.7	94.4	94.7	94.4	92.6
ja_gsd	95.3	94.9	95.0	95.1	94.9	94.3	93.1	94.3	94.0	93.8	93.9	93.5	92.9	91.5
ko_gsd	87.4	89.1	89.3	90.3	89.3	89.3	85.8	83.9	86.0	86.3	87.2	86.1	86.1	81.9
lt_alksnis	88.8	85.3	81.9	88.3	85.0	81.1	71.6	85.6	81.6	77.5	84.9	81.0	76.6	64.9
lv_lvrb	94.1	92.1	91.5	93.8	92.4	91.7	88.3	91.6	89.4	88.5	91.4	89.7	88.8	84.9
nl_alpino	94.0	94.8	94.1	95.3	94.6	94.3	91.8	91.3	92.7	91.8	93.4	92.5	92.1	88.6
no_bokmaal	94.7	95.2	94.2	95.6	95.1	94.6	93.2	93.5	94.0	93.0	94.6	94.0	93.4	91.5
pl_pdb	97.0	96.6	96.2	97.0	96.7	96.4	94.1	95.7	95.1	94.7	95.6	95.2	94.8	91.9
pt_gsd	94.0	93.7	93.4	94.0	93.9	93.5	92.6	92.1	91.6	91.1	92.1	92.0	91.4	90.0
ro_rrt	92.5	94.1	92.6	94.5	93.8	92.8	92.0	88.9	90.7	89.3	91.3	90.6	89.5	88.3
ru_syntagrus	95.2	95.5	95.6	95.8	95.5	95.6	94.7	93.1	93.5	93.7	93.9	93.5	93.6	92.5
sk_snk	96.2	95.3	94.6	95.8	94.8	94.6	89.6	94.9	93.8	92.9	94.4	93.3	92.9	86.8
sl_ssj	94.0	95.0	94.3	95.7	95.3	94.6	91.4	92.4	93.5	92.7	94.4	93.9	93.0	89.1
sr_set	95.1	94.5	93.8	95.2	94.2	94.1	91.6	93.0	92.2	91.3	92.9	91.7	91.4	88.7
sv_talbanken	94.4	93.1	92.6	94.3	93.6	92.8	90.6	92.5	90.9	90.1	92.3	91.4	90.6	87.1
tr_boun	83.0	82.1	80.4	83.2	81.9	80.8	78.0	76.6	75.5	73.5	76.7	75.2	73.9	70.8
uk_iu	94.6	93.6	92.8	94.1	92.8	92.4	86.3	92.8	91.5	90.5	92.2	90.5	89.9	82.8
zh_gsdsimp	89.1	88.2	88.2	88.1	88.6	89.1	83.0	86.7	85.5	85.6	85.5	85.8	86.5	79.4
Macro Avg.	93.1	92.8	92.0	<b>93.4</b>	92.8	92.1	89.6	90.8	90.4	89.5	<b>91.1</b>	90.3	89.5	86.4

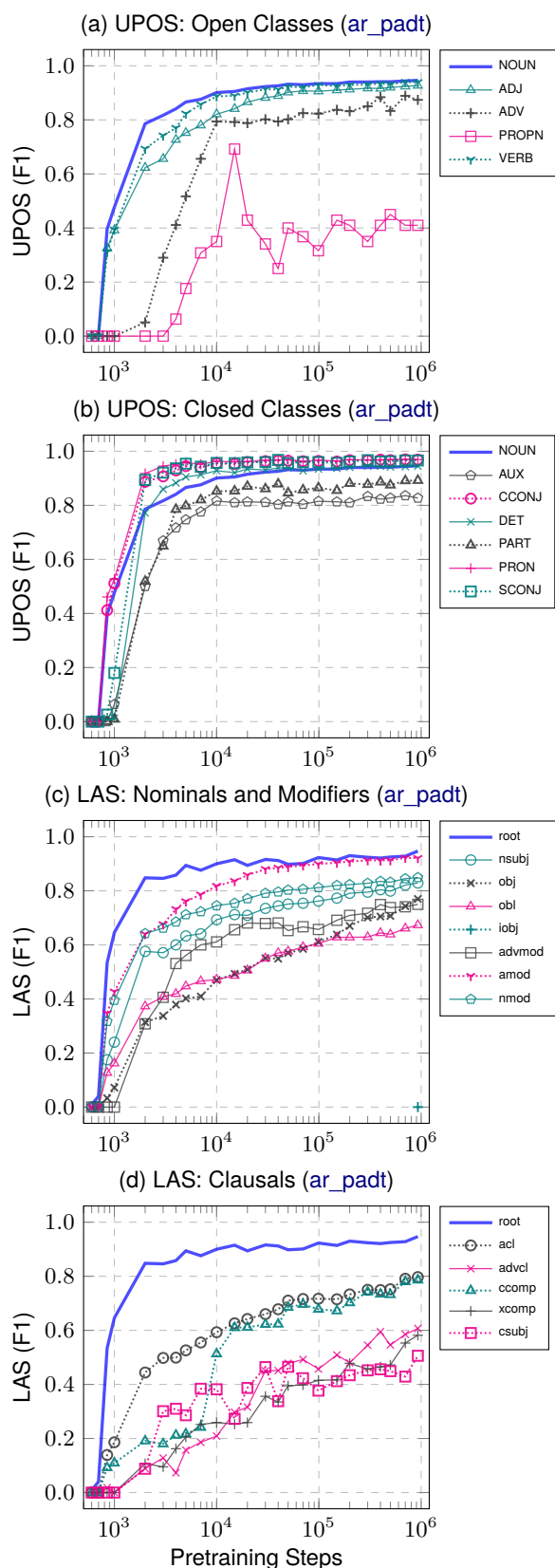
Table 8: Comparison of UAS (F1) and LAS (F1) between 3-step and 2-step-no-form among the four models; gemma-2-9b, Llama-3.1-8B, Qwen2.5-7B, and OLMo-2-7B.

## C. Fine-grained Analysis of Language-wise Syntactic Acquisition Process

In this section, we present the results of the analysis of the syntactic acquisition process, organized into subsections for each language. Each subsection includes the learning curves of major UPOS tags and DEPREL labels, as well as statistics of the treebank used in the experiments. The subsections are sorted in alphabetical order by language name. The directories categorized by language family are as follows.

- Germanic
  - Danish ([da\\_ddt](#)): C.8
  - Dutch ([nl\\_alpino](#)): C.9
  - English ([en\\_ewt](#)): C.10
  - German ([de\\_gsd](#)): C.14
  - Norwegian ([no\\_bokmaal](#)): C.22
  - Swedish ([sv\\_talbanken](#)): C.31
- Romance
  - Catalan ([ca\\_ancora](#)): C.4
  - French ([fr\\_gsd](#)): C.13
  - Italian ([it\\_isdt](#)): C.17
  - Portuguese ([pt\\_gsd](#)): C.24
  - Romanian ([ro\\_rrt](#)): C.25
  - Spanish ([es\\_ancora](#)): C.30
- Slavic
  - Belarusian ([be\\_hse](#)): C.2
  - Bulgarian ([bg\\_btb](#)): C.3
  - Croatian ([hr\\_set](#)): C.6
  - Czech ([cs\\_pdt](#)): C.7
  - Polish ([pl\\_pdb](#)): C.23
  - Russian ([ru\\_syntagrus](#)): C.26
  - Serbian ([sr\\_set](#)): C.27
  - Slovak ([sk\\_snk](#)): C.28
  - Slovenian ([sl\\_ssj](#)): C.29
  - Ukrainian ([uk\\_iu](#)): C.33
- Greek
  - Greek ([el\\_gdt](#)): C.15
- Finnic
  - Estonian ([et\\_edt](#)): C.11
  - Finnish ([fi\\_ftb](#)): C.12
- Celtic
  - Irish ([ga\\_idt](#)): C.16
- Baltic
  - Latvian ([lv\\_lvtb](#)): C.20
  - Lithuanian ([lt\\_alksnis](#)): C.21
- Turkic
  - Turkish ([tr\\_boun](#)): C.32
- Afro-Asiatic
  - Arabic ([ar\\_padt](#)): C.1
- Japanese
  - Japanese ([ja\\_gsd](#)): C.18
- Korean
  - Korean ([ko\\_gsd](#)): C.19
- Sino-Tibetan
  - Chinese ([zh\\_gsdsimp](#)): C.5

## C.1. Arabic

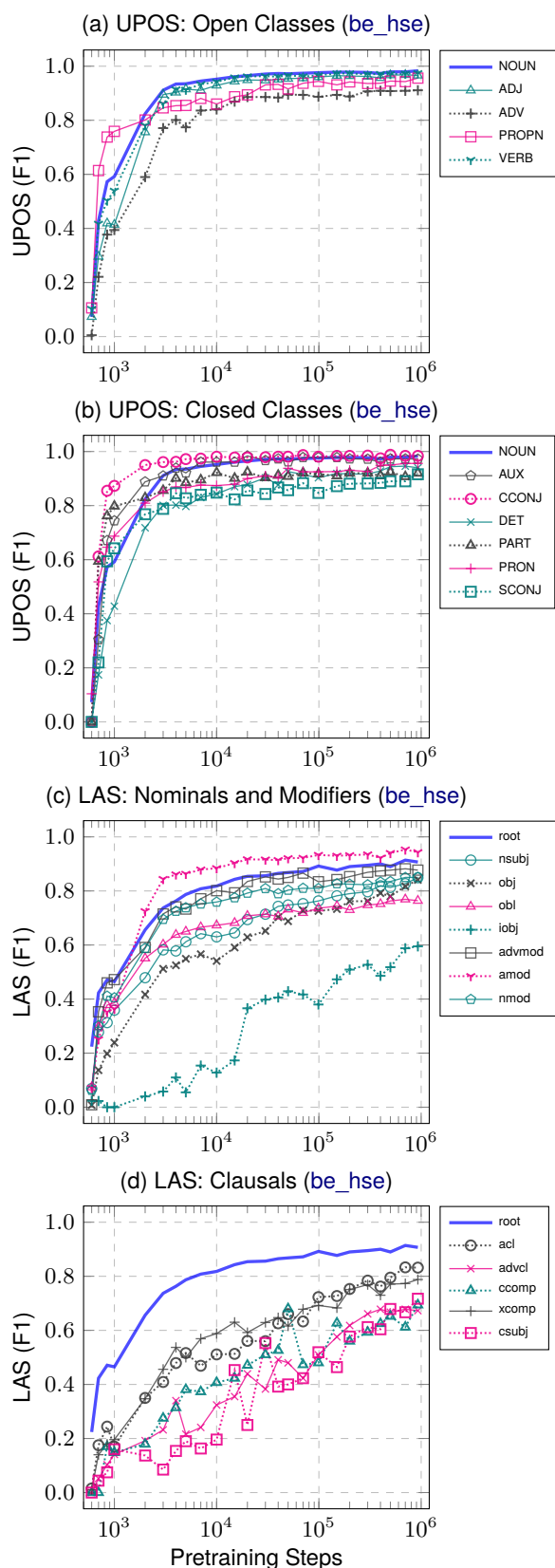


	Train	Dev	Test
Number of sentences	6,075	909	680
Number of words	223,881	30,239	28,264
Number of label occurrences			
<i>ADJ</i>	23,498	2,916	2,937
<i>ADP</i>	33,617	4,410	4,528
<i>ADV</i>	880	108	104
<i>AUX</i>	1,699	258	197
<i>CCONJ</i>	15,803	2,018	1,963
<i>DET</i>	4,648	625	623
<i>INTJ</i>	7	1	0
<i>NOUN</i>	74,546	9,612	9,547
<i>NUM</i>	6,010	969	779
<i>PART</i>	1,709	270	226
<i>PRON</i>	8,533	1,211	1,133
<i>PROPN</i>	187	27	31
<i>PUNCT</i>	17,511	2,882	2,052
<i>SCONJ</i>	4,368	555	534
<i>SYM</i>	329	18	41
<i>VERB</i>	16,789	2,318	2,189
<i>X</i>	13,747	2,041	1,380
<i>acl</i>	4,193	577	567
<i>advcl</i>	1,302	185	208
<i>advmod</i>	2,546	333	337
<i>amod</i>	19,555	2,419	2,447
<i>appos</i>	617	137	68
<i>aux</i>	1,333	214	147
<i>case</i>	31,889	4,171	4,279
<i>cc</i>	11,646	1,349	1,447
<i>ccomp</i>	2,537	356	319
<i>conj</i>	10,640	1,303	1,223
<i>cop</i>	366	44	50
<i>csubj</i>	442	51	42
<i>dep</i>	1,131	331	104
<i>det</i>	1,707	222	221
<i>discourse</i>	1	0	0
<i>dislocated</i>	172	18	13
<i>fixed</i>	2,175	267	263
<i>flat</i>	596	120	29
<i>iobj</i>	89	12	6
<i>mark</i>	6,087	791	764
<i>nmod</i>	50,215	6,501	6,351
<i>nsubj</i>	14,683	2,035	1,890
<i>nummod</i>	2,939	422	382
<i>obj</i>	6,710	882	879
<i>obl</i>	20,786	2,783	2,786
<i>orphan</i>	174	17	5
<i>parataxis</i>	4,275	663	514
<i>punct</i>	17,506	2,882	2,052
<i>root</i>	6,075	909	680
<i>xcomp</i>	1,494	245	191

Table 9: Statistics of *ar\_padt*

Figure 9: Syntactic acquisition through pretraining of OLMo-2-7B on Arabic (*ar\_padt*).

## C.2. Belarusian

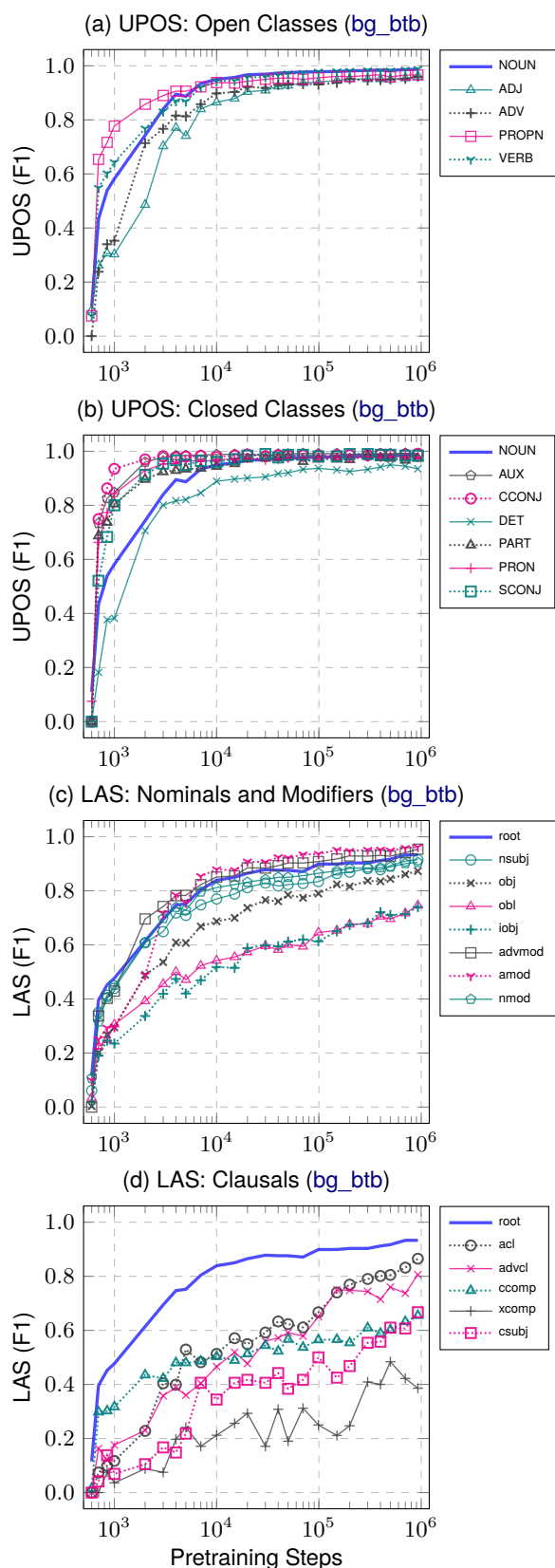


	Train	Dev	Test
Number of sentences	22,852	1,301	1,077
Number of words	273,181	15,931	15,997
Number of label occurrences			
<i>ADJ</i>	23,293	1,630	1,893
<i>ADP</i>	26,423	1,605	1,643
<i>ADV</i>	10,148	495	397
<i>AUX</i>	1,893	101	90
<i>CCONJ</i>	7,922	437	641
<i>DET</i>	6,093	288	349
<i>INTJ</i>	86	3	1
<i>NOUN</i>	63,986	4,080	4,620
<i>NUM</i>	5,326	287	233
<i>PART</i>	4,534	219	134
<i>PRON</i>	9,479	477	367
<i>PROPN</i>	18,323	1,114	946
<i>PUNCT</i>	51,192	3,069	2,754
<i>SCONJ</i>	3,146	161	95
<i>SYM</i>	2,454	88	70
<i>VERB</i>	28,254	1,469	1,369
<i>X</i>	10,629	408	395
<i>acl</i>	3,195	180	239
<i>advcl</i>	1,308	64	56
<i>advmod</i>	12,652	620	426
<i>amod</i>	18,585	1,290	1,510
<i>appos</i>	5,327	318	398
<i>aux</i>	990	62	50
<i>case</i>	26,582	1,596	1,630
<i>cc</i>	7,973	438	628
<i>ccomp</i>	1,296	61	25
<i>compound</i>	537	29	12
<i>conj</i>	11,866	767	984
<i>cop</i>	828	38	35
<i>csubj</i>	828	48	31
<i>dep</i>	5,629	158	175
<i>det</i>	5,212	227	306
<i>discourse</i>	269	13	38
<i>dislocated</i>	14	0	0
<i>expl</i>	196	10	1
<i>fixed</i>	771	80	158
<i>flat</i>	5,582	370	310
<i>goeswith</i>	1	1	0
<i>iobj</i>	2,269	112	82
<i>list</i>	1,704	64	42
<i>mark</i>	2,774	141	82
<i>nmod</i>	24,364	1,598	2,044
<i>nsubj</i>	18,057	1,012	866
<i>nummod</i>	3,219	204	124
<i>obj</i>	10,183	544	476
<i>obl</i>	17,276	1,042	1,011
<i>orphan</i>	219	23	13
<i>parataxis</i>	6,278	290	255
<i>punct</i>	51,174	3,069	2,754
<i>reparandum</i>	5	0	0
<i>root</i>	22,853	1,301	1,077
<i>vocative</i>	155	2	11
<i>xcomp</i>	3,010	159	148

Table 10: Statistics of *be\_hse*

Figure 10: Syntactic acquisition through pretraining of OLMo-2-7B on Belarusian (*be\_hse*).

### C.3. Bulgarian

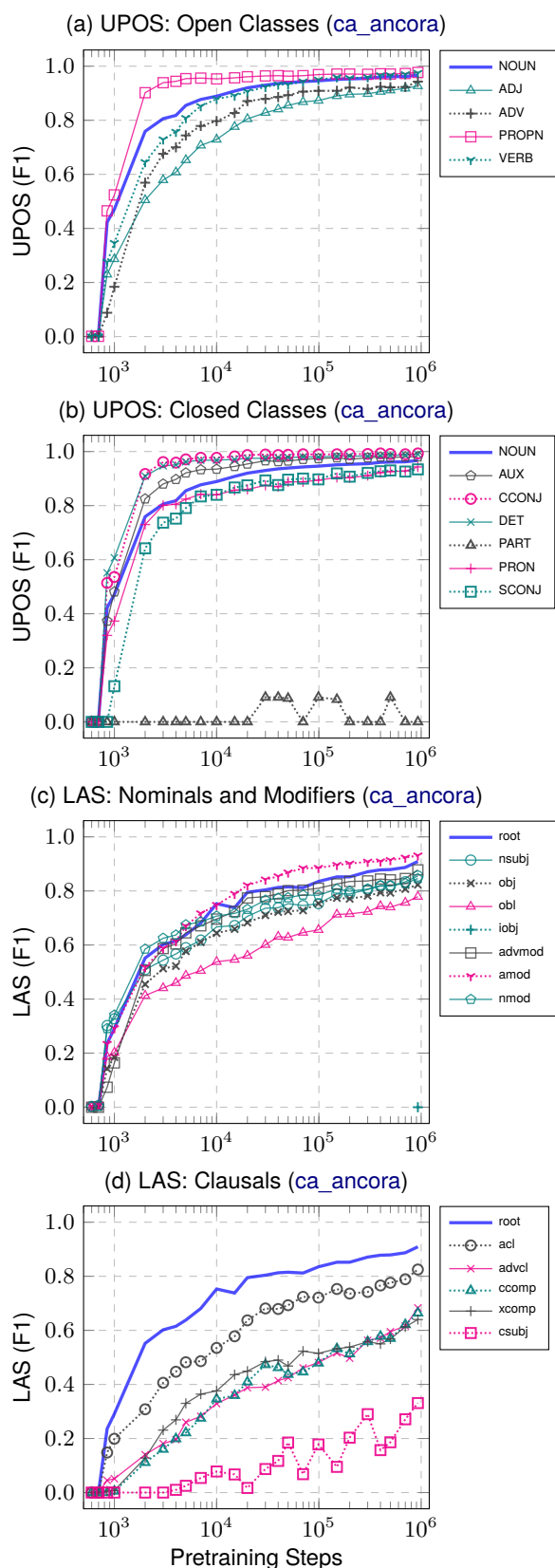


	Train	Dev	Test
Number of sentences	8,907	1,115	1,116
Number of words	124,336	16,089	15,724
Number of label occurrences			
<i>ADJ</i>	10,813	1,401	1,377
<i>ADP</i>	17,572	2,287	2,237
<i>ADV</i>	5,250	637	671
<i>AUX</i>	7,255	963	916
<i>CCONJ</i>	3,885	508	468
<i>DET</i>	1,872	288	273
<i>INTJ</i>	26	1	1
<i>NOUN</i>	27,147	3,519	3,486
<i>NUM</i>	1,668	215	222
<i>PART</i>	1,728	229	210
<i>PRON</i>	8,052	1,061	981
<i>PROPN</i>	6,797	833	805
<i>PUNCT</i>	17,523	2,267	2,269
<i>SCONJ</i>	1,276	174	156
<i>VERB</i>	13,470	1,706	1,652
<i>X</i>	2	0	0
<i>acl</i>	1,282	159	161
<i>advcl</i>	1,276	173	164
<i>advmod</i>	5,691	710	718
<i>amod</i>	9,400	1,240	1,195
<i>appos</i>	28	2	4
<i>aux</i>	4,994	646	585
<i>case</i>	17,210	2,225	2,182
<i>cc</i>	3,891	507	468
<i>ccomp</i>	1,625	298	266
<i>compound</i>	2	0	0
<i>conj</i>	4,415	593	562
<i>cop</i>	1,899	216	270
<i>csubj</i>	358	50	43
<i>det</i>	2,699	401	362
<i>discourse</i>	557	66	68
<i>expl</i>	2,676	389	320
<i>fixed</i>	560	96	83
<i>flat</i>	1,424	136	135
<i>iobj</i>	2,818	399	346
<i>mark</i>	1,460	200	179
<i>nmod</i>	11,374	1,509	1,527
<i>nsubj</i>	9,177	1,167	1,157
<i>nummod</i>	1,385	175	169
<i>obj</i>	5,366	642	713
<i>obl</i>	5,195	661	622
<i>parataxis</i>	426	4	10
<i>punct</i>	17,523	2,267	2,269
<i>root</i>	8,907	1,115	1,116
<i>vocative</i>	55	1	2
<i>xcomp</i>	663	42	28

Table 11: Statistics of *bg\_btB*

Figure 11: Syntactic acquisition through pretraining of OLMo-2-7B on Bulgarian (*bg\_btB*).

