Preferences for Idiomatic Language are Acquired Slowly — and Forgotten Quickly: A Case Study on Swedish

Jenny Kunz
Department of Computer and Information Science
Linköping University
jenny.kunz@liu.se

Abstract

In this study, we investigate how language models develop preferences for *idiomatic* as compared to *linguistically acceptable* Swedish, both during pretraining and when adapting a model from English to Swedish. To do so, we train models on Swedish from scratch and by fine-tuning English-pretrained models, probing their preferences at various checkpoints using minimal pairs that differ in linguistic acceptability or idiomaticity. For linguistic acceptability, we adapt existing benchmarks into a minimal-pair format. To assess idiomaticity, we introduce two novel datasets: one contrasting conventionalized idioms with plausible variants, and another contrasting idiomatic Swedish with Translationese. Our findings suggest that idiomatic competence emerges more slowly than other linguistic abilities, including grammatical and lexical correctness. While longer training yields diminishing returns for most tasks, idiom-related performance continues to improve, particularly in the largest model tested (8B). However, instruction tuning on data machine-translated from English — the common approach for languages with little or no native instruction data — causes models to rapidly lose their preference for idiomatic language.

1 Introduction

Even grammatically correct text can vary in how idiomatic or natural it sounds. Native speakers often prefer certain words or constructions over others, even when both are linguistically acceptable. These preferences influence how fluent or native-like a text *feels*.

In translation studies, the tendency of translated text to retain source-language features rather than adopting more natural target-language expressions is known as *Translationese* (Gellerstam, 1986). While initially observed in human translations, Translationese is now primarily discussed in relation to machine-translated text. We hypothesize that a similar phenomenon arises in language model outputs for non-English languages. Since English dominates LLM pretraining data, transfer effects — which are generally desirable! — can introduce anglicized patterns. Furthermore, parts of the training data in lower-resource languages may themselves be translations from English. This influence continues post-training, where steps such as instruction tuning are often performed using machine-translated datasets, as native data is rarely available (Muennighoff et al., 2023; Holmström and Doostmohammadi, 2023; Chen et al., 2024).

The idiomaticity of LLM outputs in non-English languages remains however understudied. This property is less straightforward to define or evaluate than grammaticality or factuality, as idiomatic preferences are more subjective and variable. Factors such as age, region, and exposure to English shape what speakers perceive as natural.

While previous work has examined Translationese in LLM-generated *translations* (Raunak et al., 2023; Li et al., 2025), to our knowledge, no existing research has investigated general idiomaticity preferences in LLMs. Consequently, there are no existing datasets for Swedish that address this issue explicitly. To address this gap, we introduce two Swedish datasets designed to probe LLMs' idiomatic preferences: one contrasts conventionalized idioms with plausible alternatives; the other contrasts Translationese Swedish with idiomatic alternatives.

We use these datasets to track small (135M, 8B) LLMs' preferences for idiomatic language during pretraining and continued pretraining. Our findings show that idiomatic language is acquired more slowly than other linguistic properties — slower than syntax, but also slower than appropri-

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1266–1285, 2026. https://doi.org/10.1162/TACL.a.727
Action Editor: Einat Minkov. Submission batch: 6/2025; Revision batch: 10/2025; Published 7/2026.
© 2026 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

ate lexical choices. Longer continued pretraining results in steady gains in idiomaticity for English-initialized models, whereas a model trained from scratch plateaus earlier and lower. This suggests that English pretraining not only boosts overall performance but also facilitates the gradual acquisition of idiomatic language use.

Finally, we examine the effect of instruction-tuning these models with machine-translated data. While performance on the linguistic acceptability benchmarks declines slightly, idiomatic preference drops strongly and early in tuning. This highlights that it is challenging to preserve idiomaticity when tuning the model on translated data.

Summary

We tackle the following two research questions:

RQ1 How well and how quickly do language models acquire idiomatic preferences, as compared to general linguistic acceptability?

RQ2 How does instruction tuning with machine-translated data influence a model's idiomatic preferences?

We make the following contributions:

1. A manually curated dataset of conventionalized idioms, where one keyword in each expression is replaced with either a semantically similar, high-frequency alternative or a variant inspired by idioms in higher-resourced languages.

2. A dataset of sentence pairs contrasting Translationese Swedish with idiomatic alternatives.[1]

3. Empirical findings demonstrating the slow but steady acquisition of preferences for idiomatic language, and their vulnerability to fine-tuning with translated data.

2 Related Work

This work is related to four main strands of research: how linguistic skills are acquired during pre-training (§2.1), how language transfer and adaptation are done and what the challenges are (§2.2), how minimal pair setups are used to probe linguistic competence (§2.3), and challenges and evaluations of idiomatic language in NLP (§2.4).

[1]Due to copyright restrictions, we cannot release this dataset publicly but it can be easily reproduced from the source book, and we can provide it upon request.