## C.4. Catalan

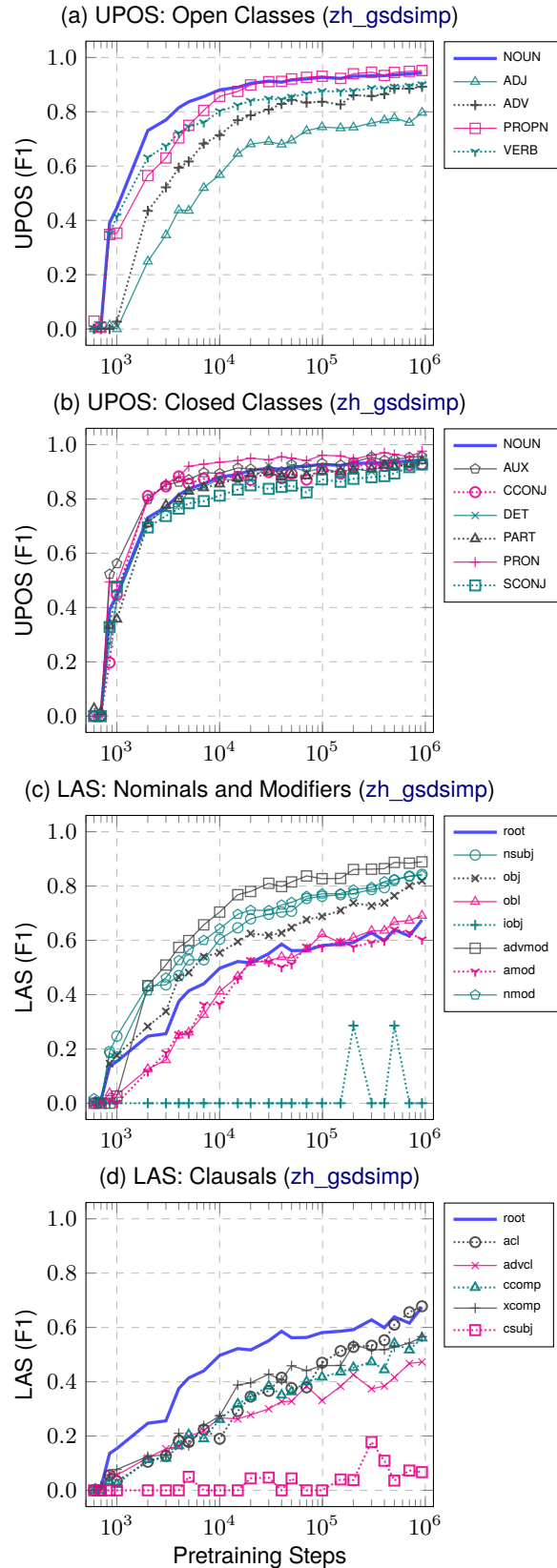


	Train	Dev	Test
Number of sentences	13,123	1,709	1,846
Number of words	429,578	58,073	59,610
Number of label occurrences			
<i>ADJ</i>	23,607	3,218	3,257
<i>ADP</i>	69,040	9,329	9,607
<i>ADV</i>	12,149	1,733	1,600
<i>AUX</i>	17,370	2,382	2,300
<i>CCONJ</i>	11,797	1,647	1,605
<i>DET</i>	68,331	9,369	9,567
<i>INTJ</i>	11	1	3
<i>NOUN</i>	77,539	10,453	10,654
<i>NUM</i>	7,917	917	1,128
<i>PART</i>	113	13	21
<i>PRON</i>	18,498	2,469	2,487
<i>PROPN</i>	36,245	4,782	5,563
<i>PUNCT</i>	44,968	5,961	6,341
<i>SCONJ</i>	8,366	1,185	999
<i>SYM</i>	648	68	105
<i>VERB</i>	32,979	4,546	4,373
<i>acl</i>	7,866	1,081	1,058
<i>advcl</i>	5,381	788	684
<i>advmod</i>	11,961	1,695	1,564
<i>amod</i>	19,615	2,653	2,735
<i>appos</i>	5,960	784	934
<i>aux</i>	13,153	1,837	1,760
<i>case</i>	61,122	8,251	8,560
<i>cc</i>	11,656	1,609	1,612
<i>ccomp</i>	3,370	437	385
<i>compound</i>	2,092	229	331
<i>conj</i>	12,992	1,747	1,786
<i>cop</i>	3,749	463	467
<i>csubj</i>	736	91	101
<i>dep</i>	71	11	7
<i>det</i>	67,526	9,258	9,450
<i>discourse</i>	1	0	0
<i>dislocated</i>	1	0	1
<i>expl</i>	459	54	59
<i>fixed</i>	7,746	1,079	1,047
<i>flat</i>	13,816	1,825	2,373
<i>mark</i>	12,083	1,705	1,466
<i>nmod</i>	32,522	4,304	4,342
<i>nsubj</i>	21,624	2,901	2,962
<i>nummod</i>	4,912	622	737
<i>obj</i>	23,997	3,350	3,209
<i>obl</i>	23,449	3,096	3,304
<i>parataxis</i>	438	69	71
<i>punct</i>	44,968	5,961	6,341
<i>root</i>	13,123	1,709	1,846
<i>vocative</i>	1	0	0
<i>xcomp</i>	3,188	464	418

Table 12: Statistics of *ca\_ancora*

Figure 12: Syntactic acquisition through pretraining of OLMo-2-7B on Catalan (*ca\_ancora*).

## C.5. Chinese

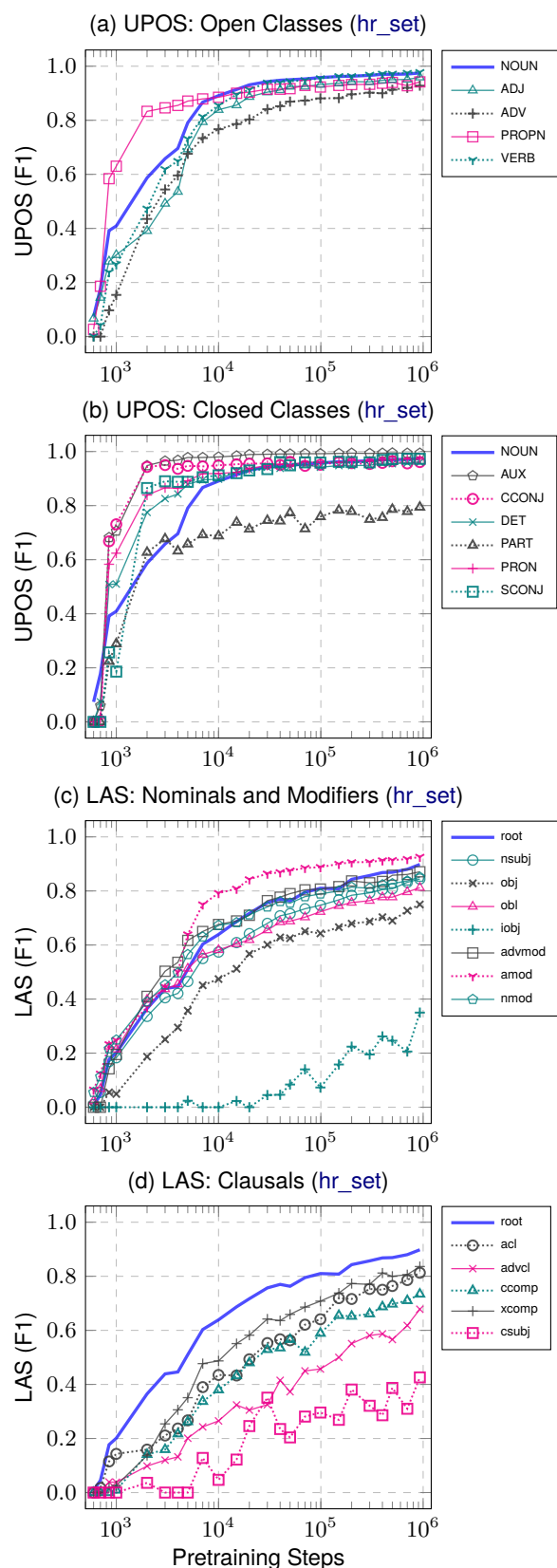


	Train	Dev	Test
Number of sentences	3,997	500	500
Number of words	98,616	12,663	12,012
Number of label occurrences			
<i>ADJ</i>	2,446	312	272
<i>ADP</i>	4,539	555	535
<i>ADV</i>	2,467	322	292
<i>AUX</i>	3,105	365	422
<i>CCONJ</i>	1,385	164	191
<i>DET</i>	1,071	119	139
<i>NOUN</i>	27,097	3,637	3,312
<i>NUM</i>	5,353	654	653
<i>PART</i>	7,847	1,041	994
<i>PRON</i>	1,433	175	168
<i>PROPN</i>	8,678	1,056	1,006
<i>PUNCT</i>	13,627	1,770	1,691
<i>SCONJ</i>	4,137	560	505
<i>SYM</i>	11	3	0
<i>VERB</i>	14,623	1,851	1,745
<i>X</i>	797	79	87
<i>acl</i>	2,034	268	252
<i>advcl</i>	3,232	447	343
<i>advmod</i>	2,552	329	304
<i>amod</i>	1,659	196	190
<i>appos</i>	1,122	128	103
<i>aux</i>	1,812	192	248
<i>case</i>	7,287	919	862
<i>cc</i>	1,391	164	191
<i>ccomp</i>	1,739	218	203
<i>clf</i>	1,814	200	233
<i>compound</i>	4,402	598	590
<i>conj</i>	2,558	329	341
<i>cop</i>	1,285	172	173
<i>csubj</i>	296	47	37
<i>det</i>	1,217	137	155
<i>discourse</i>	158	24	15
<i>dislocated</i>	47	5	10
<i>flat</i>	1,291	142	108
<i>iobj</i>	63	8	6
<i>mark</i>	5,652	749	680
<i>nmod</i>	14,198	1,975	1,653
<i>nsubj</i>	7,736	966	999
<i>nummod</i>	5,006	614	623
<i>obj</i>	6,175	819	754
<i>obl</i>	2,848	347	316
<i>orphan</i>	0	2	0
<i>parataxis</i>	1,829	235	258
<i>punct</i>	13,627	1,770	1,691
<i>reparandum</i>	1	0	0
<i>root</i>	3,997	500	500
<i>vocative</i>	1	0	0
<i>xcomp</i>	1,587	163	174

Table 13: Statistics of *zh\_gsdsimp*

Figure 13: Syntactic acquisition through pretraining of OLMo-2-7B on Chinese (*zh\_gsdsimp*).

## C.6. Croatian

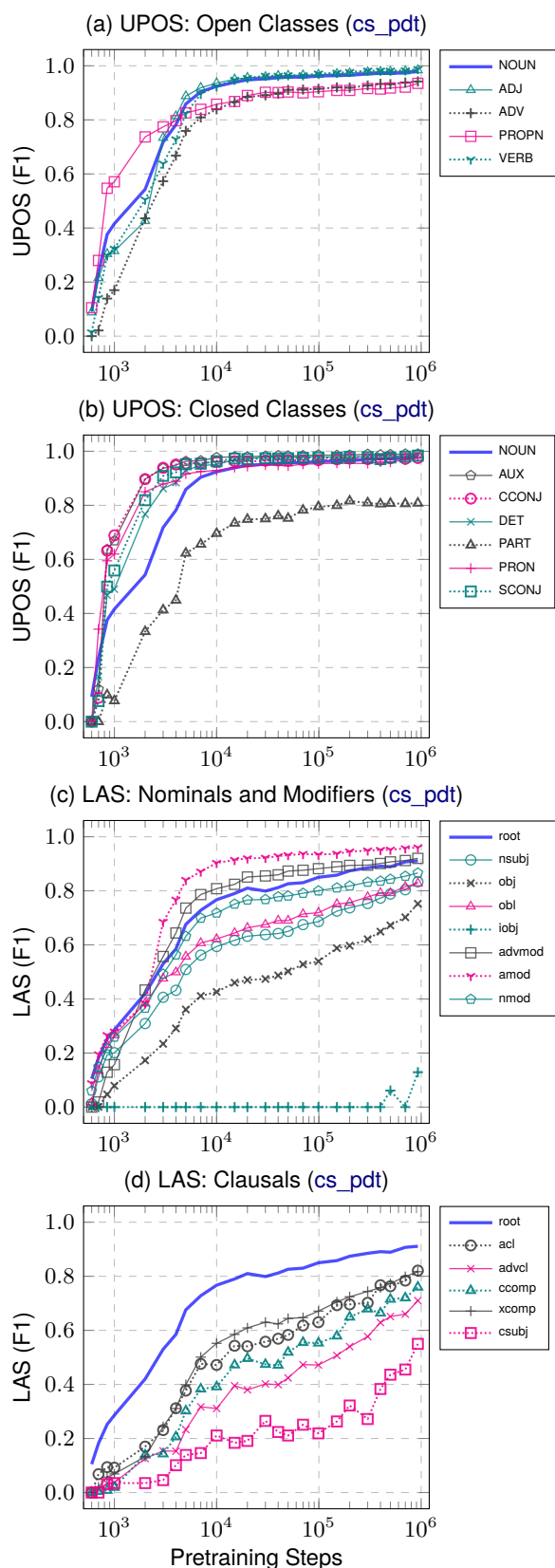


	Train	Dev	Test
Number of sentences	6,914	960	1,136
Number of words	152,857	22,292	24,260
Number of label occurrences			
<i>ADJ</i>	18,515	2,754	2,870
<i>ADP</i>	14,651	2,123	2,315
<i>ADV</i>	6,552	930	946
<i>AUX</i>	9,613	1,437	1,516
<i>CCONJ</i>	6,285	907	950
<i>DET</i>	5,980	839	875
<i>INTJ</i>	12	0	0
<i>NOUN</i>	36,963	5,445	6,169
<i>NUM</i>	2,335	460	353
<i>PART</i>	1,651	226	215
<i>PRON</i>	4,128	564	610
<i>PROPN</i>	9,817	1,413	1,618
<i>PUNCT</i>	18,444	2,684	3,037
<i>SCONJ</i>	3,811	529	589
<i>SYM</i>	94	23	2
<i>VERB</i>	13,382	1,886	2,119
<i>X</i>	624	72	76
<i>acl</i>	2,956	443	452
<i>advcl</i>	1,395	192	198
<i>advmod</i>	5,980	843	830
<i>amod</i>	14,959	2,225	2,355
<i>appos</i>	760	109	130
<i>aux</i>	6,374	1,035	1,037
<i>case</i>	14,846	2,159	2,364
<i>cc</i>	5,899	837	887
<i>ccomp</i>	1,374	178	230
<i>compound</i>	8	1	1
<i>conj</i>	7,333	1,017	1,134
<i>cop</i>	2,750	353	415
<i>csubj</i>	271	49	40
<i>dep</i>	7	1	0
<i>det</i>	2,893	409	429
<i>discourse</i>	1,606	232	208
<i>dislocated</i>	7	0	0
<i>expl</i>	1,795	263	302
<i>fixed</i>	704	91	100
<i>flat</i>	4,182	684	693
<i>iobj</i>	538	67	77
<i>list</i>	23	1	0
<i>mark</i>	3,160	421	471
<i>nmod</i>	14,328	2,032	2,437
<i>nsubj</i>	10,731	1,619	1,725
<i>nummod</i>	2,010	380	304
<i>obj</i>	6,745	983	1,072
<i>obl</i>	9,687	1,427	1,530
<i>orphan</i>	90	39	13
<i>parataxis</i>	1,735	261	300
<i>punct</i>	18,443	2,684	3,037
<i>root</i>	6,914	960	1,136
<i>vocative</i>	20	0	3
<i>xcomp</i>	2,334	297	350

Table 14: Statistics of *hr\_set*

Figure 14: Syntactic acquisition through pretraining of OLMo-2-7B on Croatian (*hr\_set*).

## C.7. Czech

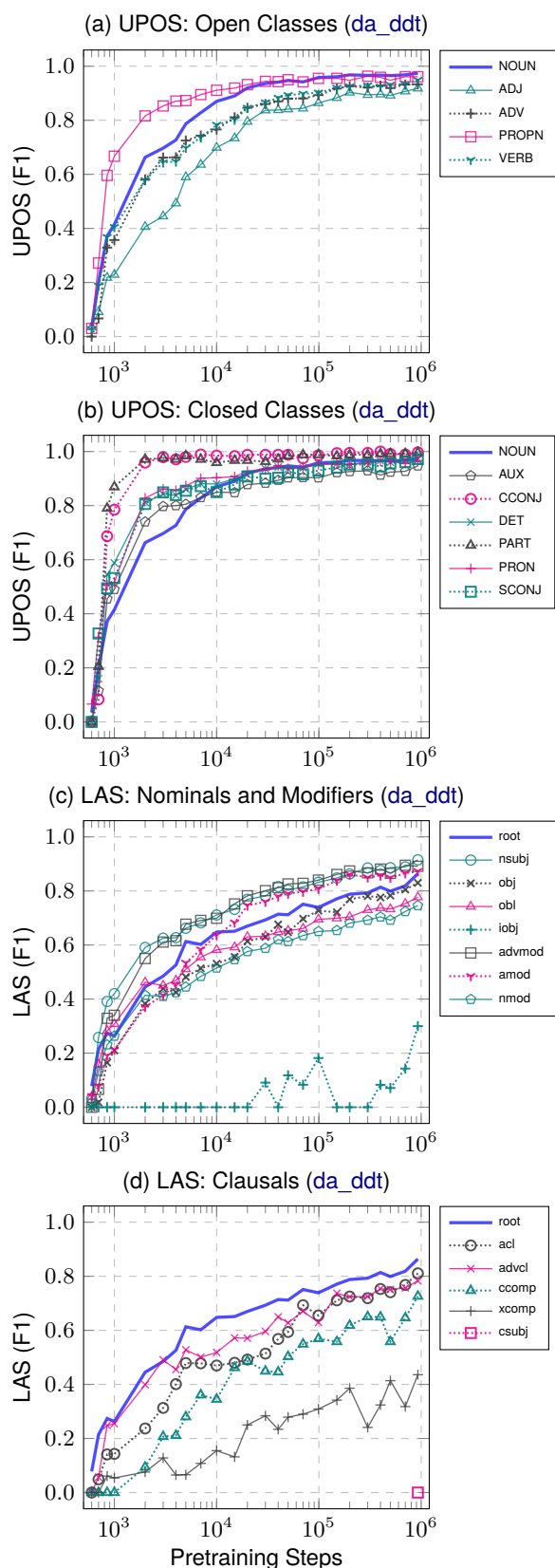


	Train	Dev	Test
Number of sentences	68,491	9,270	10,146
Number of words	1,173,285	159,283	173,918
Number of label occurrences			
<i>ADJ</i>	144,444	19,742	21,025
<i>ADP</i>	113,566	15,420	16,523
<i>ADV</i>	58,687	7,960	8,850
<i>AUX</i>	35,834	5,105	5,648
<i>CCONJ</i>	45,058	6,309	6,721
<i>DET</i>	43,380	6,156	6,642
<i>INTJ</i>	69	5	6
<i>NOUN</i>	292,282	39,508	43,665
<i>NUM</i>	34,897	4,307	4,951
<i>PART</i>	8,696	1,318	1,340
<i>PRON</i>	34,651	4,855	5,241
<i>PROPN</i>	58,837	7,584	8,157
<i>PUNCT</i>	172,565	22,909	25,409
<i>SCONJ</i>	21,800	3,021	3,432
<i>SYM</i>	989	157	222
<i>VERB</i>	100,925	13,999	15,220
<i>X</i>	6,605	928	866
<i>acl</i>	16,875	2,344	2,567
<i>advcl</i>	8,179	1,130	1,346
<i>advmod</i>	63,379	8,765	9,699
<i>amod</i>	122,414	16,662	17,679
<i>appos</i>	6,119	824	827
<i>aux</i>	15,987	2,325	2,477
<i>case</i>	113,328	15,379	16,479
<i>cc</i>	41,616	5,843	6,214
<i>ccomp</i>	8,526	1,147	1,282
<i>compound</i>	3,261	409	344
<i>conj</i>	53,491	7,228	7,722
<i>cop</i>	19,074	2,661	3,055
<i>csubj</i>	4,870	658	680
<i>dep</i>	9,887	1,181	1,572
<i>det</i>	23,797	3,352	3,691
<i>discourse</i>	290	47	45
<i>expl</i>	17,222	2,359	2,505
<i>fixed</i>	3,820	513	564
<i>flat</i>	17,829	2,408	2,486
<i>iobj</i>	455	72	66
<i>mark</i>	21,765	3,036	3,432
<i>nmod</i>	115,178	15,492	17,151
<i>nsubj</i>	73,237	9,901	10,973
<i>nummod</i>	22,349	2,745	3,125
<i>obj</i>	41,524	5,824	6,277
<i>obl</i>	89,975	12,250	13,414
<i>orphan</i>	2,054	306	277
<i>parataxis</i>	1,499	154	233
<i>punct</i>	172,551	22,906	25,402
<i>root</i>	68,491	9,270	10,146
<i>vocative</i>	56	13	12
<i>xcomp</i>	14,187	2,079	2,176

Table 15: Statistics of *cs\_pdt*

Figure 15: Syntactic acquisition through pretraining of OLMo-2-7B on Czech (*cs\_pdt*).