2.1 Skill Acquisition during Pretraining

Tracking how language models acquire linguistic skills during pretraining has been an active area of research, particularly for encoder-based models, with similar patterns now observed for decoder-based LLMs (van der Wal et al., 2025).

Prior work often contrasts syntactic skills with higher-level capabilities such as semantic understanding, factual recall, and commonsense reasoning. For instance, Liu et al. (2021) find that linguistic knowledge is acquired early and robustly, whereas factual and reasoning abilities emerge later and less reliably across domains. Similarly, Saphra and Lopez (2019) show that syntax is learned earlier than topic-level representations. They also link this to model architecture: lower layers, which converge earlier in training, encode lower-level linguistic information. Choshen et al. (2022) find that syntactic abilities are learned in a consistent order across different model sizes and architectures, and that this order even correlates with difficulty for human learners. In a multilingual context, Blevins et al. (2022) show that monolingual capabilities are learned early, while cross-lingual generalization develops later. While much of this work focuses on syntax, semantics, and reasoning, we are, to our knowledge, the first to compare the acquisition of idiomatic versus linguistically generally acceptable language.

In terms of methodology, many of these studies use probing classifiers (Hupkes and Zuidema, 2018) — small supervised classifiers trained to predict linguistic properties from internal representations — as in van der Wal et al. (2025), Liu et al. (2021), and Blevins et al. (2022). Saphra and Lopez (2019) instead measure representational similarity to task-specific models. Our work is closest in method to Choshen et al. (2022), who use a minimal pair setup; an approach we extend to the domain of idiomaticity.

2.2 Language Transfer

Enabling languages to transfer (linguistic and other) capabilities between each other is a central desideratum of multilingual language models (Costa-jussà et al., 2024; Hu et al., 2020). Multilingual models are however often evaluated on translated benchmarks, which may not fully capture the quality of their language-specific capabilities,e.g. due to missing cultural context (Kuulmets and Fishel, 2023). Moreover, performance con-

sistency across languages is a known challenge, with models typically performing better on high-resource languages and worse on morphologically rich (Arnett and Bergen, 2025) or typologically distant ones (Hu et al., 2020).

Beyond multilingual models, a growing line of work focuses on post-hoc adaptation, that is, continuing the pretraining of large English-based models on non-English data. For example, the Norwegian Mistral (Samuel et al., 2025) and AI Sweden LLaMA[2] are adapted versions of English-centric models pretrained on data in the languages of Norway and on the mainland Scandinavian languages, respectively. Samuel et al. (2025)'s results show that continued pretraining enables the effective use of larger models than what traditional scaling laws would predict, given the small size of available training data. This suggests that leveraging existing cross-lingual representations from English can be highly data-efficient.

2.3 Minimal Pair Probes

Minimal pair probes (Linzen et al., 2016; Marvin and Linzen, 2018) evaluate linguistic competence by presenting models with pairs of sentences that differ in only a minimal aspect — typically, one is grammatical or more natural, and the other is not. A model is expected to assign higher probability (i.e., lower perplexity) to the preferred variant (Marvin and Linzen, 2018). For example, consider this subject–verb agreement pair from Linzen et al. (2016): "The key is on the table." vs. "The key are on the table.". The first sentence is syntactically correct and should thus be assigned higher probability than the second, incorrect one.

Early work applied this method to recurrent neural networks (RNNs). Linzen et al. (2016) demonstrated that LSTMs (Hochreiter and Schmidhuber, 1997) can model subject–verb agreement by feeding sentences incrementally and comparing the probabilities of competing verb forms at the critical position. To test syntactic generalization rather than surface memorization, Gulordava et al. (2018) introduced "colorless green ideas"–style stimuli, replacing content words in natural sentences with random words of the same syntactic category. RNNs still performed remarkably well under this setup. Marvin and Linzen (2018) extended this approach to a broader set of syntactic phenomena, including reflexive binding and negative polarity licensing.

With the rise of Transformer-based models (Vaswani et al., 2017) like BERT (Devlin et al., 2019), minimal pair probing has been adapted to masked language modeling. Goldberg (2019) test BERT's syntactic knowledge using cloze-style minimal pairs, while Talmor et al. (2020) apply similar setups to probe commonsense reasoning, including age and object comparisons.

Several benchmark datasets have also been created to systematize this methodology. BLiMP (Benchmark of Linguistic Minimal Pairs) (Warstadt et al., 2020) provides automatically generated minimal pairs targeting twelve linguistic phenomena, such as determiner–noun agreement, anaphora binding, and ellipsis. These benchmarks provide a controlled setting for evaluating specific linguistic capabilities and are especially suited to tracking how models acquire such skills over training.

2.4 Idioms

Conventionalized idioms are fixed expressions whose meanings deviate from their literal forms. As culturally grounded and figurative constructions, idioms can be challenging for both human learners and language models. Figurative language — including idioms — has long been recognized as a bottleneck in natural language understanding (Shutova, 2011).

While some idioms have cross-lingual equivalents due to shared religious or literary heritage (e.g., the biblical idiom "cast pearls before swine" has the Swedish equivalent "kasta pärlor åt svin"), others are culture-specific. Literal translations frequently produce unnatural output, even when the intended meaning is clear. Swedish idioms such as "gå åt skogen" ("go to the forest," meaning "go wrong") or "gå av stapeln" ("to be launched", originally a maritime term) reflect ties to nature, local customs, and daily life.

This cultural specificity makes idioms particularly challenging for multilingual LLMs. In a qualitative study, Pedersen et al. (2025) show that LLMs explain English metaphors more effectively than Danish-specific ones, raising concerns about linguistic diversity in LLMs. Idioms are also a well-known challenge in machine translation, where literal renderings are common (Fadaee et al., 2018; Dankers et al., 2022; Liu et al., 2023),

[2] https://my.ai.se/resurser/3070; Last accessed: Feb 03, 2026

and a lack of correct interpretation even affects downstream tasks tasks such as textual entailment (Chakrabarty et al., 2021; Stowe et al., 2022).

Idioms are more difficult that other linguistic phenomena even in human second-language acquisition. Learners tend to avoid using idioms, especially opaque or culture-bound ones, and prefer transparent idioms or those with equivalents in their native language (Laufer, 2000; Tarabrina, 2019). As Pinnavaia et al. (2002) notes, idioms are the "patrimony of a culture", deeply embedded in specific sociolinguistic and geographical contexts.

Most NLP work on idioms focuses on detection (Tayyar Madabushi et al., 2022; Yayavaram et al., 2024), disambiguation (Kurfalı and Östling, 2020; Garcia et al., 2021), translation (Baziotis et al., 2023), or explanation (Gluck et al., 2025; Pedersen et al., 2025). In contrast, we address active idiomatic competence: the ability to produce idiomatic expressions from inputs that give a definition or explanation of the idiom. This represents a higher bar than translating or explaining a given idiom as it ensures that the model is able to actively *produce* the idiom, while a correct interpretation can often be inferred from context. Perhaps most similar to our work is a subset of the Norwegian benchmark NorEval (Mikhailov et al., 2025), where the last word of an idiom is cut off and has to be predicted. The EuroEval benchmark (Nielsen, 2023) even includes a multiple-choice variant of this dataset, including LLM-generated incorrect alternatives.

Finally, Liu and Lareau (2024) find that CamemBERT predicts idiomatic tokens (in French) *more* reliably than literal ones, especially in longer expressions. This suggests that idioms, once learned, function as cohesive units with strong internal constraints, making them more predictable despite their semantic opacity.

3 Minimal Pair Benchmarks

We develop two sets of minimal pair benchmarks: A baseline with *linguistic acceptability* datasets from existing work (§3.1), and one with our own datasets for *idiomatic preferences* (§3.2).

3.1 Baseline: Linguistic Acceptability

As a baseline for monitoring the acquisition of idiomatic preferences, we construct a linguistic acceptability benchmark. We follow the minimal pair paradigm, where each item consists of one correct sentence and a closely related ungrammatical counterpart. This formulation allows us to assess whether models prefer the well-formed sentence over its erroneous variant — an approach grounded in prior work on linguistic minimal pairs (Linzen et al., 2016; Marvin and Linzen, 2018).

In contrast to English, subject-verb agreement (the most common target in minimal pair probing) does not exist in modern Swedish, so that we cannot make use of this classical task. Instead, we focus on naturally occurring and synthetically induced grammatical errors that reflect Swedish morphosyntax and word order phenomena.

3.1.1 Naturally Occurring Errors

The DaLAJ (Dataset for Linguistic Acceptability Judgments) benchmark (Volodina et al., 2023) provides a basis for constructing minimal pairs from real-world learner language. It is based on essays written by second-language learners of Swedish across different proficiency levels. Each sentence in the dataset has been manually corrected, modified so that it contains exactly one error, and annotated for the type of linguistic error.

As the DaLAJ dataset contains sentences with error annotations, but not a direct reference to the correction, we need to align each incorrect sentence with its correct correspondence to construct minimal pairs. We do so by searching the correct sentence with the smallest Levenshtein distance (Levenshtein, 1966) for each incorrect sentence. We concatenate the original training, validation and test splits to maximize coverage, resulting in 20,000 sentence pairs (with 6,500 unique correct sentences) of which 2122 are *punctuation* errors, 3924 are *orthography* errors, 3754 are *lexical choice* errors, 6482 are *morphological* errors and 4666 are *syntax* errors.