## C.8. Danish

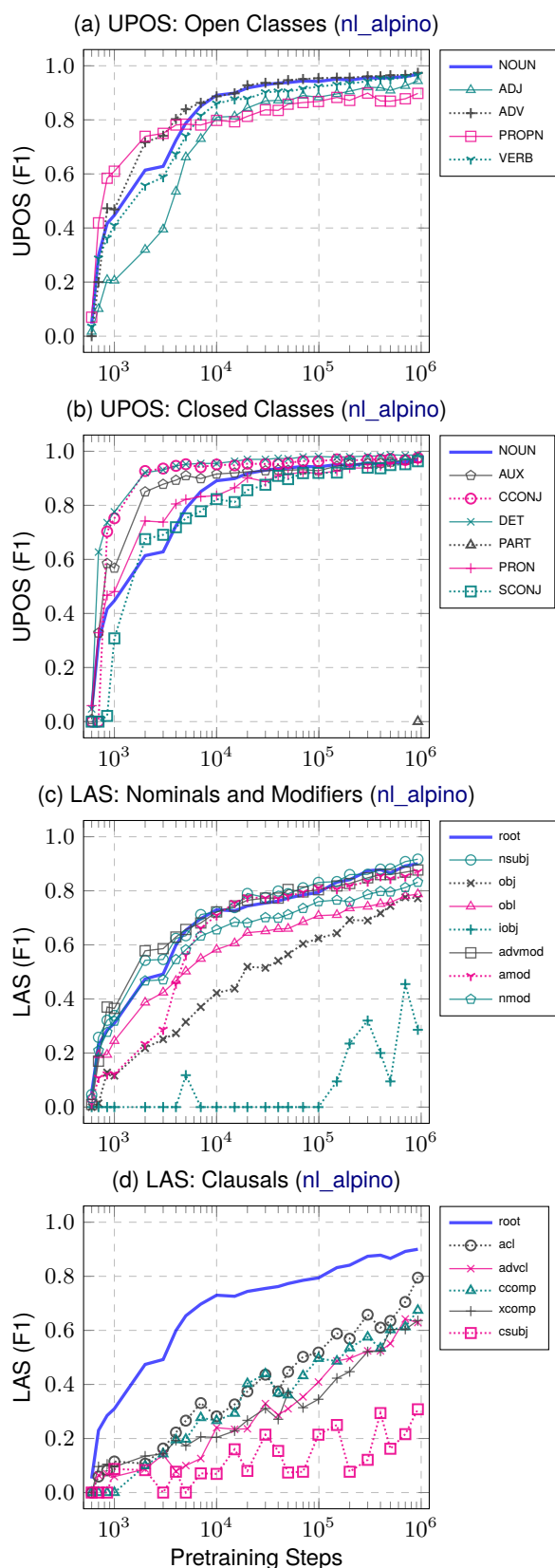


	Train	Dev	Test
Number of sentences	4,383	564	565
Number of words	80,378	10,332	10,023
Number of label occurrences			
ADJ	5,246	675	638
ADP	9,413	1,188	1,145
ADV	6,474	818	778
AUX	3,742	500	488
CCONJ	2,693	367	323
DET	4,396	609	499
INTJ	75	11	8
NOUN	14,956	1,947	1,823
NUM	1,212	140	153
PART	956	137	122
PRON	5,759	720	723
PROPN	3,940	488	550
PUNCT	11,111	1,381	1,444
SCONJ	1,427	180	183
SYM	48	0	6
VERB	8,632	1,148	1,118
X	298	23	22
<i>acl</i>	1,446	187	183
<i>advcl</i>	1,301	173	156
<i>advmod</i>	5,856	728	715
<i>amod</i>	4,618	583	545
<i>appos</i>	288	34	38
<i>aux</i>	2,487	342	325
<i>case</i>	7,947	1,013	953
<i>cc</i>	2,514	346	302
<i>ccomp</i>	461	64	79
<i>compound</i>	321	40	35
<i>conj</i>	2,809	378	346
<i>cop</i>	1,255	158	163
<i>dep</i>	405	50	30
<i>det</i>	4,373	607	496
<i>discourse</i>	25	4	3
<i>dislocated</i>	1	0	1
<i>expl</i>	368	34	39
<i>fixed</i>	390	40	62
<i>flat</i>	1,305	151	188
<i>iobj</i>	120	22	15
<i>list</i>	164	18	17
<i>mark</i>	3,518	461	449
<i>nmod</i>	5,026	616	567
<i>nsubj</i>	7,180	949	945
<i>nummod</i>	1,016	119	113
<i>obj</i>	4,037	521	525
<i>obl</i>	5,063	672	650
<i>orphan</i>	11	0	3
<i>parataxis</i>	2	3	3
<i>punct</i>	11,110	1,379	1,444
<i>reparandum</i>	2	0	0
<i>root</i>	4,383	564	565
<i>vocative</i>	23	5	4
<i>xcomp</i>	553	71	64

Table 16: Statistics of *da\_ddt*

Figure 16: Syntactic acquisition through pretraining of OLMo-2-7B on Danish (*da\_ddt*).

## C.9. Dutch

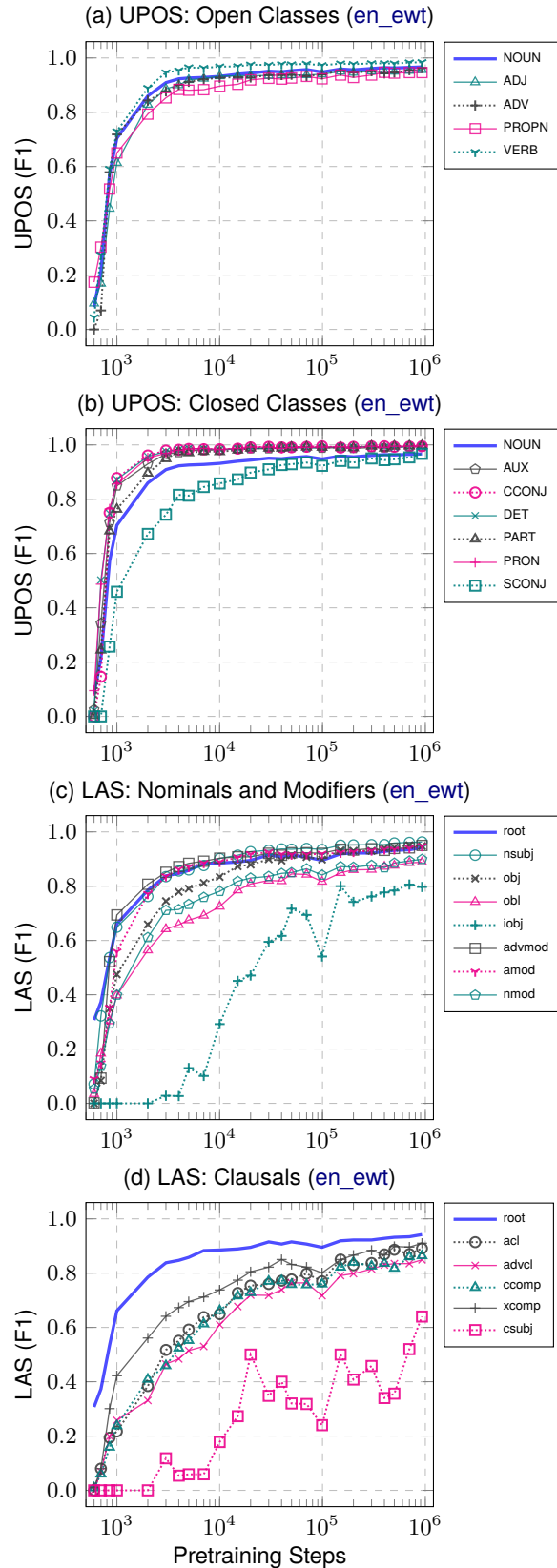


	Train	Dev	Test
Number of sentences	12,289	718	596
Number of words	186,027	11,541	11,046
Number of label occurrences			
<i>ADJ</i>	12,111	761	761
<i>ADP</i>	23,314	1,523	1,463
<i>ADV</i>	11,483	629	601
<i>AUX</i>	8,881	408	527
<i>CCONJ</i>	3,907	253	337
<i>DET</i>	22,561	1,314	1,348
<i>INTJ</i>	94	9	0
<i>NOUN</i>	30,855	1,809	2,090
<i>NUM</i>	3,107	286	156
<i>PRON</i>	12,052	639	451
<i>PROPN</i>	13,609	1,119	660
<i>PUNCT</i>	20,040	1,382	1,191
<i>SCONJ</i>	3,056	163	191
<i>SYM</i>	453	35	57
<i>VERB</i>	19,676	1,152	1,144
<i>X</i>	828	59	69
<i>acl</i>	2,934	114	128
<i>advcl</i>	2,080	109	116
<i>advmod</i>	10,607	607	570
<i>amod</i>	9,025	561	593
<i>appos</i>	2,071	155	115
<i>aux</i>	6,233	278	393
<i>case</i>	19,145	1,267	1,232
<i>cc</i>	3,538	226	326
<i>ccomp</i>	1,224	73	89
<i>compound</i>	1,734	136	103
<i>conj</i>	4,312	271	425
<i>cop</i>	2,640	130	133
<i>csubj</i>	400	14	21
<i>det</i>	22,055	1,289	1,317
<i>expl</i>	810	40	33
<i>fixed</i>	3,033	212	266
<i>flat</i>	4,807	321	230
<i>iobj</i>	509	23	13
<i>mark</i>	5,574	303	318
<i>nmod</i>	11,151	691	768
<i>nsubj</i>	16,148	966	823
<i>nummod</i>	1,657	151	80
<i>obj</i>	6,239	375	274
<i>obl</i>	11,641	856	651
<i>orphan</i>	68	1	3
<i>parataxis</i>	1,345	123	115
<i>punct</i>	20,040	1,382	1,191
<i>root</i>	12,289	718	596
<i>xcomp</i>	2,718	149	124

Table 17: Statistics of *nl\_alpino*

Figure 17: Syntactic acquisition through pretraining of OLMo-2-7B on Dutch (*nl\_alpino*).

## C.10. English

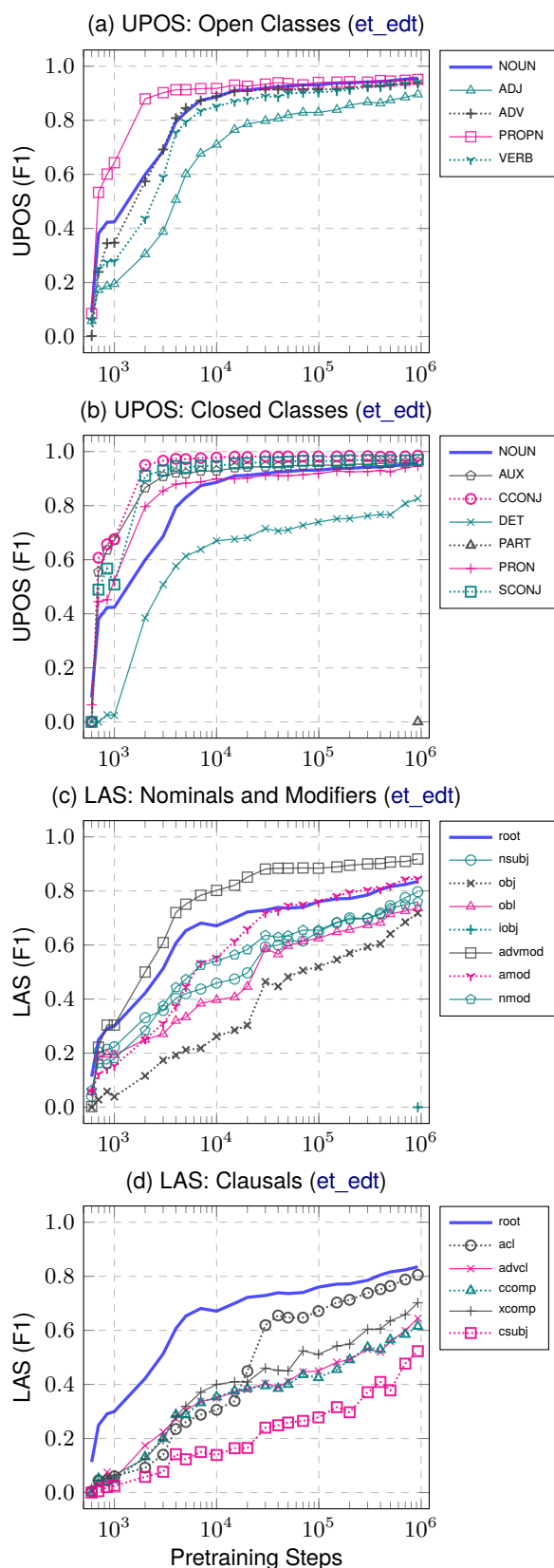


	Train	Dev	Test
Number of sentences	12,544	2,001	2,077
Number of words	204,579	25,149	25,094
Number of label occurrences			
<i>ADJ</i>	13,186	1,873	1,794
<i>ADP</i>	17,739	2,038	2,030
<i>ADV</i>	10,117	1,224	1,183
<i>AUX</i>	12,815	1,567	1,543
<i>CCONJ</i>	6,687	779	736
<i>DET</i>	16,299	1,900	1,896
<i>INTJ</i>	695	115	121
<i>NOUN</i>	34,755	4,212	4,123
<i>NUM</i>	4,127	383	542
<i>PART</i>	5,748	647	649
<i>PRON</i>	18,677	2,225	2,165
<i>PROPN</i>	12,618	1,865	2,076
<i>PUNCT</i>	23,596	3,075	3,096
<i>SCONJ</i>	3,822	397	384
<i>SYM</i>	722	83	109
<i>VERB</i>	22,577	2,707	2,605
<i>X</i>	399	59	42
<i>acl</i>	3,403	380	376
<i>advcl</i>	3,942	372	367
<i>advmod</i>	11,262	1,344	1,315
<i>amod</i>	9,638	1,334	1,253
<i>appos</i>	1,544	172	179
<i>aux</i>	7,859	911	939
<i>case</i>	17,423	2,014	1,969
<i>cc</i>	6,837	798	755
<i>ccomp</i>	2,107	200	223
<i>compound</i>	8,120	990	1,103
<i>conj</i>	7,572	916	861
<i>cop</i>	4,717	623	576
<i>csubj</i>	293	38	25
<i>dep</i>	3	2	0
<i>det</i>	15,953	1,851	1,860
<i>discourse</i>	813	127	126
<i>dislocated</i>	5	1	0
<i>expl</i>	601	87	68
<i>fixed</i>	544	58	64
<i>flat</i>	1,785	266	335
<i>goeswith</i>	130	26	15
<i>iobj</i>	649	74	71
<i>list</i>	444	75	251
<i>mark</i>	7,126	757	752
<i>nmod</i>	11,028	1,250	1,202
<i>nsubj</i>	17,377	2,136	2,074
<i>nummod</i>	2,430	264	254
<i>obj</i>	9,902	1,213	1,153
<i>obl</i>	10,117	1,145	1,165
<i>orphan</i>	26	2	1
<i>parataxis</i>	1,571	249	232
<i>punct</i>	23,556	3,061	3,065
<i>reparandum</i>	35	9	4
<i>root</i>	12,544	2,001	2,077
<i>vocative</i>	139	22	21
<i>xcomp</i>	3,084	381	363

Table 18: Statistics of *en\_ewt*

Figure 18: Syntactic acquisition through pretraining of OLMo-2-7B on English (*en\_ewt*).

## C.11. Estonian

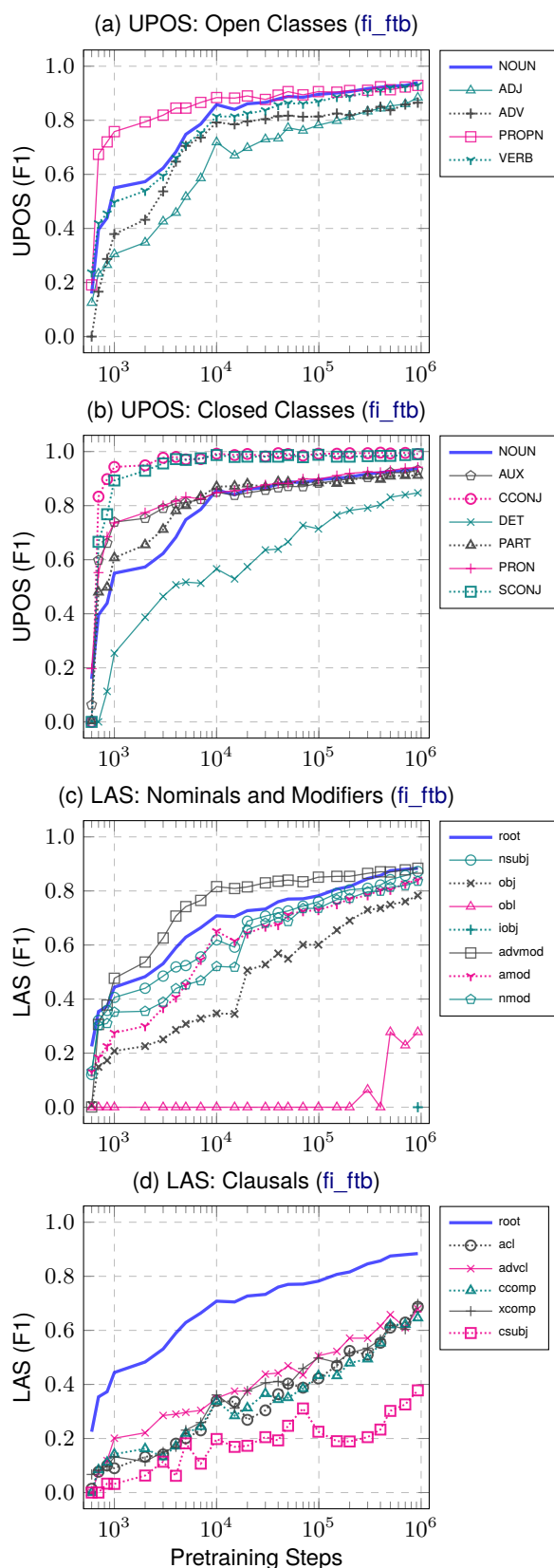


	Train	Dev	Test
Number of sentences	24,601	3,122	3,207
Number of words	344,581	44,742	48,465
Number of label occurrences			
<i>ADJ</i>	28,769	3,999	4,102
<i>ADP</i>	7,485	896	923
<i>ADV</i>	33,058	4,193	4,887
<i>AUX</i>	17,493	2,250	2,583
<i>CCONJ</i>	12,403	1,680	2,008
<i>DET</i>	5,646	688	850
<i>INTJ</i>	256	21	52
<i>NOUN</i>	91,404	11,665	12,616
<i>NUM</i>	7,099	1,077	840
<i>PART</i>	2	0	0
<i>PRON</i>	17,853	2,453	2,500
<i>PROPN</i>	20,712	2,521	3,047
<i>PUNCT</i>	56,395	7,498	7,672
<i>SCONJ</i>	6,737	860	1,112
<i>SYM</i>	599	115	26
<i>VERB</i>	38,015	4,779	5,184
<i>X</i>	655	47	63
<i>acl</i>	9,657	1,227	1,452
<i>advcl</i>	5,381	716	784
<i>advmod</i>	26,103	3,295	3,949
<i>amod</i>	17,373	2,529	2,373
<i>appos</i>	2,871	389	415
<i>aux</i>	9,187	1,154	1,376
<i>case</i>	7,463	894	920
<i>cc</i>	12,765	1,734	2,053
<i>ccomp</i>	3,473	384	478
<i>compound</i>	3,915	485	535
<i>conj</i>	17,311	2,432	2,436
<i>cop</i>	8,301	1,096	1,206
<i>csubj</i>	1,546	193	243
<i>dep</i>	9	0	9
<i>det</i>	5,636	686	844
<i>discourse</i>	239	18	49
<i>fixed</i>	355	56	35
<i>flat</i>	3,815	547	618
<i>goeswith</i>	238	12	21
<i>list</i>	8	2	1
<i>mark</i>	7,671	969	1,268
<i>nmod</i>	30,616	3,903	4,315
<i>nsubj</i>	28,226	3,585	3,764
<i>nummod</i>	4,696	690	555
<i>obj</i>	16,860	2,046	2,439
<i>obl</i>	31,177	3,907	4,138
<i>orphan</i>	479	72	51
<i>parataxis</i>	3,496	513	609
<i>punct</i>	56,395	7,498	7,672
<i>root</i>	24,601	3,122	3,207
<i>vocative</i>	120	18	13
<i>xcomp</i>	4,598	570	637

Table 19: Statistics of *et\_edt*

Figure 19: Syntactic acquisition through pretraining of OLMo-2-7B on Estonian (*et\_edt*).