3.1.2 Synthetic Corruptions

The ScaLA dataset (Nielsen, 2023) provides another source of acceptability judgments by using synthetic corruption of grammatical sentences. In this dataset, well-formed sentences are modified by introducing controlled perturbations: Either two adjacent tokens are swapped (called *flip neighbours* in this paper), or a single token is deleted (*delete*). For each corrupted sentence, we identify its closest grammatical source sentence, again using the minimum Levenshtein distance, to form minimal pairs. As with DaLAJ, we combine all splits (train, validation, test) to construct the fi-

nal set, resulting in 552 pairs for the *flip neighbors* and 601 pairs for the *delete* category.

3.2 Idiomatic Preferences

As no existing benchmarks evaluate a model's preference for idiomatic language use in Swedish, we introduce two new datasets using the same minimal pair setup as in the syntactic tasks.

3.2.1 Swedish Conventionalized Idioms

We construct a benchmark of Swedish idioms initially based on a completion task, and then reformulated into minimal pairs. Our aim is to test whether models prefer the idiomatic keyword over a plausible but non-idiomatic alternative. We briefly describe the creation process here; more details and examples can be found in Appendix A.

Data Collection. We manually curate a list of Swedish idioms and their definitions by browsing relevant Wiktionary categories and lists[3] and selecting those that:

- can be reasonably framed as a cloze task with a single missing word;
- have a clearly identifiable idiomatic answer (without being overly obvious or trivially inferrable from context).

This resulted in 479 idioms. We formulated each instance as a short sentence or question that includes the idiom's definition or explanation and is aimed at eliciting a specific idiomatic keyword. An example for such a sentence is: *Någon som är väldigt rik är rik som ett troll* ("Someone who is very rich is rich like a troll"). In this formulation, clarity was prioritized over syntactic or lexical diversity. The dataset was initially compiled by one annotator and then reviewed by two others for the full dataset $\text{Idioms}_{\text{all}}$.

Minimal Pair Construction. For each idiom, we manually construct a plausible but less idiomatic alternative keyword. Whenever possible, the less idiomatic option (the distractor) is based on a similar idiom in a closely related high-resource language, such as English or German. For example, in Swedish the idiom is *jämföra äpplen och päron* ("comparing apples and pears"), whereas in English it is "comparing apples and oranges." Thus, *apelsiner* ("oranges") becomes the distractor. To find such counterparts, annotators use their linguistic knowledge, web search, and resources such as the Oxford Dictionary of Idioms. When no direct cross-lingual idiom exists, the distractor is chosen from semantically and syntactically plausible alternatives in the same lexical category. Absurd alternatives are discouraged.

We then use the chatbot assistant Claude (Anthropic, 2024) to rewrite prompts into minimal pair sentences with both the intended idiomatic keyword and the distractor inserted. These keywords can appear in any part of the sentence. All generated pairs are manually checked and revised to ensure fluency and idiomatic language.

Challenge Subset. We additionally construct a challenge subset $\text{Idioms}_{\text{challenge}}$ of idioms that:

1. have no exact counterparts in either English or German;
2. lack strong surface cues that would make them easy to guess.

To verify the absence of equivalents in English and German, our annotators conduct web searches and consult the Oxford Dictionary of Idioms. This subset is reviewed and enhanced by three annotators, including two native speakers of German.

Release. We make this dataset publicly available on the Hugging Face Hub under the identifier *liu-nlp/swedish-idioms* and on https://github.com/jekunz/idiomatic-language.

3.2.2 Translationese

Translationese refers to the systematic influence of a source language on the target language in translation. The concept was introduced by Gellerstam (1986), who defined it as a set of linguistic patterns that distinguish translated text from native language use.

According to Baker (1993), professional translations often exhibit a strong preference for conventional grammaticality and style. This bias may lead to stylistically flattened or simplified output that, paradoxically, results in lower perplexity compared to more idiomatic or natural native expressions, making a dataset comparing Translationese with non-Translationese more challenging.

Dataset Construction. We base our dataset on Katourgi (2022), a popular science book that systematically contrasts "naturally occurring"

[3] https://sv.wikipedia.org/wiki/Lista_Ãűver_svenska_idiomatiska_uttryck, https://sv.wiktionary.org/wiki/Appendix:Idiomatiska_uttryck/Svenska, https://sv.wiktionary.org/wiki/Kategori:Svenska/Idiomatiskt; Last access: Feb. 03, 2026

Swedish translations influenced by English with more natural Swedish formulations proposed by the author. For example, it contrasts the Translationese sentence *Låt oss ta en titt*, which is a literal translation of the English sentence "Let us take a look", with the more idiomatic *Kom så kollar vi* (Literally: "Come then we check"). We transcribe all sentence pairs from the book into a structured dataset. In cases where two alternative idiomatic versions are provided, we include both, resulting in a total of 967 sentence pairs. We denote this dataset $\text{Translationese}_{\text{all}}$.