## C.12. Finnish

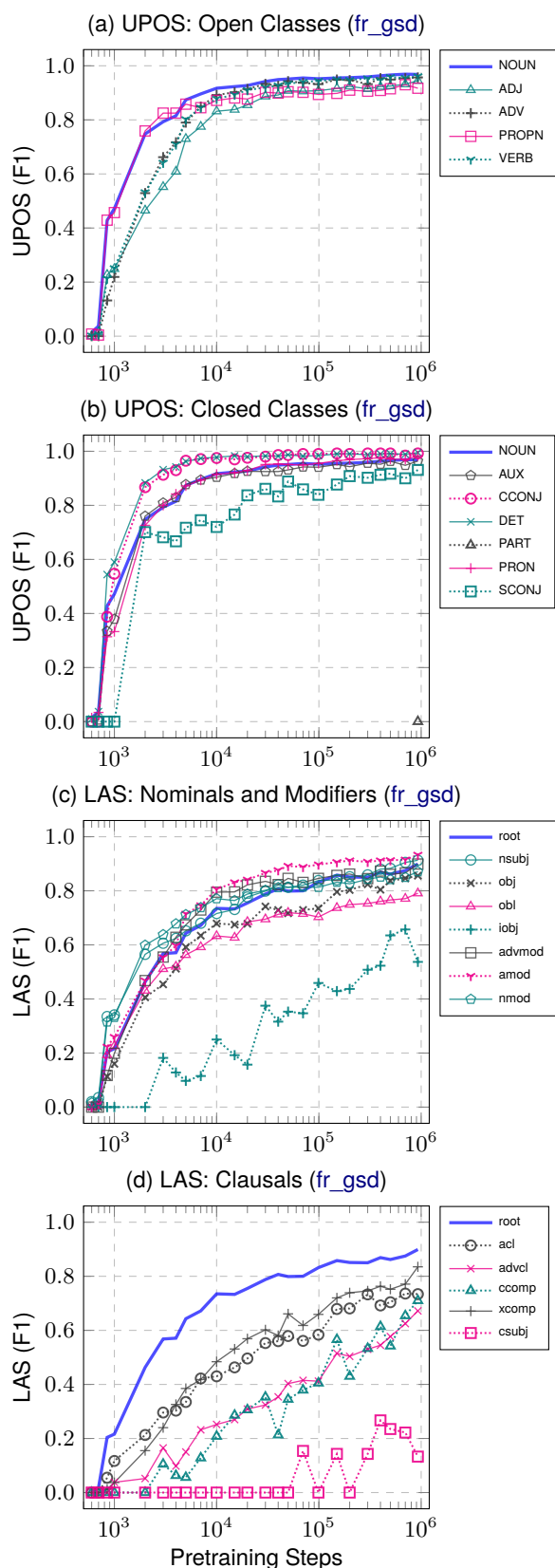


	Train	Dev	Test
Number of sentences	14,981	1,875	1,867
Number of words	127,602	15,724	16,286
Number of label occurrences			
<i>ADJ</i>	8,327	1,056	1,106
<i>ADP</i>	2,203	254	274
<i>ADV</i>	8,089	1,020	1,023
<i>AUX</i>	8,318	1,071	1,017
<i>CCONJ</i>	3,815	453	511
<i>DET</i>	2,940	353	398
<i>INTJ</i>	270	39	32
<i>NOUN</i>	29,650	3,595	3,788
<i>NUM</i>	1,826	223	203
<i>PART</i>	4,536	596	550
<i>PRON</i>	8,818	1,153	1,095
<i>PROPN</i>	5,482	610	680
<i>PUNCT</i>	18,062	2,190	2,313
<i>SCONJ</i>	3,270	431	445
<i>SYM</i>	17	4	1
<i>VERB</i>	21,728	2,639	2,827
<i>X</i>	251	37	23
<i>acl</i>	2,364	291	324
<i>advcl</i>	2,687	340	365
<i>advmod</i>	11,539	1,464	1,412
<i>amod</i>	5,500	695	698
<i>appos</i>	2	0	0
<i>aux</i>	5,456	711	674
<i>case</i>	2,162	253	272
<i>cc</i>	3,774	450	506
<i>ccomp</i>	2,007	244	261
<i>compound</i>	613	59	70
<i>conj</i>	5,230	600	668
<i>cop</i>	2,859	360	343
<i>csubj</i>	402	53	58
<i>dep</i>	229	37	25
<i>det</i>	2,765	333	381
<i>discourse</i>	240	33	35
<i>expl</i>	405	60	59
<i>fixed</i>	425	63	67
<i>flat</i>	682	61	75
<i>goeswith</i>	7	0	0
<i>mark</i>	3,514	469	490
<i>nmod</i>	18,989	2,340	2,396
<i>nsubj</i>	12,595	1,537	1,570
<i>nummod</i>	1,459	162	156
<i>obj</i>	6,861	833	967
<i>obl</i>	143	15	25
<i>orphan</i>	3	0	0
<i>punct</i>	18,062	2,190	2,313
<i>reparandum</i>	17	1	1
<i>root</i>	14,981	1,875	1,867
<i>vocative</i>	119	24	16
<i>xcomp</i>	1,511	171	192

Table 20: Statistics of *fi\_ftb*

Figure 20: Syntactic acquisition through pretraining of OLMo-2-7B on Finnish (*fi\_ftb*).

### C.13. French

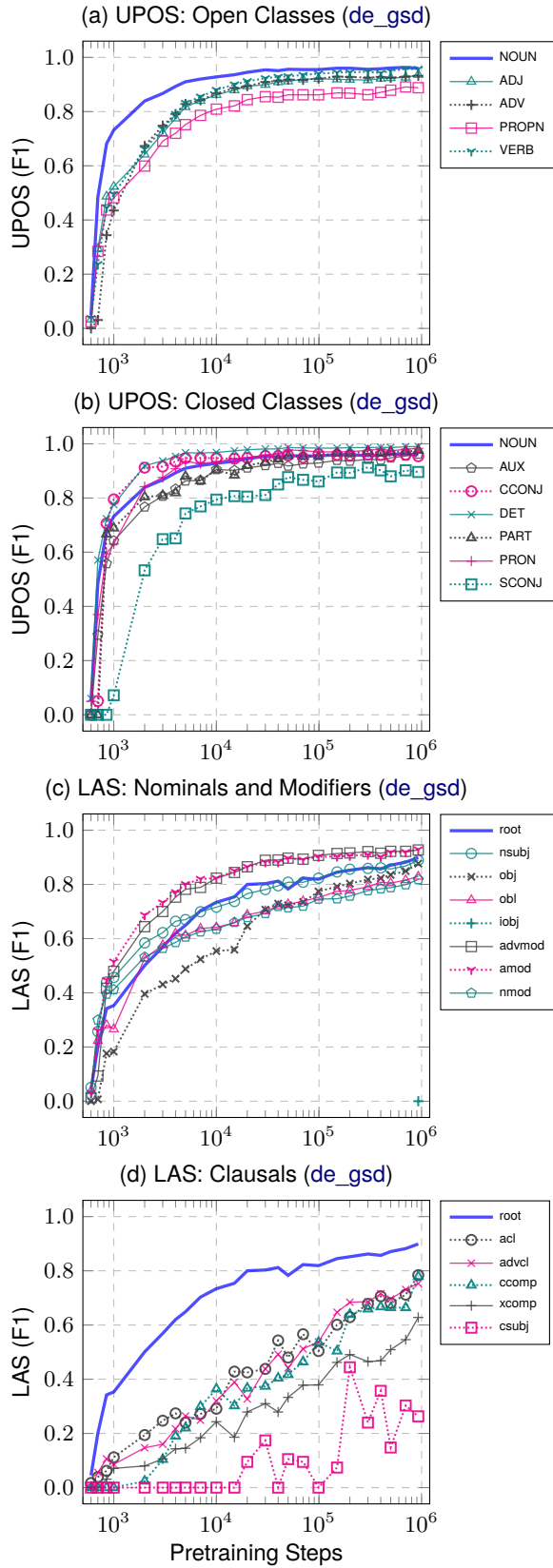


	Train	Dev	Test
Number of sentences	14,450	1,476	416
Number of words	354,652	35,721	10,018
Number of label occurrences			
<i>ADJ</i>	20,999	2,212	609
<i>ADP</i>	56,363	5,729	1,479
<i>ADV</i>	12,430	1,225	487
<i>AUX</i>	11,586	1,136	359
<i>CCONJ</i>	9,303	932	249
<i>DET</i>	54,109	5,506	1,479
<i>INTJ</i>	70	6	9
<i>NOUN</i>	66,624	6,786	1,870
<i>NUM</i>	9,325	968	229
<i>PRON</i>	16,050	1,576	559
<i>PROPN</i>	24,680	2,551	487
<i>PUNCT</i>	39,017	3,802	1,186
<i>SCONJ</i>	2,600	253	128
<i>SYM</i>	616	62	39
<i>VERB</i>	28,189	2,763	821
<i>X</i>	2,691	214	28
<i>acl</i>	7,302	715	178
<i>advcl</i>	3,094	315	117
<i>advmod</i>	12,133	1,188	462
<i>amod</i>	17,692	1,890	485
<i>appos</i>	5,536	574	111
<i>aux</i>	6,651	649	206
<i>case</i>	51,383	5,250	1,309
<i>cc</i>	9,515	955	255
<i>ccomp</i>	1,276	102	55
<i>compound</i>	6	0	0
<i>conj</i>	12,714	1,201	319
<i>cop</i>	4,914	485	153
<i>csubj</i>	195	18	12
<i>dep</i>	44	4	0
<i>det</i>	53,723	5,462	1,459
<i>discourse</i>	39	3	7
<i>dislocated</i>	62	6	8
<i>expl</i>	2,487	260	98
<i>fixed</i>	3,041	323	105
<i>flat</i>	7,201	749	127
<i>goeswith</i>	29	5	3
<i>iobj</i>	813	73	34
<i>mark</i>	5,908	569	244
<i>nmod</i>	31,534	3,210	848
<i>nsubj</i>	20,208	2,000	589
<i>nummod</i>	3,190	339	110
<i>obj</i>	12,047	1,183	368
<i>obl</i>	23,303	2,423	578
<i>orphan</i>	179	17	11
<i>parataxis</i>	981	86	50
<i>punct</i>	39,017	3,802	1,186
<i>root</i>	14,450	1,476	416
<i>vocative</i>	14	1	0
<i>xcomp</i>	3,971	388	115

Table 21: Statistics of *fr\_gsd*

Figure 21: Syntactic acquisition through pretraining of OLMo-2-7B on French (*fr\_gsd*).

## C.14. German

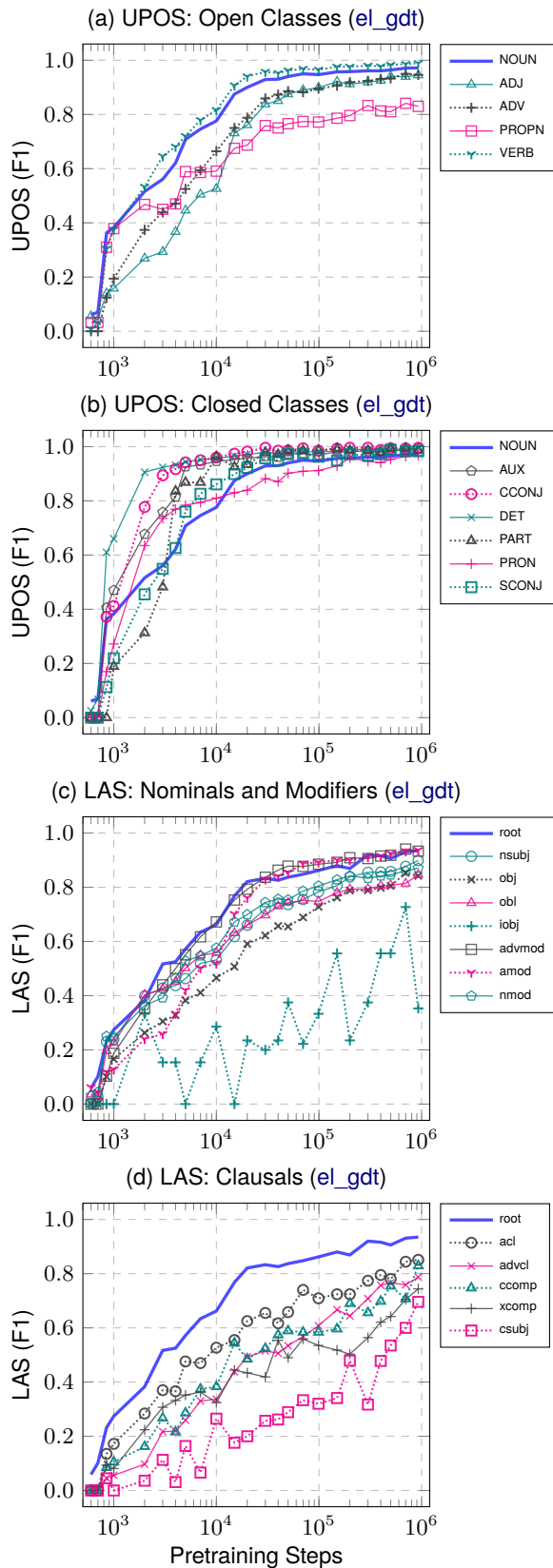


	Train	Dev	Test
Number of sentences	13,814	799	977
Number of words	263,791	12,480	16,498
Number of label occurrences			
<i>ADJ</i>	19,489	1,001	1,249
<i>ADP</i>	29,374	1,125	1,606
<i>ADV</i>	11,900	922	1,062
<i>AUX</i>	9,333	641	691
<i>CCONJ</i>	7,962	417	462
<i>DET</i>	37,305	1,639	2,262
<i>INTJ</i>	3	1	4
<i>NOUN</i>	46,986	2,200	3,111
<i>NUM</i>	6,929	175	233
<i>PART</i>	1,789	150	210
<i>PRON</i>	9,433	707	705
<i>PROPN</i>	28,785	612	1,022
<i>PUNCT</i>	34,453	1,672	2,365
<i>SCONJ</i>	1,451	115	161
<i>SYM</i>	93	3	4
<i>VERB</i>	18,237	1,081	1,326
<i>X</i>	269	19	25
<i>acl</i>	2,475	87	115
<i>advcl</i>	1,273	80	117
<i>advmod</i>	15,082	1,140	1,350
<i>amod</i>	14,790	534	793
<i>appos</i>	6,567	123	240
<i>aux</i>	5,769	428	509
<i>case</i>	28,090	1,070	1,524
<i>cc</i>	7,930	403	449
<i>ccomp</i>	612	110	150
<i>compound</i>	2,537	204	306
<i>conj</i>	10,932	471	524
<i>cop</i>	3,517	210	176
<i>csubj</i>	214	10	16
<i>dep</i>	605	80	206
<i>det</i>	36,095	1,541	2,044
<i>discourse</i>	7	1	3
<i>expl</i>	579	17	16
<i>fixed</i>	108	12	8
<i>flat</i>	7,808	128	223
<i>goeswith</i>	2	0	0
<i>mark</i>	2,671	190	258
<i>nmod</i>	20,237	568	996
<i>nsubj</i>	18,234	1,064	1,251
<i>nummod</i>	2,623	145	147
<i>obj</i>	7,248	502	587
<i>obl</i>	17,339	756	971
<i>orphan</i>	19	0	0
<i>parataxis</i>	483	30	56
<i>punct</i>	34,453	1,672	2,364
<i>reparandum</i>	3	0	0
<i>root</i>	13,814	799	977
<i>vocative</i>	2	2	3
<i>xcomp</i>	1,673	103	119

Table 22: Statistics of *de\_gsd*

Figure 22: Syntactic acquisition through pretraining of OLMo-2-7B on German (*de\_gsd*).

### C.15. Greek

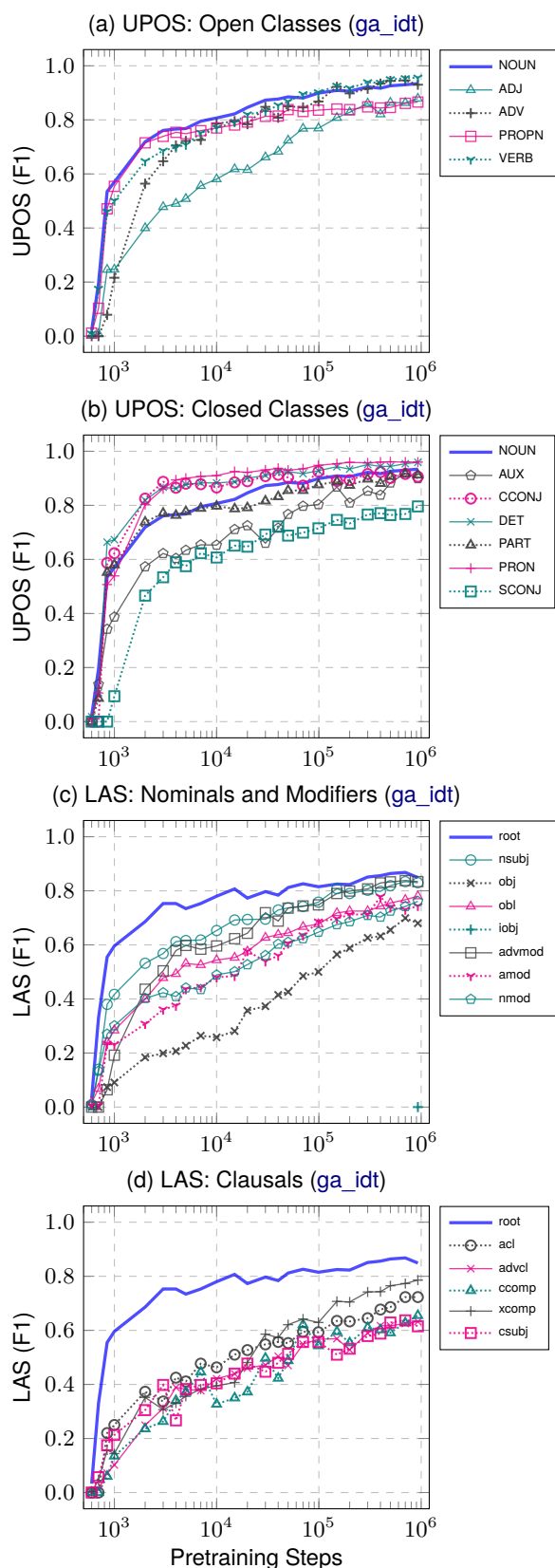


	Train	Dev	Test
Number of sentences	1,662	403	456
Number of words	42,326	10,443	10,672
Number of label occurrences			
<i>ADJ</i>	3,308	918	883
<i>ADP</i>	3,531	918	864
<i>ADV</i>	1,913	376	448
<i>AUX</i>	2,141	373	505
<i>CCONJ</i>	1,363	324	363
<i>DET</i>	7,957	2,146	2,016
<i>NOUN</i>	8,934	2,306	2,304
<i>NUM</i>	484	152	156
<i>PART</i>	312	41	70
<i>PRON</i>	1,584	342	367
<i>PROPN</i>	1,145	370	325
<i>PUNCT</i>	4,187	1,039	1,093
<i>SCONJ</i>	738	126	178
<i>SYM</i>	1	0	0
<i>VERB</i>	4,042	851	973
<i>X</i>	686	161	127
<i>acl</i>	865	228	210
<i>advcl</i>	564	106	142
<i>advmod</i>	1,953	355	474
<i>amod</i>	2,827	824	748
<i>appos</i>	177	49	38
<i>aux</i>	1,759	272	389
<i>case</i>	3,575	936	863
<i>cc</i>	1,345	321	360
<i>ccomp</i>	491	70	98
<i>compound</i>	5	1	16
<i>conj</i>	1,397	364	371
<i>cop</i>	382	101	116
<i>csubj</i>	144	28	40
<i>dep</i>	1	0	0
<i>det</i>	7,953	2,143	2,016
<i>discourse</i>	2	0	0
<i>expl</i>	15	1	4
<i>fixed</i>	67	14	8
<i>flat</i>	461	99	104
<i>iobj</i>	59	2	9
<i>mark</i>	817	143	202
<i>nmod</i>	4,247	1,166	1,038
<i>nsubj</i>	2,358	593	628
<i>nummod</i>	293	83	90
<i>obj</i>	1,642	330	410
<i>obl</i>	2,516	659	615
<i>orphan</i>	27	7	8
<i>parataxis</i>	60	15	15
<i>punct</i>	4,187	1,039	1,093
<i>root</i>	1,662	403	456
<i>vocative</i>	57	7	10
<i>xcomp</i>	418	84	101

Table 23: Statistics of *el\_gdt*

Figure 23: Syntactic acquisition through pretraining of OLMo-2-7B on Greek (*el\_gdt*).

## C.16. Irish

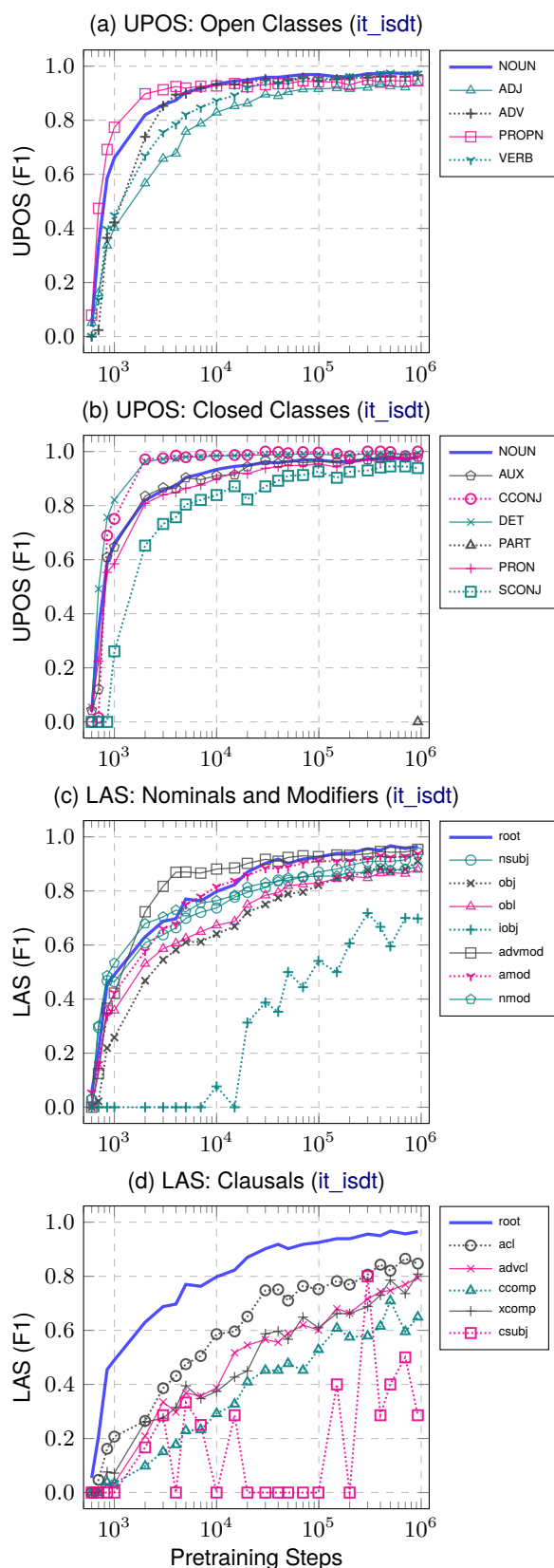


	Train	Dev	Test
Number of sentences	4,005	451	454
Number of words	95,881	10,000	10,109
Number of label occurrences			
<i>ADJ</i>	5,501	531	489
<i>ADP</i>	15,259	1,503	1,514
<i>ADV</i>	1,478	200	208
<i>AUX</i>	1,224	169	164
<i>CCONJ</i>	3,241	292	294
<i>DET</i>	8,549	871	867
<i>INTJ</i>	14	9	7
<i>NOUN</i>	28,141	2,640	2,592
<i>NUM</i>	1,625	140	176
<i>PART</i>	5,635	656	685
<i>PRON</i>	2,759	417	445
<i>PROPN</i>	4,724	487	515
<i>PUNCT</i>	8,993	1,082	1,108
<i>SCONJ</i>	1,304	162	163
<i>SYM</i>	22	1	1
<i>VERB</i>	7,157	793	825
<i>X</i>	255	47	56
<i>acl</i>	1,879	172	212
<i>advcl</i>	1,083	137	146
<i>advmod</i>	1,969	269	269
<i>amod</i>	3,595	323	291
<i>appos</i>	336	58	62
<i>case</i>	13,135	1,221	1,227
<i>cc</i>	3,129	268	246
<i>ccomp</i>	695	83	83
<i>compound</i>	162	16	28
<i>conj</i>	3,948	308	287
<i>cop</i>	1,213	162	163
<i>csubj</i>	764	101	95
<i>det</i>	7,900	798	756
<i>discourse</i>	22	8	7
<i>dislocated</i>	47	0	8
<i>fixed</i>	1,366	153	126
<i>flat</i>	908	202	157
<i>goeswith</i>	3	1	0
<i>list</i>	93	7	66
<i>mark</i>	5,439	627	657
<i>nmod</i>	11,605	1,068	1,151
<i>nsubj</i>	5,310	641	665
<i>nummod</i>	554	44	50
<i>obj</i>	3,825	367	368
<i>obl</i>	9,146	939	891
<i>orphan</i>	9	0	2
<i>parataxis</i>	542	119	95
<i>punct</i>	8,986	1,082	1,105
<i>root</i>	4,005	451	454
<i>vocative</i>	30	14	11
<i>xcomp</i>	4,183	361	431

Table 24: Statistics of *ga\_idt*

Figure 24: Syntactic acquisition through pretraining of OLMo-2-7B on Irish (*ga\_idt*).

## C.17. Italian

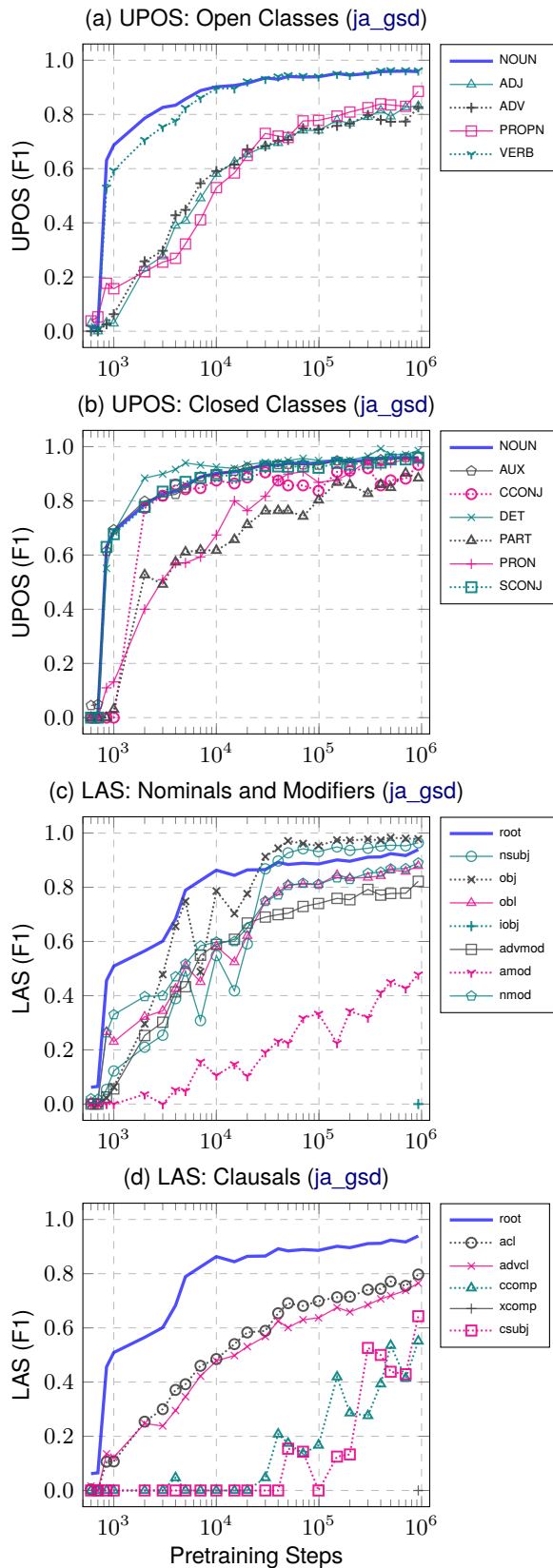


	Train	Dev	Test
Number of sentences	13,121	564	482
Number of words	276,014	11,907	10,417
Number of label occurrences			
<i>ADJ</i>	18,321	778	680
<i>ADP</i>	41,816	1,776	1,638
<i>ADV</i>	10,622	482	406
<i>AUX</i>	10,877	422	405
<i>CCONJ</i>	7,555	323	263
<i>DET</i>	45,035	1,895	1,712
<i>INTJ</i>	62	3	0
<i>NOUN</i>	54,968	2,390	2,070
<i>NUM</i>	4,791	226	171
<i>PART</i>	24	2	0
<i>PRON</i>	10,445	456	411
<i>PROPN</i>	13,676	595	505
<i>PUNCT</i>	31,289	1,413	1,175
<i>SCONJ</i>	2,817	115	100
<i>SYM</i>	93	4	5
<i>VERB</i>	23,364	1,021	863
<i>X</i>	259	6	13
<i>acl</i>	5,653	253	213
<i>advcl</i>	3,444	141	130
<i>advmod</i>	9,662	439	367
<i>amod</i>	15,435	667	586
<i>appos</i>	837	39	40
<i>aux</i>	7,674	296	300
<i>case</i>	38,639	1,643	1,540
<i>cc</i>	7,574	324	263
<i>ccomp</i>	1,349	61	38
<i>compound</i>	709	26	23
<i>conj</i>	9,444	370	311
<i>cop</i>	3,201	126	105
<i>csubj</i>	300	13	3
<i>dep</i>	13	0	1
<i>det</i>	45,020	1,895	1,711
<i>discourse</i>	59	3	0
<i>dislocated</i>	31	0	1
<i>expl</i>	2,757	118	105
<i>fixed</i>	939	41	45
<i>flat</i>	3,887	175	143
<i>goeswith</i>	0	0	1
<i>iobj</i>	648	26	20
<i>mark</i>	5,859	242	190
<i>nmod</i>	22,133	977	842
<i>nsubj</i>	13,986	598	535
<i>nummod</i>	3,263	184	113
<i>obj</i>	9,500	415	336
<i>obl</i>	16,943	726	699
<i>orphan</i>	38	1	1
<i>parataxis</i>	413	31	17
<i>punct</i>	31,289	1,413	1,175
<i>root</i>	13,121	564	482
<i>vocative</i>	96	3	3
<i>xcomp</i>	2,098	97	78

Table 25: Statistics of *it\_isdt*

Figure 25: Syntactic acquisition through pretraining of OLMo-2-7B on Italian (*it\_isdt*).

## C.18. Japanese

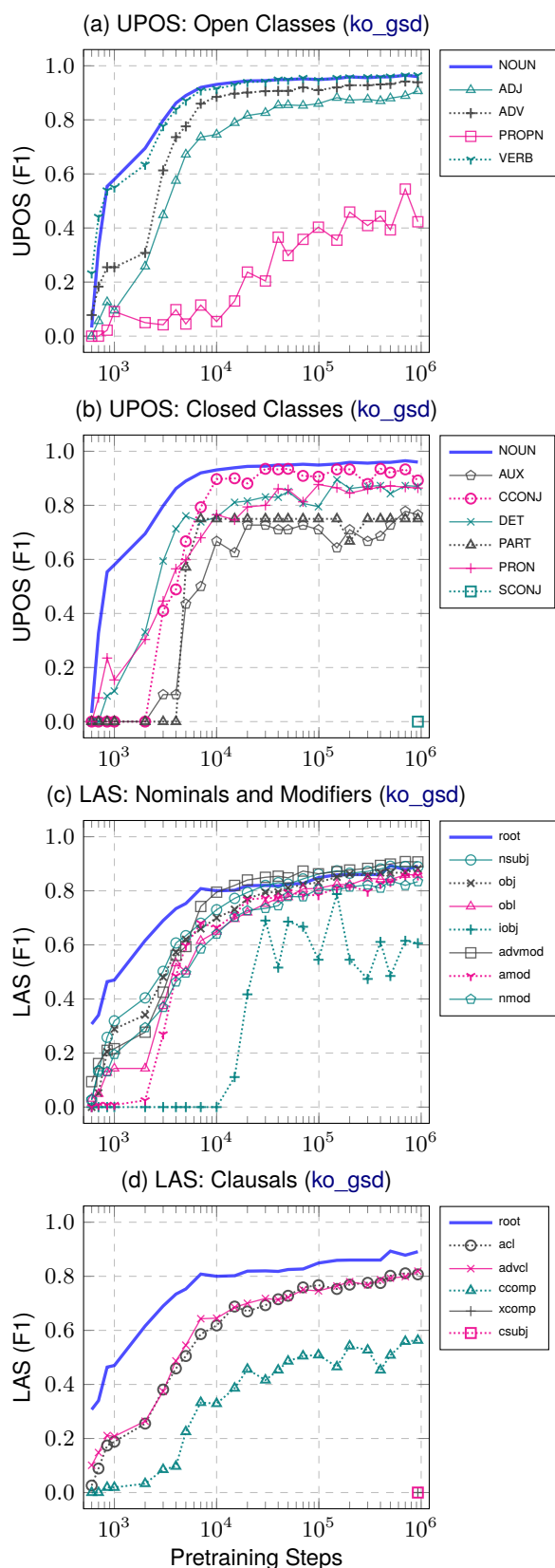


	Train	Dev	Test
Number of sentences	7,050	507	543
Number of words	168,333	12,287	13,034
Number of label occurrences			
<i>ADJ</i>	3,226	264	349
<i>ADP</i>	36,388	2,669	2,807
<i>ADV</i>	1,999	142	223
<i>AUX</i>	18,325	1,349	1,484
<i>CCONJ</i>	727	48	44
<i>DET</i>	863	53	71
<i>INTJ</i>	12	0	1
<i>NOUN</i>	50,730	3,765	3,689
<i>NUM</i>	4,607	305	251
<i>PART</i>	1,061	70	128
<i>PRON</i>	935	75	98
<i>PROPN</i>	6,351	477	313
<i>PUNCT</i>	16,776	1,166	1,291
<i>SCONJ</i>	6,809	507	679
<i>SYM</i>	1,153	78	70
<i>VERB</i>	18,371	1,319	1,536
<i>acl</i>	6,029	456	513
<i>advcl</i>	6,166	435	596
<i>advmod</i>	1,925	140	215
<i>amod</i>	370	37	38
<i>aux</i>	14,962	1,077	1,196
<i>case</i>	35,898	2,632	2,774
<i>cc</i>	727	48	44
<i>ccomp</i>	328	22	40
<i>compound</i>	24,192	1,783	1,514
<i>cop</i>	2,102	172	167
<i>csubj</i>	138	13	12
<i>dep</i>	60	6	10
<i>det</i>	863	53	71
<i>discourse</i>	15	0	1
<i>fixed</i>	7,459	561	603
<i>mark</i>	6,647	500	713
<i>nmod</i>	11,372	846	752
<i>nsubj</i>	7,110	533	593
<i>nummod</i>	2,487	169	144
<i>obj</i>	4,629	331	349
<i>obl</i>	11,028	800	855
<i>punct</i>	16,776	1,166	1,291
<i>root</i>	7,050	507	543

Table 26: Statistics of *ja\_gsd*

Figure 26: Syntactic acquisition through pretraining of OLMo-2-7B on Japanese (*ja\_gsd*).

## C.19. Korean

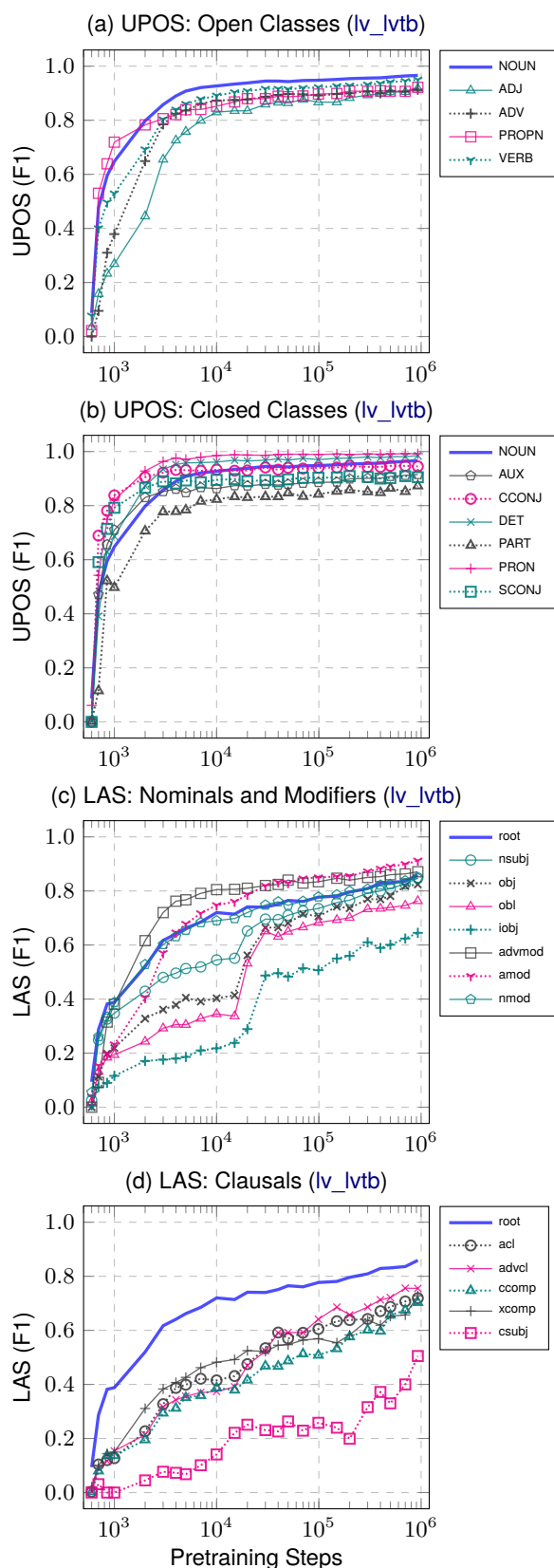


	Train	Dev	Test
Number of sentences	4,400	950	989
Number of words	56,687	11,958	11,677
Number of label occurrences			
<i>ADJ</i>	1,912	405	443
<i>ADP</i>	1,288	253	252
<i>ADV</i>	7,950	1,729	1,683
<i>AUX</i>	87	20	18
<i>CCONJ</i>	168	24	31
<i>DET</i>	393	89	91
<i>INTJ</i>	3	0	0
<i>NOUN</i>	22,915	4,810	4,622
<i>NUM</i>	610	123	115
<i>PART</i>	25	1	5
<i>PRON</i>	480	88	114
<i>PROPN</i>	359	72	64
<i>PUNCT</i>	7,335	1,583	1,493
<i>SYM</i>	197	38	40
<i>VERB</i>	12,955	2,720	2,706
<i>X</i>	10	3	0
<i>acl</i>	2,271	473	459
<i>advcl</i>	3,230	661	664
<i>advmod</i>	4,468	986	1,014
<i>amod</i>	1,114	221	231
<i>appos</i>	1,085	232	193
<i>aux</i>	33	10	4
<i>case</i>	1,162	243	226
<i>cc</i>	168	24	31
<i>ccomp</i>	450	98	103
<i>compound</i>	4	1	2
<i>conj</i>	2,715	598	553
<i>cop</i>	49	10	14
<i>csubj</i>	12	4	5
<i>dep</i>	1,750	365	318
<i>det</i>	513	114	122
<i>discourse</i>	1	0	0
<i>fixed</i>	10	2	1
<i>flat</i>	8,700	1,852	1,690
<i>iobj</i>	77	16	15
<i>list</i>	1	0	0
<i>mark</i>	65	11	10
<i>nmod</i>	2,936	628	526
<i>nsubj</i>	5,771	1,195	1,323
<i>nummod</i>	397	75	83
<i>obj</i>	4,122	810	868
<i>obl</i>	3,847	795	740
<i>parataxis</i>	1	0	0
<i>punct</i>	7,335	1,583	1,493
<i>root</i>	4,400	950	989
<i>xcomp</i>	0	1	0

Table 27: Statistics of *ko\_gsd*

Figure 27: Syntactic acquisition through pretraining of OLMo-2-7B on Korean (*ko\_gsd*).

## C.20. Latvian

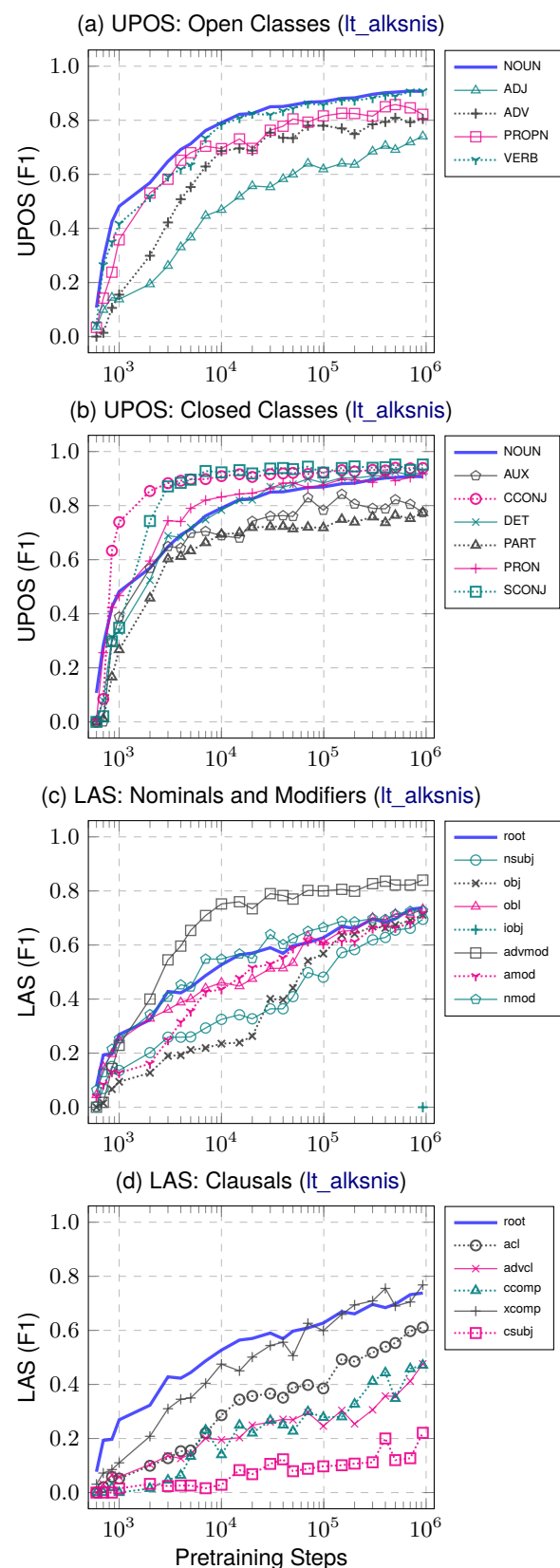


	Train	Dev	Test
Number of sentences	14,887	2,080	2,400
Number of words	255,954	34,348	37,162
Number of label occurrences			
<i>ADJ</i>	13,907	1,762	1,808
<i>ADP</i>	11,432	1,465	1,581
<i>ADV</i>	14,230	1,931	2,197
<i>AUX</i>	7,086	882	950
<i>CCONJ</i>	10,435	1,363	1,482
<i>DET</i>	11,917	1,490	1,765
<i>INTJ</i>	107	46	68
<i>NOUN</i>	69,455	9,143	9,695
<i>NUM</i>	3,343	415	402
<i>PART</i>	5,070	739	938
<i>PRON</i>	8,006	1,144	1,217
<i>PROPN</i>	11,109	1,348	1,521
<i>PUNCT</i>	44,053	6,423	6,831
<i>SCONJ</i>	5,998	743	847
<i>SYM</i>	495	45	47
<i>VERB</i>	38,374	5,282	5,620
<i>X</i>	937	127	193
<i>acl</i>	4,266	545	611
<i>advcl</i>	4,810	635	654
<i>advmod</i>	14,346	1,947	2,187
<i>amod</i>	13,882	1,701	1,834
<i>appos</i>	501	80	67
<i>aux</i>	3,264	420	470
<i>case</i>	11,610	1,490	1,598
<i>cc</i>	10,544	1,379	1,511
<i>ccomp</i>	4,342	561	614
<i>compound</i>	57	7	2
<i>conj</i>	14,635	2,151	2,128
<i>cop</i>	3,779	455	475
<i>csubj</i>	853	103	115
<i>dep</i>	691	79	96
<i>det</i>	5,988	762	916
<i>discourse</i>	3,497	532	684
<i>dislocated</i>	11	2	1
<i>fixed</i>	1,217	156	204
<i>flat</i>	2,474	275	391
<i>goeswith</i>	18	1	3
<i>iobj</i>	6,141	828	874
<i>mark</i>	5,512	666	758
<i>nmod</i>	27,975	3,525	3,837
<i>nsubj</i>	21,561	2,816	3,014
<i>nummod</i>	2,400	326	282
<i>obj</i>	10,929	1,571	1,555
<i>obl</i>	16,338	2,074	2,278
<i>orphan</i>	406	86	100
<i>parataxis</i>	997	141	111
<i>punct</i>	44,051	6,419	6,831
<i>reparandum</i>	10	4	3
<i>root</i>	14,887	2,080	2,400
<i>vocative</i>	71	11	29
<i>xcomp</i>	3,891	520	529

Table 28: Statistics of *lv\_lvtb*

Figure 28: Syntactic acquisition through pretraining of OLMo-2-7B on Latvian (*lv\_lvtb*).