Subset for More Reliable Evaluation. It is important to note that the examples from the book are not per se minimal: the idiomatic rewritings tend to be shorter and more freely adapted than the original Translationese versions. On average:

- The Translationese variants contain 5.09 whitespace-tokenized tokens.
- The idiomatic alternatives contain 4.41 tokens.

Even though we normalize the perplexity by sequence length (see Section 4.3), this difference can add systematic biases, e.g. by longer sequences containing more easy-to-predict function words. In subword tokenizers (which most current models, including those on our experiments, use), later subwords in a word have lower perplexities because they have a strong cue, that is, the preceding subword(s) (Chang et al., 2024).

For a fair minimal pair evaluation, we therefore extract a filtered subset $\text{Translationese}_{\text{filtered}}$ based on the following two criteria:

1. The two variants have the same length according to the tokenizer of the respective model.
2. The idiomatic variant is clearly preferable based on general Swedish language norms, even without additional context. This criterion has been implemented with one primary rater and one reviewer of the annotations.

The intersection of both subsets results in 98 samples for the SmolLM models and 134 samples for the Llama model.[4]

[4]Note that due to the different subsets resulting from different tokenizers, the results of the SmolLM models are not directly comparable to the Llama model. As we focus on intra-model comparisons, this is of less importance in this paper. For inter-model comparisons, basing the length restriction on whitespace tokenizer can be a compromise. This approach results in a larger dataset but retains a stronger bias towards the Translationese samples than the model-tokenizer based filtering, as visible in Table 3 in Appendix B.

Data Availability. Because the dataset is derived from a copyrighted book, we are unable to publish it freely. However, it can be reproduced by transcribing all sentences from Katourgi (2022); we share the indices of the samples included in the challenge set at https://github.com/jekunz/idiomatic-language. We are able to share both the full and filtered datasets upon request with researchers with legal access to the source material.

3.3 Annotators

The primary annotators and raters of the datasets are two student assistants (paid after collective agreement) who are native-level speakers of Swedish and have undergone basic training in linguistics in their Cognitive Science study program. The first author reviewed all annotations.

4 Experimental Setup

All models and checkpoints used in our experiments are available on the HuggingFace Hub in the collection https://huggingface.co/collections/jekunz/idiomatic-language-acquisition. The code is available at https://github.com/jekunz/idiomatic-language.

4.1 Training

To answer our research questions, we train in two phases: (continued) pre-training (RQ1) and instruction tuning (RQ2).

Pretraining (RQ1). We use the Swedish portion of the deduplicated FineWeb2 dataset (Penedo et al., 2025), which contains approximately 25 billion tokens, to train two models with 135M parameters based on the SmolLM2 architecture (Allal et al., 2025): $\textit{SmolLM}_{\textit{CPT}}$ is a continued pretraining (CPT) of the original SmolLM2 135M model, which was trained on the English Fineweb dataset (Penedo et al., 2024). $\textit{SmolLM}_{\textit{scratch}}$ is trained only on Swedish from random initialization.

Both models are trained with identical hyperparameters: we use the AdamW optimizer (Loshchilov and Hutter, 2019), a cosine learning rate scheduler with a warmup rate of 5%, an effective batch size of 256, and a context window size of 8192 tokens (the same as in the SmolLM2 pre-training). Checkpoints are saved every 200 training steps, corresponding to every 200 million tokens processed. For training on the full 25

billion tokens, the models required approximately 720 GPU hours on NVIDIA A100 40GB cards.

Instruction Tuning (RQ2). To see the effect of instruction tuning with translated data, we translate Smol-Smoltalk (Allal et al., 2025), a synthetic instruction-tuning dataset specifically adapted for models with less than 1B parameters, to Swedish using Gemma-27B (Mesnard et al., 2024)[5]. In the prompt, we instruct the model to phrase the translations as naturally and idiomatically in Swedish as possible. A manual inspection of 20 translated conversations indicated reasonable translation quality for current standards, with little obvious Translationese. We train the SmolLM models on the dataset (456,780 conversations) for one epoch with the same parameters as for the pretraining, requiring 12 A100 hours per model.

4.2 Existing Models

We evaluate AI-Sweden-Llama, a version of Llama 3 8B (Meta AI, 2024) further trained on the Swedish, Danish, and Norwegian subsets of the Nordic Pile (Öhman et al., 2023). The model has ten publicly released checkpoints recorded after 1500, 2700, 3900, 5325, 6550, 8200, 11525, 14375, 16000, and 18833 training steps, allowing us to tackle RQ1. For comparison, we also evaluate the base Llama 3 8B model, which is officially an English model, as checkpoint 0. A certain exposure to Swedish already in the base model's pre-training is however likely. The model was instruction-tuned on a Swedish translation of OpenHermes 2.5 (Teknium, 2023), produced using GPT-3.5[6], which allows us to tackle RQ2.