## C.21. Lithuanian

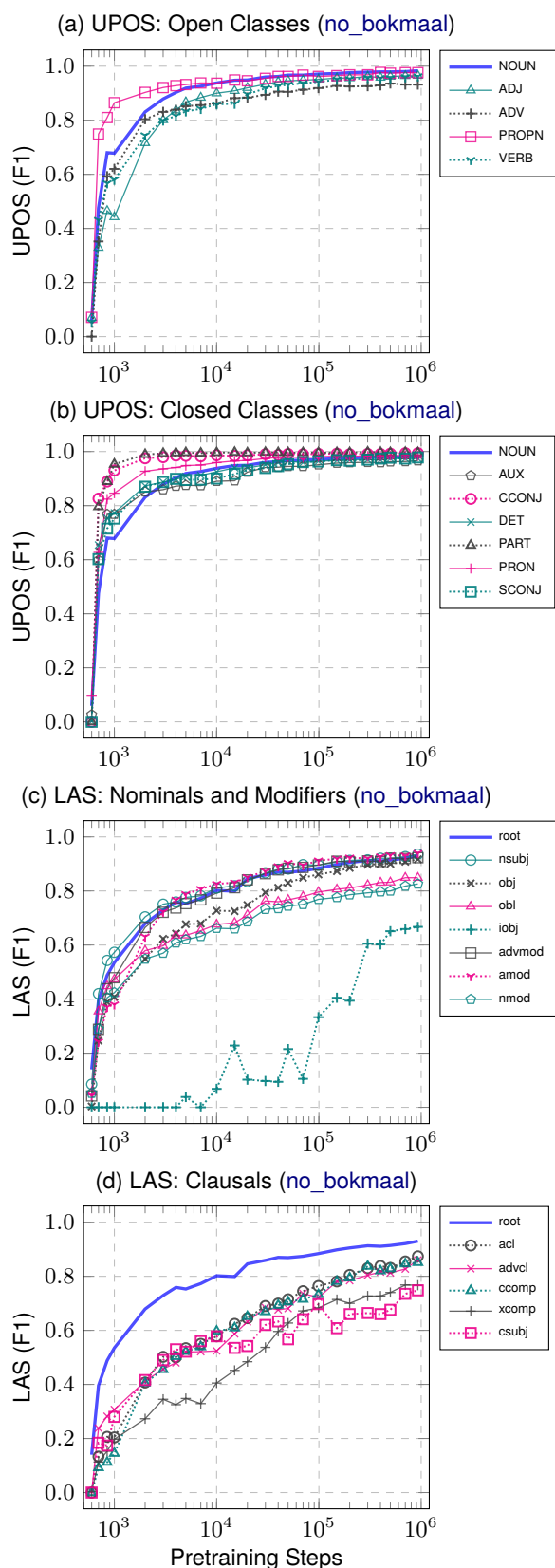


	Train	Dev	Test
Number of sentences	2,341	617	684
Number of words	47,641	11,560	10,846
Number of label occurrences			
<i>ADJ</i>	3,274	768	609
<i>ADP</i>	1,490	402	439
<i>ADV</i>	1,826	432	570
<i>AUX</i>	453	129	102
<i>CCONJ</i>	2,136	471	474
<i>DET</i>	1,181	311	288
<i>INTJ</i>	9	2	8
<i>NOUN</i>	14,933	3,534	2,810
<i>NUM</i>	1,328	202	169
<i>PART</i>	951	246	356
<i>PRON</i>	1,688	283	448
<i>PROPN</i>	983	365	245
<i>PUNCT</i>	8,756	2,081	2,059
<i>SCONJ</i>	917	227	279
<i>SYM</i>	45	9	8
<i>VERB</i>	6,604	1,758	1,818
<i>X</i>	1,067	340	164
<i>acl</i>	1,586	379	297
<i>advcl</i>	851	252	226
<i>advmod</i>	2,569	620	870
<i>amod</i>	2,653	661	481
<i>appos</i>	125	26	43
<i>case</i>	1,476	398	432
<i>cc</i>	2,112	467	470
<i>ccomp</i>	267	88	108
<i>compound</i>	1	0	1
<i>conj</i>	3,650	663	761
<i>cop</i>	453	129	102
<i>csubj</i>	322	81	64
<i>dep</i>	79	7	17
<i>det</i>	664	159	165
<i>discourse</i>	1	0	4
<i>flat</i>	212	117	52
<i>iobj</i>	2	2	0
<i>mark</i>	949	236	289
<i>nmod</i>	7,052	1,663	1,055
<i>nsubj</i>	2,579	703	624
<i>nummod</i>	1,011	153	113
<i>obj</i>	1,742	408	379
<i>obl</i>	5,057	1,328	1,210
<i>orphan</i>	3	0	0
<i>parataxis</i>	402	88	116
<i>punct</i>	8,756	2,081	2,059
<i>root</i>	2,341	617	684
<i>xcomp</i>	726	234	224

Table 29: Statistics of *lt\_alksnis*

Figure 29: Syntactic acquisition through pretraining of OLMo-2-7B on Lithuanian (*lt\_alksnis*).

## C.22. Norwegian

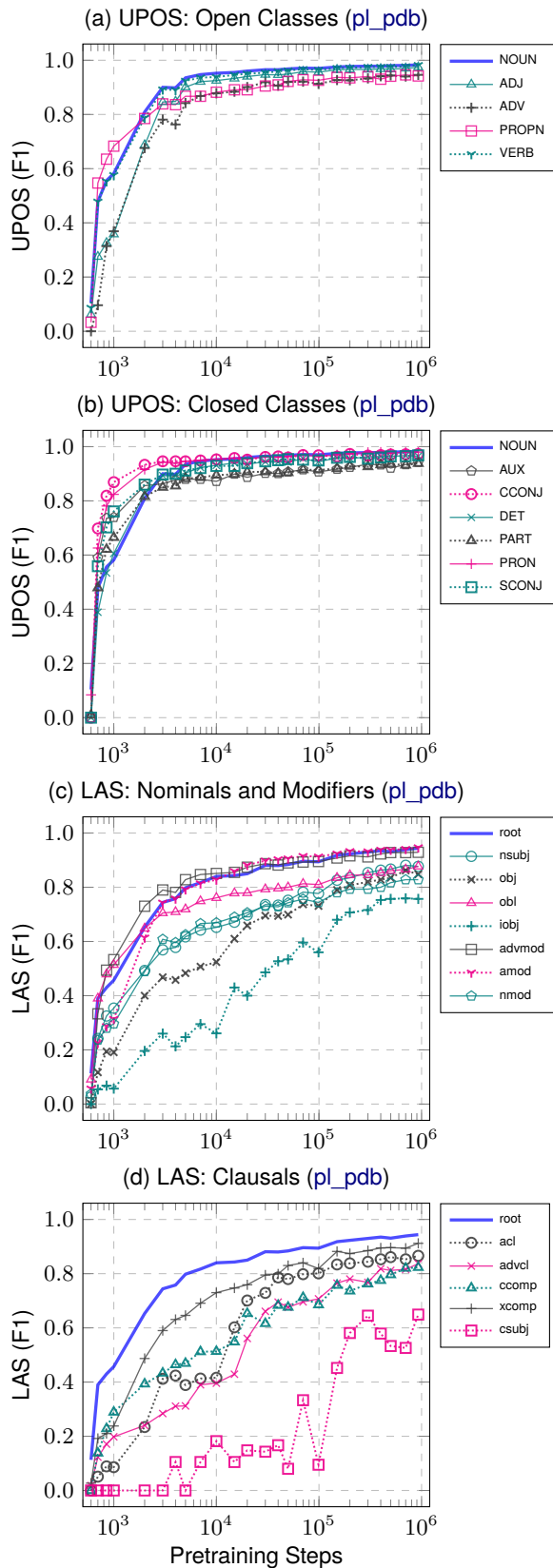


	Train	Dev	Test
Number of sentences	15,696	2,409	1,939
Number of words	243,886	36,369	29,966
Number of label occurrences			
<i>ADJ</i>	21,215	3,145	2,450
<i>ADP</i>	29,753	4,497	3,773
<i>ADV</i>	10,038	1,440	1,220
<i>AUX</i>	13,394	2,064	1,723
<i>CCONJ</i>	8,474	1,307	1,003
<i>DET</i>	11,416	1,671	1,309
<i>INTJ</i>	167	35	22
<i>NOUN</i>	45,008	6,768	5,477
<i>NUM</i>	3,049	486	410
<i>PART</i>	5,514	808	676
<i>PRON</i>	17,817	2,621	2,207
<i>PUNCT</i>	29,128	4,352	3,501
<i>SCONJ</i>	7,845	1,074	947
<i>SYM</i>	63	7	11
<i>VERB</i>	26,154	3,868	3,329
<i>X</i>	569	51	105
<i>acl</i>	4,158	619	494
<i>advcl</i>	4,084	565	476
<i>advmod</i>	16,532	2,384	2,051
<i>amod</i>	11,260	1,684	1,257
<i>appos</i>	1,308	174	115
<i>aux</i>	6,996	1,128	954
<i>case</i>	29,238	4,420	3,705
<i>cc</i>	8,462	1,303	1,008
<i>ccomp</i>	2,322	308	315
<i>compound</i>	239	41	24
<i>conj</i>	9,452	1,377	1,138
<i>cop</i>	6,395	936	767
<i>csubj</i>	794	135	103
<i>det</i>	12,474	1,834	1,400
<i>discourse</i>	108	23	11
<i>dislocated</i>	163	33	28
<i>expl</i>	2,552	370	319
<i>flat</i>	5,018	713	646
<i>iobj</i>	471	74	51
<i>mark</i>	11,216	1,603	1,340
<i>nmod</i>	12,761	2,030	1,512
<i>nsubj</i>	21,553	3,172	2,785
<i>nummod</i>	1,831	245	222
<i>obj</i>	10,632	1,583	1,393
<i>obl</i>	14,971	2,221	1,930
<i>orphan</i>	2	1	0
<i>parataxis</i>	187	55	15
<i>punct</i>	29,128	4,352	3,501
<i>reparandum</i>	31	11	1
<i>root</i>	15,696	2,409	1,939
<i>xcomp</i>	3,852	566	466

Table 30: Statistics of *no\_bokmaal*

Figure 30: Syntactic acquisition through pretraining of OLMo-2-7B on Norwegian (*no\_bokmaal*).

### C.23. Polish

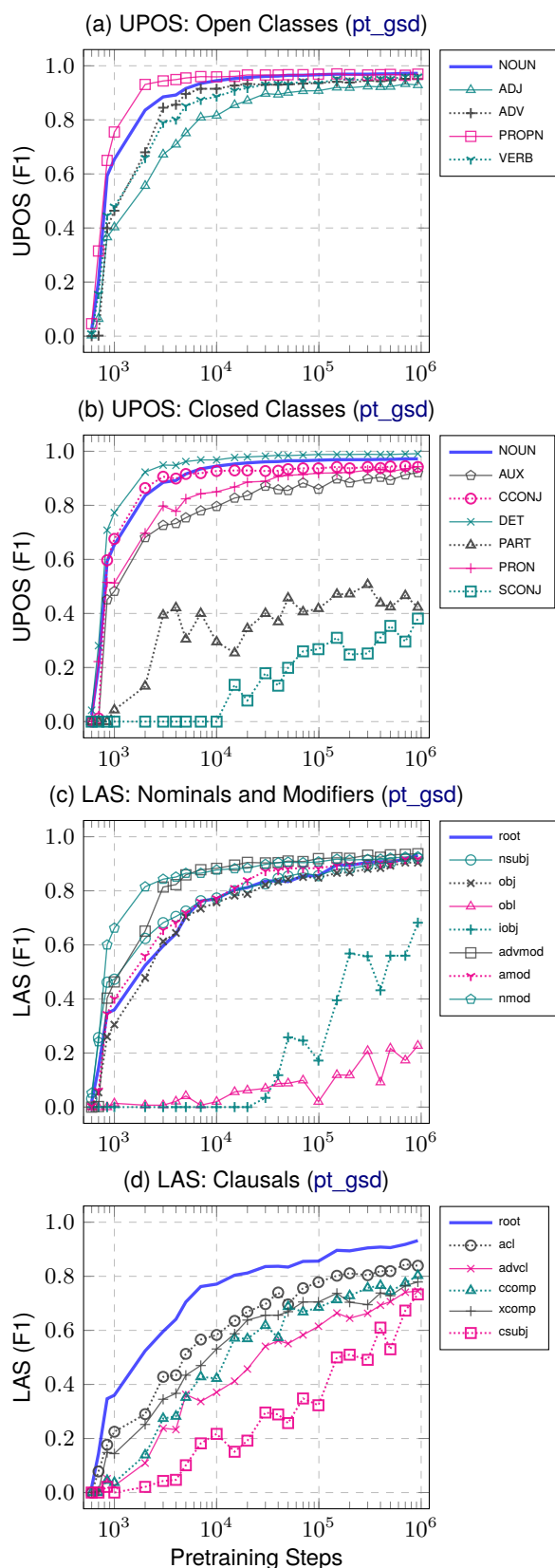


	Train	Dev	Test
Number of sentences	17,722	2,215	2,215
Number of words	281,685	34,677	33,616
Number of label occurrences			
ADJ	29,018	3,545	3,365
ADP	29,890	3,748	3,523
ADV	9,218	1,163	1,097
AUX	7,110	820	845
CCONJ	8,386	1,023	1,033
DET	7,576	920	852
INTJ	139	23	13
NOUN	71,287	8,828	8,519
NUM	2,113	262	258
PART	8,830	1,021	1,059
PRON	13,274	1,620	1,592
PROPN	9,634	1,214	1,152
PUNCT	46,497	5,748	5,628
SCONJ	6,136	706	689
SYM	14	3	4
VERB	31,725	3,922	3,880
X	838	111	107
<i>acl</i>	6,059	782	681
<i>advcl</i>	3,236	374	380
<i>advmod</i>	16,988	2,063	1,994
<i>amod</i>	20,076	2,425	2,375
<i>appos</i>	1,727	232	213
<i>aux</i>	4,352	529	527
<i>case</i>	29,129	3,662	3,440
<i>cc</i>	8,494	1,031	1,059
<i>ccomp</i>	2,735	316	313
<i>conj</i>	12,591	1,534	1,545
<i>cop</i>	2,756	291	318
<i>csubj</i>	185	22	16
<i>dep</i>	10	1	3
<i>det</i>	5,448	664	607
<i>discourse</i>	128	18	12
<i>expl</i>	4,960	600	595
<i>fixed</i>	3,157	381	357
<i>flat</i>	2,081	276	258
<i>iobj</i>	5,311	649	673
<i>list</i>	208	39	26
<i>mark</i>	6,175	711	698
<i>nmod</i>	21,074	2,562	2,454
<i>nsubj</i>	17,077	2,081	2,011
<i>nummod</i>	1,916	241	230
<i>obj</i>	12,242	1,526	1,479
<i>obl</i>	22,628	2,908	2,703
<i>orphan</i>	100	11	7
<i>parataxis</i>	2,313	262	260
<i>punct</i>	46,496	5,748	5,628
<i>root</i>	17,722	2,215	2,215
<i>vocative</i>	218	28	36
<i>xcomp</i>	4,093	495	503

Table 31: Statistics of pl\_pdb

Figure 31: Syntactic acquisition through pretraining of OLMo-2-7B on Polish (pl\_pdb).

## C.24. Portuguese

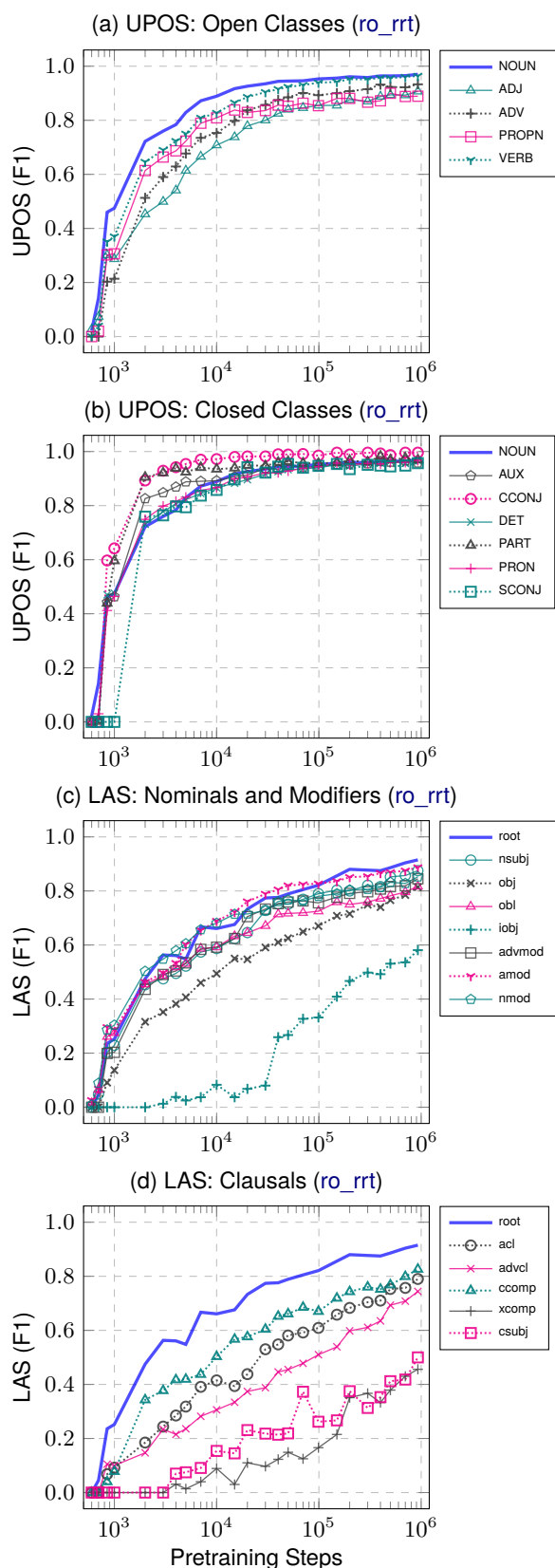


	Train	Dev	Test
Number of sentences	9,616	1,204	1,200
Number of words	255,116	32,073	31,477
Number of label occurrences			
<i>ADJ</i>	12,076	1,486	1,477
<i>ADP</i>	40,980	5,201	5,045
<i>ADV</i>	7,835	985	950
<i>AUX</i>	5,588	687	661
<i>CCONJ</i>	8,314	1,086	1,026
<i>DET</i>	37,999	4,907	4,696
<i>NOUN</i>	45,237	5,769	5,584
<i>NUM</i>	6,824	832	856
<i>PART</i>	459	60	46
<i>PRON</i>	6,214	763	743
<i>PROP</i>	25,908	3,143	3,228
<i>PUNCT</i>	33,234	4,114	4,074
<i>SCONJ</i>	975	90	120
<i>SYM</i>	813	98	107
<i>VERB</i>	22,356	2,812	2,805
<i>X</i>	304	40	59
<i>acl</i>	4,948	665	630
<i>advcl</i>	2,143	243	284
<i>advmod</i>	7,167	884	875
<i>amod</i>	12,367	1,550	1,502
<i>appos</i>	4,880	579	550
<i>aux</i>	3,786	471	454
<i>case</i>	38,458	4,862	4,727
<i>cc</i>	6,187	815	760
<i>ccomp</i>	1,838	244	219
<i>compound</i>	18	2	1
<i>conj</i>	8,253	1,046	1,045
<i>cop</i>	1,779	214	203
<i>csubj</i>	490	50	61
<i>dep</i>	195	20	16
<i>det</i>	37,918	4,902	4,685
<i>discourse</i>	2	0	1
<i>expl</i>	611	65	60
<i>fixed</i>	940	107	115
<i>flat</i>	5,948	676	770
<i>iobj</i>	377	56	45
<i>mark</i>	5,312	671	677
<i>nmod</i>	37,002	4,834	4,583
<i>nsubj</i>	13,188	1,661	1,659
<i>nummod</i>	3,777	452	483
<i>obj</i>	9,046	1,162	1,127
<i>obl</i>	2,420	189	270
<i>parataxis</i>	670	74	70
<i>punct</i>	33,234	4,114	4,074
<i>root</i>	9,616	1,204	1,200
<i>xcomp</i>	2,546	261	331

Table 32: Statistics of *pt\_gsd*

Figure 32: Syntactic acquisition through pretraining of OLMo-2-7B on Portuguese (*pt\_gsd*).

## C.25. Romanian

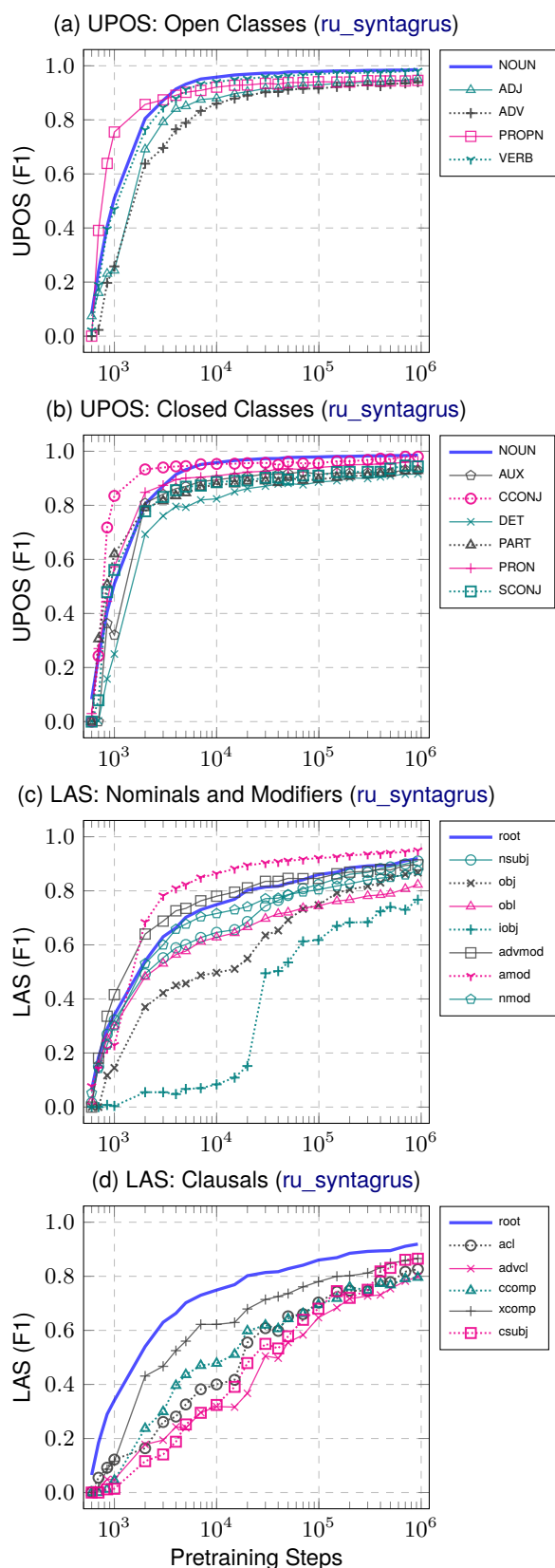


	Train	Dev	Test
Number of sentences	8,043	752	729
Number of words	185,125	17,073	16,324
Number of label occurrences			
<i>ADJ</i>	12,855	1,271	1,172
<i>ADP</i>	26,329	2,393	2,333
<i>ADV</i>	7,554	662	650
<i>AUX</i>	7,289	658	618
<i>CCONJ</i>	5,940	519	471
<i>DET</i>	10,196	931	898
<i>INTJ</i>	32	3	6
<i>NOUN</i>	45,990	4,226	4,042
<i>NUM</i>	4,675	418	456
<i>PART</i>	4,149	378	358
<i>PRON</i>	10,014	931	862
<i>PROPN</i>	4,919	511	455
<i>PUNCT</i>	23,691	2,223	2,083
<i>SCONJ</i>	1,896	151	154
<i>VERB</i>	19,466	1,784	1,749
<i>X</i>	130	14	17
<i>acl</i>	4,168	345	356
<i>advcl</i>	2,880	256	231
<i>advmod</i>	8,491	776	704
<i>amod</i>	11,093	1,118	1,040
<i>appos</i>	951	100	87
<i>aux</i>	5,414	490	478
<i>case</i>	23,007	2,064	2,072
<i>cc</i>	5,647	492	464
<i>ccomp</i>	2,115	192	224
<i>compound</i>	55	3	3
<i>conj</i>	7,417	660	614
<i>cop</i>	1,828	167	135
<i>csubj</i>	755	68	41
<i>dep</i>	32	2	3
<i>det</i>	10,090	914	879
<i>discourse</i>	16	1	3
<i>expl</i>	3,968	377	325
<i>fixed</i>	5,268	536	475
<i>flat</i>	1,237	156	91
<i>goeswith</i>	67	4	11
<i>iobj</i>	1,366	140	139
<i>list</i>	9	0	1
<i>mark</i>	5,480	457	454
<i>nmod</i>	17,054	1,582	1,532
<i>nsubj</i>	10,204	939	894
<i>nummod</i>	3,565	326	352
<i>obj</i>	6,430	588	562
<i>obl</i>	12,355	1,124	1,128
<i>orphan</i>	61	3	5
<i>parataxis</i>	1,337	133	99
<i>punct</i>	23,691	2,223	2,083
<i>reparandum</i>	1	0	0
<i>root</i>	8,043	752	729
<i>vocative</i>	54	3	4
<i>xcomp</i>	976	82	106

Table 33: Statistics of *ro\_rrt*

Figure 33: Syntactic acquisition through pretraining of OLMo-2-7B on Romanian (*ro\_rrt*).

## C.26. Russian

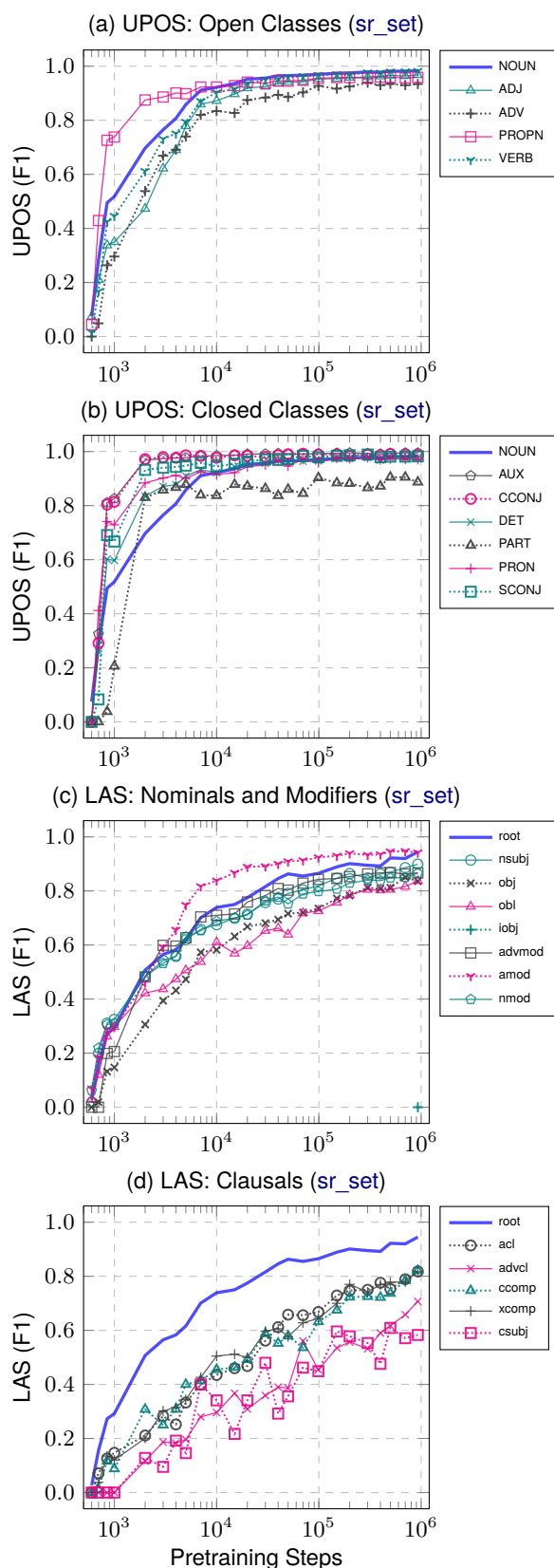


	Train	Dev	Test
Number of sentences	69,630	8,906	8,800
Number of words	1,204,640	153,325	157,718
Number of label occurrences			
<i>ADJ</i>	116,158	15,103	14,471
<i>ADP</i>	112,193	13,717	15,062
<i>ADV</i>	60,719	7,783	8,085
<i>AUX</i>	10,819	1,390	1,518
<i>CCONJ</i>	43,636	5,672	5,736
<i>DET</i>	32,780	4,265	4,094
<i>INTJ</i>	183	24	23
<i>NOUN</i>	287,221	36,238	36,568
<i>NUM</i>	15,169	1,734	2,528
<i>PART</i>	39,634	5,125	4,921
<i>PRON</i>	57,486	7,444	8,015
<i>PROPN</i>	44,399	5,473	5,883
<i>PUNCT</i>	222,074	29,186	29,463
<i>SCONJ</i>	22,668	2,865	2,992
<i>SYM</i>	1,052	62	165
<i>VERB</i>	137,570	17,110	18,146
<i>X</i>	879	134	48
<i>acl</i>	20,368	2,594	2,593
<i>advcl</i>	12,362	1,542	1,810
<i>advmod</i>	77,228	9,737	9,860
<i>amod</i>	95,863	12,319	11,954
<i>appos</i>	9,579	1,114	1,371
<i>aux</i>	5,236	583	788
<i>case</i>	110,899	13,647	14,943
<i>cc</i>	44,657	5,832	5,869
<i>ccomp</i>	6,726	803	865
<i>compound</i>	1,386	164	141
<i>conj</i>	56,740	8,034	7,640
<i>cop</i>	5,555	805	729
<i>csubj</i>	6,743	755	897
<i>dep</i>	4	0	1
<i>det</i>	30,241	3,930	3,752
<i>discourse</i>	979	127	116
<i>dislocated</i>	12	0	0
<i>expl</i>	801	107	100
<i>fixed</i>	11,125	1,354	1,445
<i>flat</i>	8,812	1,276	1,374
<i>iobj</i>	10,424	1,374	1,425
<i>list</i>	11	1	0
<i>mark</i>	21,281	2,601	2,804
<i>nmod</i>	99,066	12,409	12,179
<i>nsubj</i>	89,215	10,985	11,546
<i>nummod</i>	11,229	1,135	1,622
<i>obj</i>	39,513	4,890	5,124
<i>obl</i>	97,544	12,304	13,404
<i>orphan</i>	1,106	179	171
<i>parataxis</i>	24,841	3,125	3,285
<i>punct</i>	222,074	29,186	29,463
<i>root</i>	69,630	8,906	8,800
<i>vocative</i>	166	16	32
<i>xcomp</i>	13,224	1,491	1,615

Table 34: Statistics of *ru\_syntagrus*

Figure 34: Syntactic acquisition through pretraining of OLMo-2-7B on Russian (*ru\_syntagrus*).

## C.27. Serbian

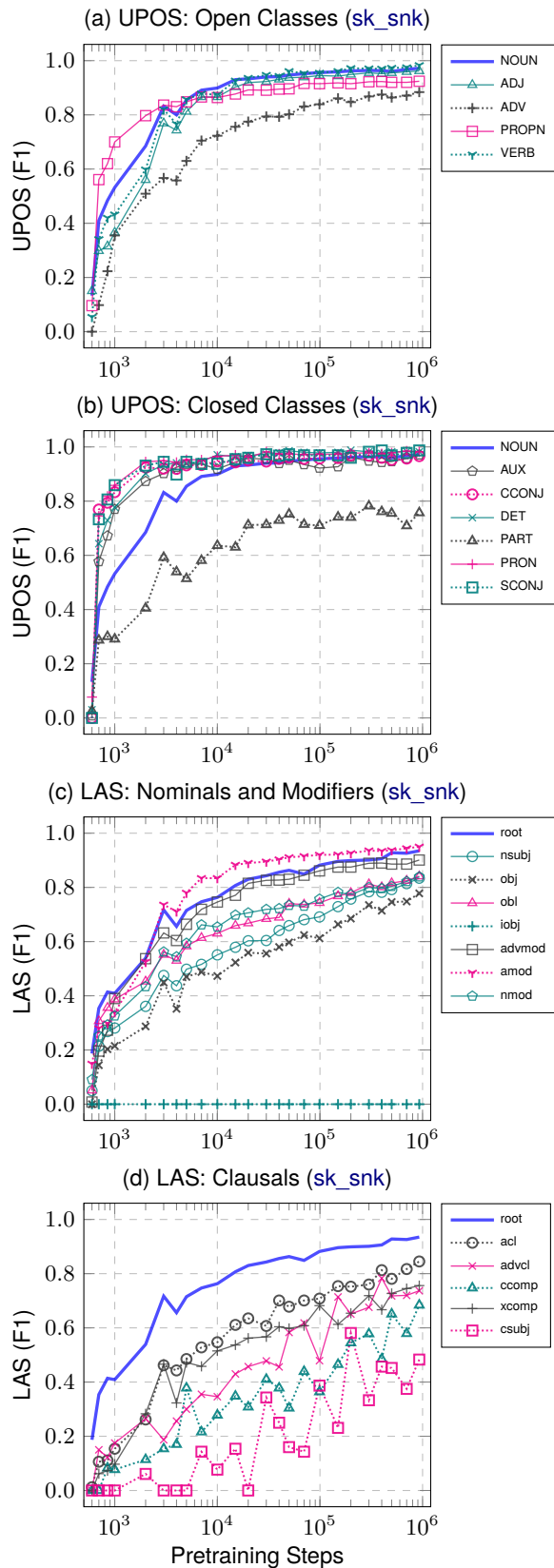


	Train	Dev	Test
Number of sentences	3,328	536	520
Number of words	74,259	11,993	11,421
Number of label occurrences			
<i>ADJ</i>	8,835	1,464	1,362
<i>ADP</i>	7,130	1,124	1,122
<i>ADV</i>	2,543	388	393
<i>AUX</i>	4,667	817	719
<i>CCONJ</i>	2,541	410	380
<i>DET</i>	2,848	433	358
<i>INTJ</i>	3	1	0
<i>NOUN</i>	18,103	2,853	2,862
<i>NUM</i>	944	179	155
<i>PART</i>	461	74	51
<i>PRON</i>	1,843	306	254
<i>PROPN</i>	5,622	860	930
<i>PUNCT</i>	9,351	1,574	1,418
<i>SCONJ</i>	2,729	422	415
<i>SYM</i>	1	2	1
<i>VERB</i>	6,406	1,036	971
<i>X</i>	232	50	30
<i>acl</i>	1,571	240	202
<i>advcl</i>	564	106	79
<i>advmod</i>	2,213	339	318
<i>amod</i>	7,102	1,144	1,105
<i>appos</i>	563	86	78
<i>aux</i>	3,350	594	535
<i>case</i>	7,233	1,131	1,138
<i>cc</i>	2,349	378	343
<i>ccomp</i>	1,012	139	163
<i>compound</i>	2	1	0
<i>conj</i>	2,920	471	432
<i>cop</i>	1,185	205	166
<i>csubj</i>	160	24	24
<i>dep</i>	2	0	0
<i>det</i>	1,476	204	198
<i>discourse</i>	390	57	53
<i>expl</i>	1,076	193	155
<i>fixed</i>	357	70	59
<i>flat</i>	2,152	368	357
<i>iobj</i>	2	0	0
<i>list</i>	6	5	0
<i>mark</i>	2,426	374	372
<i>nmod</i>	6,598	962	1,121
<i>nsubj</i>	5,714	961	859
<i>nummod</i>	808	147	130
<i>obj</i>	2,651	401	395
<i>obl</i>	5,919	959	925
<i>orphan</i>	35	6	4
<i>parataxis</i>	912	176	135
<i>punct</i>	9,351	1,574	1,418
<i>root</i>	3,328	536	520
<i>vocative</i>	1	0	0
<i>xcomp</i>	831	142	137

Table 35: Statistics of *sr\_set*

Figure 35: Syntactic acquisition through pretraining of OLMo-2-7B on Serbian (*sr\_set*).

## C.28. Slovak

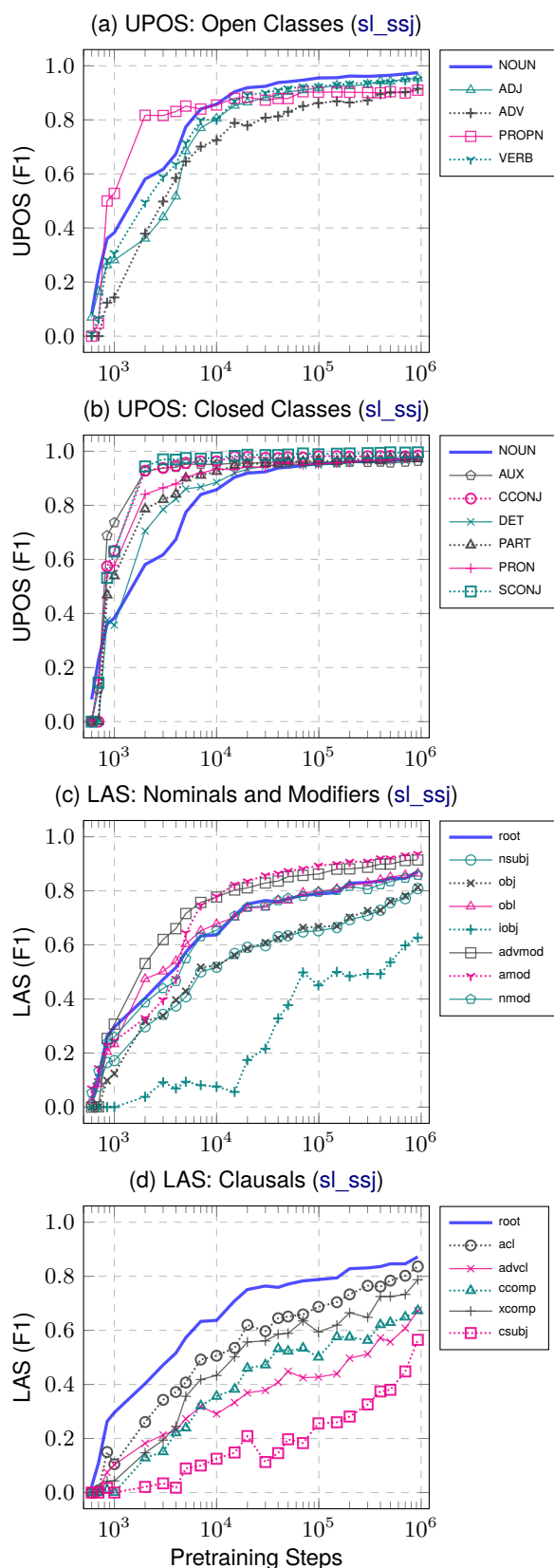


	Train	Dev	Test
Number of sentences	8,483	1,060	1,061
Number of words	80,628	12,733	12,736
Number of label occurrences			
<i>ADJ</i>	6,321	1,635	1,518
<i>ADP</i>	6,209	1,154	1,249
<i>ADV</i>	3,641	394	410
<i>AUX</i>	3,197	286	240
<i>CCONJ</i>	2,527	401	448
<i>DET</i>	3,446	507	451
<i>INTJ</i>	66	0	0
<i>NOUN</i>	15,269	3,219	3,173
<i>NUM</i>	661	442	470
<i>PART</i>	1,591	179	146
<i>PRON</i>	5,718	353	364
<i>PROPN</i>	2,916	827	1,087
<i>PUNCT</i>	15,532	1,809	1,789
<i>SCONJ</i>	1,541	213	111
<i>SYM</i>	2	16	5
<i>VERB</i>	11,757	1,187	1,124
<i>X</i>	234	111	151
<i>acl</i>	685	147	152
<i>advcl</i>	729	77	33
<i>advmod</i>	4,746	545	507
<i>amod</i>	5,232	1,429	1,359
<i>appos</i>	162	65	68
<i>aux</i>	2,033	78	28
<i>case</i>	6,204	1,153	1,254
<i>cc</i>	2,467	390	446
<i>ccomp</i>	1,182	62	17
<i>compound</i>	5	0	0
<i>conj</i>	3,004	507	593
<i>cop</i>	1,163	208	212
<i>csubj</i>	132	35	18
<i>dep</i>	500	53	52
<i>det</i>	1,991	324	288
<i>discourse</i>	45	1	0
<i>expl</i>	2,346	232	266
<i>fixed</i>	129	28	32
<i>flat</i>	31	24	44
<i>iobj</i>	80	7	1
<i>mark</i>	1,621	216	113
<i>nmod</i>	4,188	1,343	1,426
<i>nsubj</i>	5,622	902	818
<i>nummod</i>	406	311	355
<i>obj</i>	4,237	559	514
<i>obl</i>	6,565	1,031	1,143
<i>orphan</i>	60	8	9
<i>parataxis</i>	39	8	18
<i>punct</i>	15,532	1,809	1,789
<i>root</i>	8,483	1,060	1,061
<i>vocative</i>	34	0	0
<i>xcomp</i>	975	121	120

Table 36: Statistics of *sk\_snk*

Figure 36: Syntactic acquisition through pretraining of OLMo-2-7B on Slovak (*sk\_snk*).