4.3 Metric

To evaluate model preference, we compute the average per-token perplexity of the full sentence $x_{1:N}$ of length N:

$$\text{ppl}(x) = \exp\left(-\frac{1}{N}\sum_{t=1}^{N}\log p(x_t \mid x_{<t})\right).$$

For each minimal pair, we calculate perplexity for both sentences. The sentence with the lower perplexity is taken to be the model's preferred choice. Accuracy is then defined as the proportion of $\hat{y}$ in which the model prefers the better (i.e., grammatical or more idiomatic) sentence x_{better}:

$$\hat{y} = \begin{cases} 1 & \text{if ppl}(x_{\text{better}}) < \text{ppl}(x_{\text{worse}}) \\ 0 & \text{otherwise.} \end{cases}$$

5 Results and Analysis

To answer RQ1, we analyze how models acquire idiomatic language during pre-training in Section 5.1. Then, tackling RQ2, we examine the impact of instruction fine-tuning on these capabilities in Section 5.2. Finally, we analyze *why* idiomatic language preferences are more sensitive than general linguistic competence in Section 5.3.

5.1 *Acquisition* of Linguistic Skills (RQ1)

All three models are better at distinguishing linguistically (syntactically, orthographically and lexically) correct from incorrect Swedish than they are at distinguishing idiomatic from unidiomatic language: The scores for the linguistic acceptability datasets DaLaJ and ScaLA in Table 1 are higher across models than for the Idioms and Translationese datasets. Idiomaticity remains challenging to acquire and improves slowly during language adaptation, as visible in the left column in Figure 1 (Subfigures 1a, 1c and 1e).

We observe consistent performance gains with English pretraining compared to $\text{SmolLM}_{\text{scratch}}$. English pretraining does not hinder the learning of idiomaticity (or any other linguistic skill included in our evaluations) but provides a better base for learning fine-grained Swedish capabilities.

The base LLaMA 3 model (checkpoint 0 in Figure 1e), despite being trained primarily on English, already performs well across linguistic acceptability tasks: It has more than 95% of its final accuracy on the DaLaJ and ScaLA benchmarks already before adaptation, as seen in Table 2. However, its lexical and idiomatic competence is markedly weaker, and requires continued pretraining to acquire, as we can see in Figure 1e.

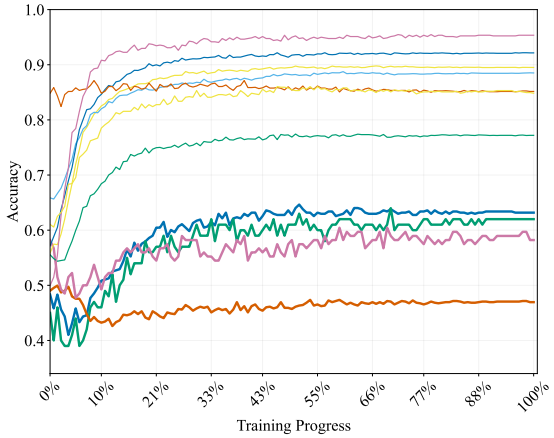
5.1.1 Idioms

This section examines how models acquire the conventionalized idioms of the *Idioms* datasets, and how this differs from their ability to learn lexical or syntactic norms.

Idioms versus Lexical Correctness. Interestingly, the LLaMa model not only catches up in idiom competence but *surpasses* its performance on

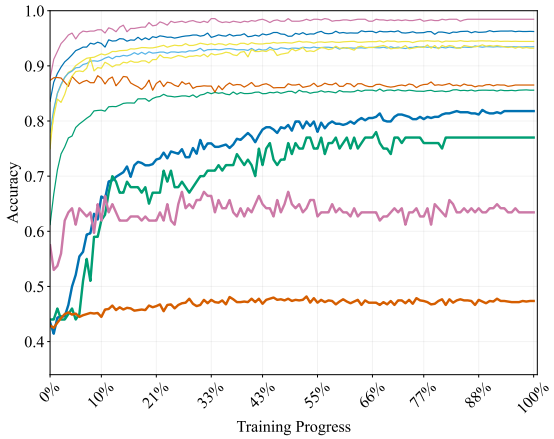
[5]We choose Gemma due to its multilingual capabilities while being relatively cheap to use: The translation of the dataset cost 144 A100 hours.

[6]Information via personal communication with model creator Tim Isbister.

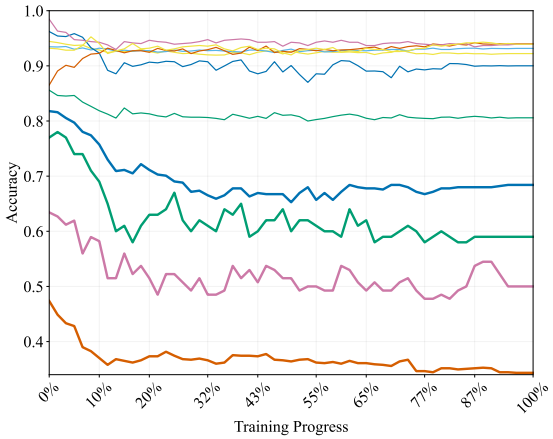


(a) SmolLM2: Training from scratch.

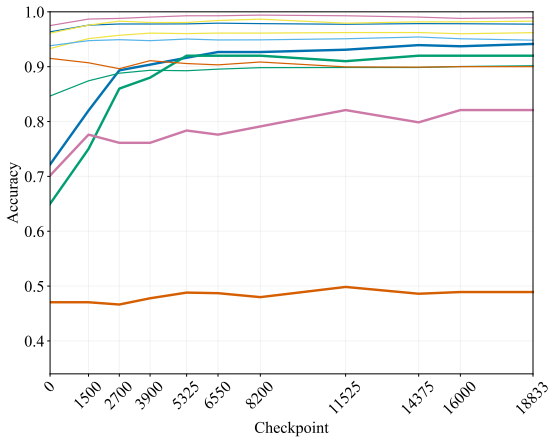
(b) SmolLM2 from scratch: Instruction tuning.



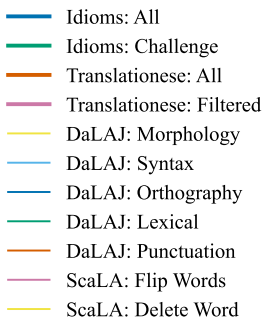
(c) SmolLM2: Continued pretraining from English.



(d) SmolLM2 CPT from English: Instruction tuning.



(e) AI Sweden LLaMA 8B: Continued pretraining.



(f) Legend for all Figures: Datasets and subsets.

Figure 1: Accuracy across training checkpoints for all SmolLM2 135M variants and for AI Sweden LLaMA 8B. The subfigures on the left are for (continued) pre-training with plain text, the subfigures on the right for instruction tuning. Subfigure 1f contains the legend for all plots.

		$\text{SmolLM}_{\text{scratch}}$		$\text{SmolLM}_{\text{CPT}}$		AI Sweden Llama	
		Before	After	Before	After	Before	After
Final Accuracy	$\text{Idioms}_{\text{all}}$	63.17	48.53	81.79	68.41	94.14	88.91
	$\text{Idioms}_{\text{challenge}}$	62.00	40.00	77.00	59.00	92.00	84.00
	$\text{Translationese}_{\text{all}}$	46.94	36.91	47.36	34.33	48.91	35.36
	$\text{Translationese}_{\text{filtered}}$	58.20	50.00	69.40	50.00	82.08	59.70
	$\text{DaLAJ}_{\text{lexical}}$	77.19	72.58	85.58	80.58	90.17	87.45
	$\text{DaLAJ}_{\text{others}}$	88.82	86.60	92.65	92.35	94.69	94.10
	ScaLA (avg.)	90.03	86.38	95.82	93.92	98.61	98.19

Table 1: Final accuracy after (continued pre-) training on Swedish. We report results *before* and *after* instruction tuning for all models and tasks.

	$\text{SmolLM}_{\text{scratch}}$	$\text{SmolLM}_{\text{CPT}}$	AI Swe. Llama
$\text{Idioms}_{\text{all}}$	6200	11600	3900
$\text{Idioms}_{\text{chall.}}$	8000	11400	3900
$\text{Transl.}_{\text{all}}$	0	3000	3900
$\text{Transl.}_{\text{filtered}}$	0	1200	5325
$\text{DaLAJ}_{\text{lexical}}$	4400	2400	1500
$\text{DaLAJ}_{\text{others}}$*	3933	1266	0
ScaLA (avg.)	3700	1300	0

Table 2: Training step at which each model reaches 95% of its total accuracy. (* We exclude DaLAJ punctuation here because it is 0 for all models.)

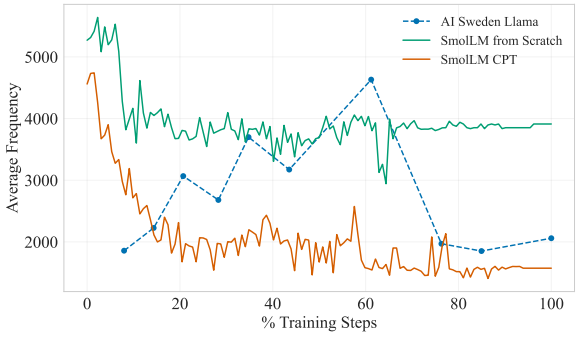


Figure 2: Average frequencies of idioms mispredicted at each checkpoint in the FineWeb2 corpus. Average over all idioms in the dataset: 5766.

the lexical acceptability benchmark ($\text{DaLAJ}_{\text{lexical}}$) relatively early in training (by step 2,700). These findings are in line with Liu and Lareau (2024), who show that idiomatic expressions in French are easier to predict in context, due to their strong internal constraints. However, this pattern does not hold for the smaller models, and suggests that both a high learning capacity (parameter count) and extensive Scandinavian-language training data are key enablers for learning conventionalized idioms.