## C.29. Slovenian

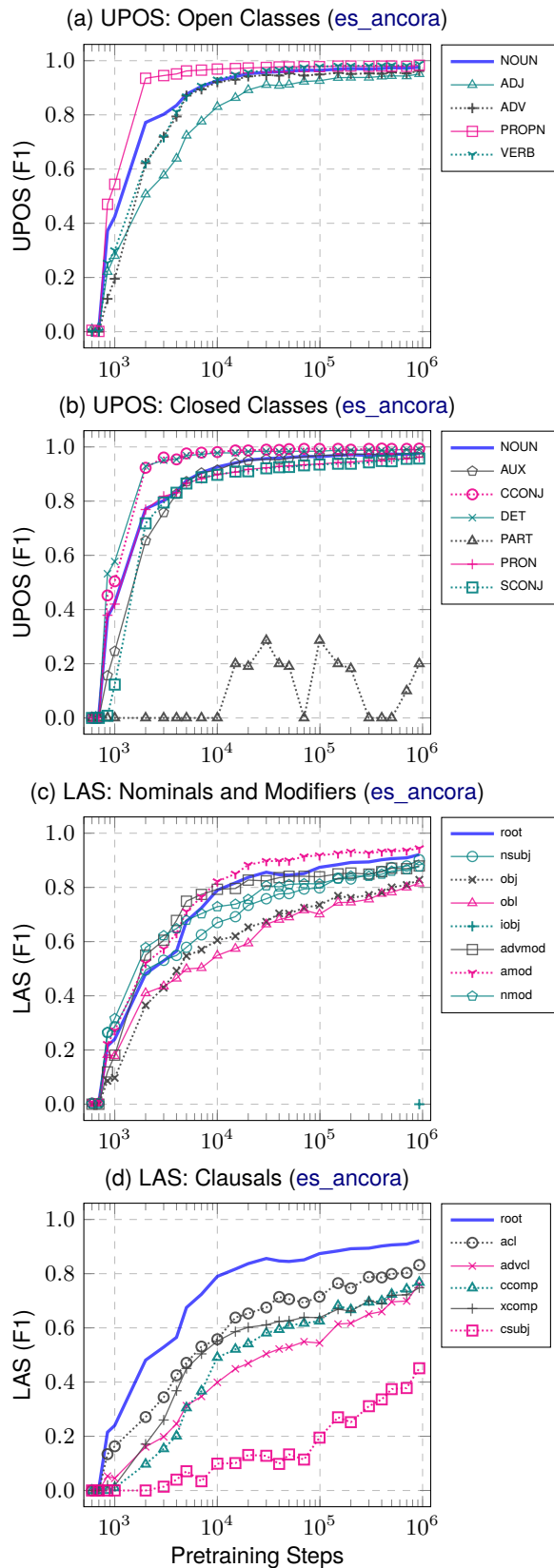


	Train	Dev	Test
Number of sentences	10,903	1,250	1,282
Number of words	215,155	26,500	25,442
Number of label occurrences			
<i>ADJ</i>	22,726	2,919	2,781
<i>ADP</i>	19,351	2,535	2,349
<i>ADV</i>	9,511	1,167	1,125
<i>AUX</i>	14,028	1,594	1,705
<i>CCONJ</i>	8,960	1,146	1,131
<i>DET</i>	7,530	867	955
<i>INTJ</i>	43	1	1
<i>NOUN</i>	45,586	5,790	5,489
<i>NUM</i>	4,625	562	398
<i>PART</i>	6,036	788	718
<i>PRON</i>	7,371	860	875
<i>PROPN</i>	8,269	1,075	895
<i>PUNCT</i>	32,110	3,745	3,623
<i>SCONJ</i>	7,171	905	898
<i>SYM</i>	171	7	21
<i>VERB</i>	19,816	2,376	2,400
<i>X</i>	1,851	163	78
<i>acl</i>	3,329	469	437
<i>advcl</i>	1,970	240	246
<i>advmod</i>	16,288	2,021	1,929
<i>amod</i>	17,627	2,316	2,165
<i>appos</i>	1,497	171	164
<i>aux</i>	9,773	1,081	1,162
<i>case</i>	19,810	2,592	2,415
<i>cc</i>	7,368	949	940
<i>ccomp</i>	1,624	195	203
<i>conj</i>	9,293	1,134	1,108
<i>cop</i>	4,241	512	542
<i>csubj</i>	722	93	78
<i>dep</i>	170	6	12
<i>det</i>	4,732	533	617
<i>discourse</i>	152	15	15
<i>dislocated</i>	10	0	0
<i>expl</i>	2,998	350	361
<i>fixed</i>	987	117	95
<i>flat</i>	2,291	248	156
<i>goeswith</i>	1	0	0
<i>iobj</i>	1,412	135	145
<i>list</i>	583	82	16
<i>mark</i>	6,417	827	804
<i>nmod</i>	15,887	2,182	1,887
<i>nsubj</i>	10,589	1,273	1,312
<i>nummod</i>	3,532	457	311
<i>obj</i>	9,191	1,134	1,085
<i>obl</i>	14,049	1,735	1,722
<i>orphan</i>	797	101	83
<i>parataxis</i>	3,152	349	325
<i>punct</i>	32,108	3,745	3,623
<i>root</i>	10,903	1,250	1,282
<i>vocative</i>	64	5	1
<i>xcomp</i>	1,588	183	201

Table 37: Statistics of *sl\_ssj*

Figure 37: Syntactic acquisition through pretraining of OLMo-2-7B on Slovenian (*sl\_ssj*).

### C.30. Spanish

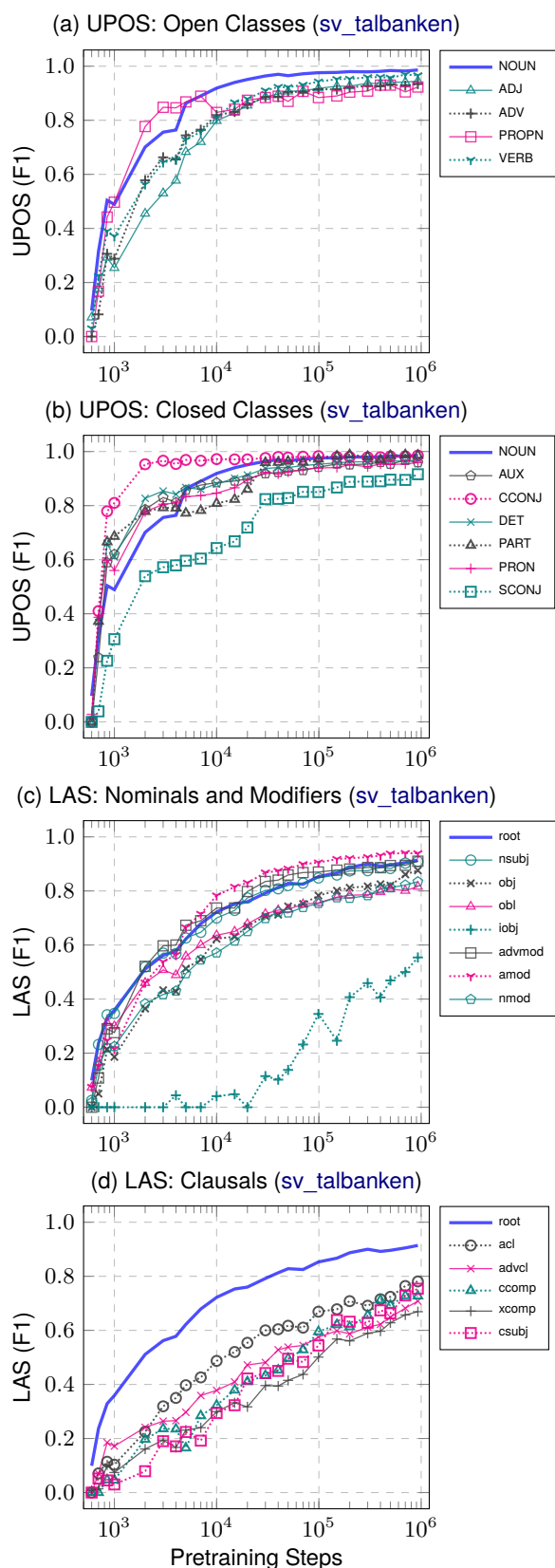


	Train	Dev	Test
Number of sentences	14,287	1,654	1,721
Number of words	453,039	53,476	53,622
Number of label occurrences			
<i>ADJ</i>	29,418	3,541	3,468
<i>ADP</i>	71,074	8,523	8,332
<i>ADV</i>	14,878	1,710	1,763
<i>AUX</i>	11,061	1,200	1,304
<i>CCONJ</i>	12,211	1,454	1,439
<i>DET</i>	68,315	8,061	8,055
<i>INTJ</i>	151	10	16
<i>NOUN</i>	81,411	9,593	9,531
<i>NUM</i>	7,277	920	977
<i>PART</i>	122	14	18
<i>PRON</i>	20,397	2,370	2,417
<i>PROPN</i>	34,259	4,037	4,094
<i>PUNCT</i>	52,988	6,303	6,334
<i>SCONJ</i>	10,128	1,136	1,210
<i>SYM</i>	358	41	28
<i>VERB</i>	38,987	4,563	4,636
<i>X</i>	4	0	0
<i>acl</i>	8,580	1,054	983
<i>advcl</i>	7,326	811	839
<i>advmod</i>	15,617	1,794	1,850
<i>amod</i>	23,784	2,873	2,822
<i>appos</i>	6,096	717	704
<i>aux</i>	5,889	668	718
<i>case</i>	61,994	7,418	7,302
<i>cc</i>	12,354	1,466	1,455
<i>ccomp</i>	4,950	533	610
<i>compound</i>	1,279	156	137
<i>conj</i>	13,161	1,559	1,560
<i>cop</i>	4,560	468	519
<i>csubj</i>	988	92	94
<i>dep</i>	238	9	24
<i>det</i>	68,456	8,095	8,110
<i>discourse</i>	4	0	0
<i>dislocated</i>	3	0	0
<i>expl</i>	4,933	573	589
<i>fixed</i>	6,661	795	711
<i>flat</i>	11,670	1,331	1,435
<i>list</i>	5	0	16
<i>mark</i>	15,700	1,811	1,839
<i>nmod</i>	31,882	3,864	3,661
<i>nsubj</i>	24,023	2,741	2,877
<i>nummod</i>	4,806	603	596
<i>obj</i>	19,966	2,371	2,356
<i>obl</i>	26,409	3,157	3,227
<i>orphan</i>	1	0	0
<i>parataxis</i>	538	84	80
<i>punct</i>	52,981	6,301	6,331
<i>root</i>	14,287	1,654	1,721
<i>vocative</i>	1	0	0
<i>xcomp</i>	3,897	478	456

Table 38: Statistics of *es\_ancora*

Figure 38: Syntactic acquisition through pretraining of OLMo-2-7B on Spanish (*es\_ancora*).

### C.31. Swedish

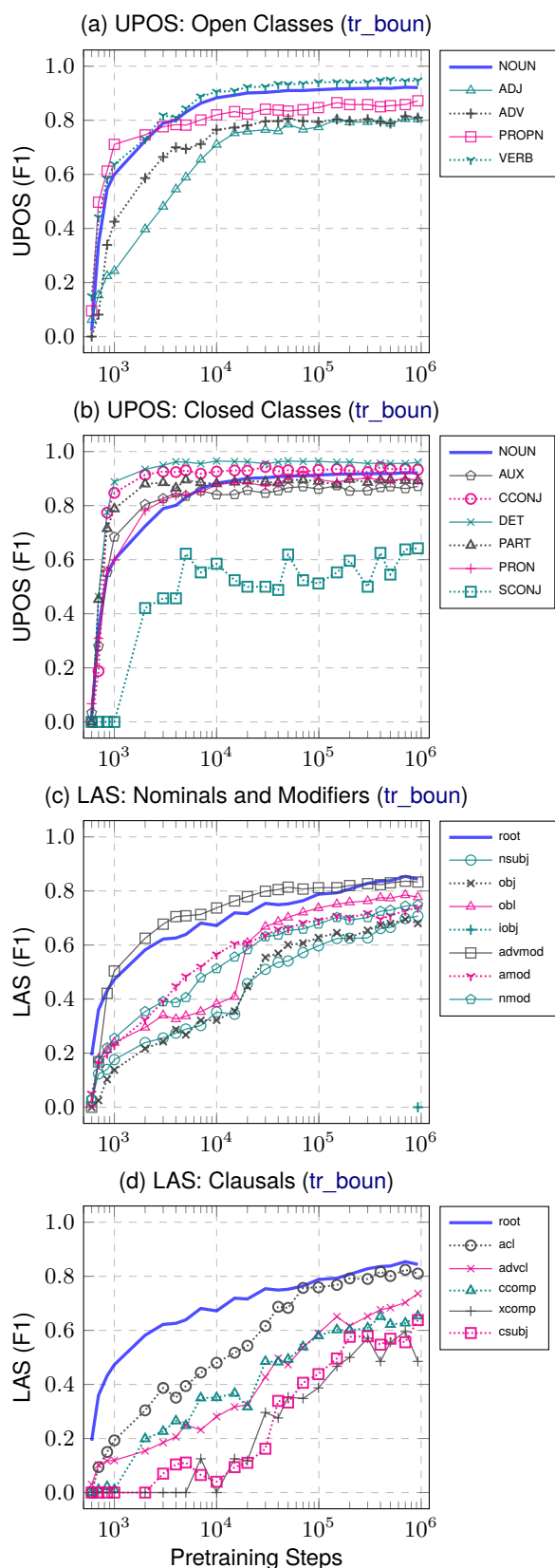


	Train	Dev	Test
Number of sentences	4,303	504	1,219
Number of words	66,646	9,797	20,377
Number of label occurrences			
<i>ADJ</i>	5,876	936	1,859
<i>ADP</i>	7,860	1,080	2,299
<i>ADV</i>	4,628	809	1,504
<i>AUX</i>	2,851	410	1,021
<i>CCONJ</i>	2,456	456	791
<i>DET</i>	3,375	499	1,003
<i>INTJ</i>	21	11	3
<i>NOUN</i>	16,096	2,190	4,711
<i>NUM</i>	1,324	59	357
<i>PART</i>	1,222	246	406
<i>PRON</i>	4,385	839	1,449
<i>PROPN</i>	1,081	36	243
<i>PUNCT</i>	7,303	962	2,104
<i>SCONJ</i>	1,368	243	491
<i>SYM</i>	75	10	1
<i>VERB</i>	6,725	1,011	2,134
<i>X</i>	0	0	1
<i>acl</i>	1,323	246	427
<i>advcl</i>	1,136	207	406
<i>advmod</i>	4,407	797	1,466
<i>amod</i>	4,251	630	1,288
<i>appos</i>	361	35	103
<i>aux</i>	1,713	237	614
<i>case</i>	7,318	978	2,121
<i>cc</i>	2,434	448	771
<i>ccomp</i>	358	47	108
<i>compound</i>	589	65	166
<i>conj</i>	3,054	484	898
<i>cop</i>	1,137	173	406
<i>csubj</i>	249	32	93
<i>dep</i>	1	0	0
<i>det</i>	3,363	501	1,008
<i>discourse</i>	12	8	2
<i>dislocated</i>	88	17	19
<i>expl</i>	381	43	94
<i>fixed</i>	104	23	26
<i>flat</i>	122	11	26
<i>goeswith</i>	0	0	1
<i>iobj</i>	105	21	37
<i>list</i>	5	0	0
<i>mark</i>	2,559	495	884
<i>nmod</i>	4,237	621	1,181
<i>nsubj</i>	6,105	873	1,972
<i>nummod</i>	1,009	49	281
<i>obj</i>	2,863	479	930
<i>obl</i>	4,778	613	1,386
<i>orphan</i>	25	2	8
<i>parataxis</i>	288	51	79
<i>punct</i>	7,303	962	2,104
<i>root</i>	4,303	504	1,219
<i>vocative</i>	0	1	0
<i>xcomp</i>	665	144	253

Table 39: Statistics of *sv\_talbanken*

Figure 39: Syntactic acquisition through pretraining of OLMo-2-7B on Swedish (*sv\_talbanken*).

### C.32. Turkish

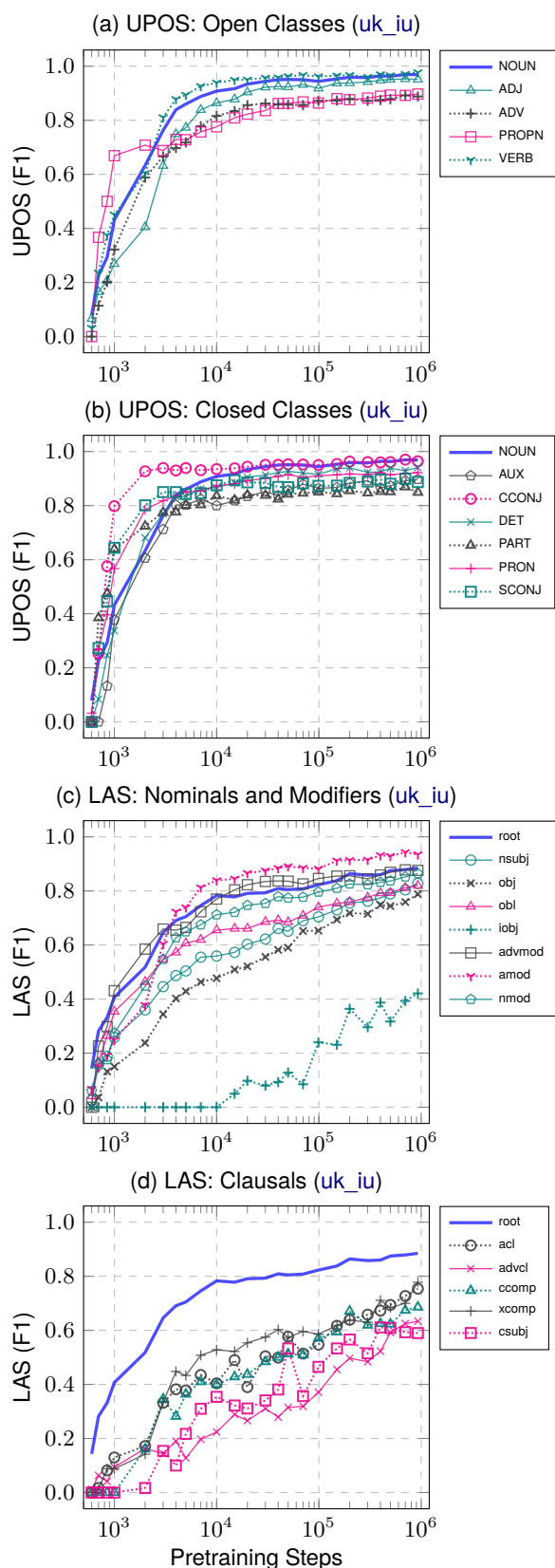


	Train	Dev	Test
Number of sentences	7,803	979	979
Number of words	100,713	12,289	12,210
Number of label occurrences			
<i>ADJ</i>	6,022	807	681
<i>ADP</i>	1,984	249	264
<i>ADV</i>	4,362	570	479
<i>AUX</i>	3,187	325	240
<i>CCONJ</i>	2,881	452	337
<i>DET</i>	4,048	505	546
<i>INTJ</i>	213	18	22
<i>NOUN</i>	30,686	3,713	3,952
<i>NUM</i>	2,105	225	276
<i>PART</i>	1,891	46	162
<i>PRON</i>	2,877	384	321
<i>PROPN</i>	6,438	766	677
<i>PUNCT</i>	16,715	2,043	2,028
<i>SCONJ</i>	179	15	26
<i>VERB</i>	17,121	2,171	2,199
<i>X</i>	4	0	0
<i>acl</i>	2,791	341	351
<i>advcl</i>	2,434	309	314
<i>advmod</i>	5,248	609	587
<i>amod</i>	5,650	759	766
<i>appos</i>	121	34	37
<i>aux</i>	425	26	11
<i>case</i>	2,214	271	294
<i>cc</i>	2,452	326	273
<i>ccomp</i>	1,247	174	175
<i>clf</i>	0	13	16
<i>compound</i>	3,101	366	385
<i>conj</i>	5,725	689	673
<i>cop</i>	2,455	262	192
<i>csubj</i>	467	43	72
<i>dep</i>	535	38	21
<i>det</i>	3,997	483	524
<i>discourse</i>	707	69	83
<i>dislocated</i>	17	3	3
<i>fixed</i>	56	13	16
<i>flat</i>	1,617	173	181
<i>iobj</i>	184	20	24
<i>list</i>	52	2	2
<i>mark</i>	179	9	14
<i>nmod</i>	10,138	1,173	1,170
<i>nsubj</i>	6,729	832	827
<i>nummod</i>	1,248	143	163
<i>obj</i>	5,914	752	740
<i>obl</i>	9,963	1,279	1,243
<i>orphan</i>	34	5	2
<i>parataxis</i>	207	22	23
<i>punct</i>	16,691	2,039	2,026
<i>root</i>	7,803	979	979
<i>vocative</i>	88	10	9
<i>xcomp</i>	224	23	14

Table 40: Statistics of *tr\_boun*

Figure 40: Syntactic acquisition through pretraining of OLMo-2-7B on Turkish (*tr\_boun*).

### C.33. Ukrainian



	Train	Dev	Test
Number of sentences	5,521	673	898
Number of words	92,927	12,606	17,217
Number of label occurrences			
<i>ADJ</i>	8,831	1,258	1,976
<i>ADP</i>	8,313	1,157	1,588
<i>ADV</i>	5,138	635	718
<i>AUX</i>	856	82	120
<i>CCONJ</i>	3,722	446	630
<i>DET</i>	3,581	461	630
<i>INTJ</i>	109	4	10
<i>NOUN</i>	21,571	3,263	4,537
<i>NUM</i>	1,164	211	390
<i>PART</i>	2,876	292	374
<i>PRON</i>	4,151	404	519
<i>PROPN</i>	2,484	406	633
<i>PUNCT</i>	17,702	2,423	3,130
<i>SCONJ</i>	1,838	245	261
<i>SYM</i>	83	8	17
<i>VERB</i>	10,133	1,209	1,559
<i>X</i>	375	102	125
<i>acl</i>	1,215	220	205
<i>advcl</i>	1,315	139	164
<i>advmod</i>	5,227	617	752
<i>amod</i>	7,065	1,022	1,659
<i>appos</i>	636	97	128
<i>aux</i>	208	20	27
<i>case</i>	8,302	1,155	1,584
<i>cc</i>	3,748	451	630
<i>ccomp</i>	475	55	89
<i>compound</i>	350	77	101
<i>conj</i>	4,877	566	890
<i>cop</i>	648	62	93
<i>csubj</i>	426	70	57
<i>det</i>	2,589	301	494
<i>discourse</i>	1,911	183	227
<i>dislocated</i>	11	7	1
<i>expl</i>	94	9	13
<i>fixed</i>	218	31	39
<i>flat</i>	1,430	289	441
<i>goeswith</i>	2	1	2
<i>iobj</i>	298	15	33
<i>list</i>	13	3	9
<i>mark</i>	1,809	241	265
<i>nmod</i>	6,403	1,211	1,902
<i>nsubj</i>	6,012	793	917
<i>nummod</i>	735	111	185
<i>obj</i>	4,422	589	701
<i>obl</i>	6,609	849	1,172
<i>orphan</i>	208	15	33
<i>parataxis</i>	1,282	156	225
<i>punct</i>	17,702	2,423	3,130
<i>reparandum</i>	1	0	0
<i>root</i>	5,521	673	898
<i>vocative</i>	95	1	6
<i>xcomp</i>	1,070	154	145

Table 41: Statistics of *uk\_iu*

Figure 41: Syntactic acquisition through pretraining of OLMo-2-7B on Ukrainian (*uk\_iu*).