Learning Curves. For AI Sweden LLaMA and $\text{SmolLM}_{\text{CPT}}$, idiom preferences continue to improve steadily throughout training, even as performance on other tasks (including $\text{DaLAJ}_{\text{lexical}}$, the most similar task) plateaus. This suggests that the Idioms dataset requires more training data and longer exposure to perform well on, but remains learnable with sufficient capacity and training. In contrast, $\text{SmolLM}_{\text{scratch}}$ plateaus early on the Idioms tasks and reach significantly lower performance levels, suggesting that the rarer idioms are not successfully learned.

The commonly observed breakpoint around 2–5B tokens, which is linked to the emergence of circuits responsible for core linguistic structure (Tigges et al., 2024; van der Wal et al., 2025), is reflected in the linguistic acceptability tasks, where performance jumps occur early (roughly at 10–25% of training). Idiom acquisition, by contrast, lacks a sharp inflection point and instead improves gradually, with substantial gains continuing well beyond the convergence of other tasks. This suggests that idioms require sustained exposure and are acquired more slowly than structural aspects of language.

Effect of Idiom Frequency. We analyze the frequency of idioms in the FineWeb2 corpus using a lexical search, identifying each idiom by the shortest substring that likely uniquely identifies it.[7] For

[7]This method is imperfect: it captures only a subset of instances for compositional idioms and may fail to fully exclude literal uses. A more recall-optimized strategy would however require careful engineering for each idiom and is outside the scope of this article.

each training checkpoint, we collect the idioms predicted incorrectly and report their average frequency in FineWeb2 in Figure 2.

For both SmolLM models, incorrectly predicted idioms initially have a higher average frequency and then gradually converge towards a lower frequency. The model trained from scratch converges faster and stabilizes at a higher frequency, whereas the continually pre-trained model's curve decreases slightly almost until the end. These trends closely mirror the training curves for idioms shown in Figures 1a and 1c: as performance improves, the average frequency of mispredicted idioms declines. In contrast, AI Sweden Llama shows a different pattern: the average frequency of mispredicted idioms is low at the earliest checkpoints, increases at intermediate steps, and then declines again toward the final checkpoints. Moreover, idioms that remain incorrectly predicted after full training are not less frequent than those missed at the first checkpoint: the average frequency of mispredicted idioms is 1857 at the first and 2059 at the final checkpoint. This difference could have two explanations. First, AI Sweden Llama is trained on a different, non-public corpus, which may have a distinct idiom distribution. Second, its higher learning capacity and overall better performance mean that fewer idioms are mispredicted, so other factors beyond frequency may play a larger role. Examining the mispredicted idioms for AI Sweden Llama reveals a pattern: many rarer, non-compositional idioms that are initially missed are eventually learned, likely because their strong lexical cues make them easier to memorize from limited examples compared to more compositional idioms. For instance, the idiom *inte för allt smör i Småland* (literally, "not for all butter in Småland", roughly equivalent to English "not for all the money in the world") appears only 1,269 times in the corpus, making it a quite rare idiom, yet its highly non-compositional form and distinctive word sequence allow the model to learn it with relatively small exposure: It is mispredicted at the first checkpoint, but then memorized.

Swedish-specific idioms ($\text{Idioms}_{\text{challenge}}$) show a learning trajectory similar to that of $\text{Idioms}_{\text{all}}$, although with slightly lower overall performance. This suggests that idioms that do not exist in the higher-resource languages English and German are learned at the same pace as those with English or German counterparts. We find no evidence that the model relies on translation cues; idiomatic competence appears to stem from direct exposure rather than inference from English equivalents. Even in $\text{SmolLM}_{\text{scratch}}$, performance on $\text{Idioms}_{\text{challenge}}$ remains lower.

Error Analysis. We manually inspect the idioms that the best-performing model, AI Sweden LLaMA, does not prefer over the corrupted counterpart. A recurring pattern is that many of these idioms are infrequently used in contemporary Swedish. For instance, *glad i hatten* (literally "happy in the hat", meaning "drunk") is widely recognized by Swedish speakers (according to an informal survey among friends and colleagues) but rarely used in active speech. Other idioms, such as *få korgen* (literally "get the basket", equivalent to the English "get the mitten", meaning "to be turned down"), are unfamiliar to some speakers altogether, and considered old-fashioned by those who do know the expression. In such cases, it may be reasonable that the model assigns these expressions a low probability. We conclude that AI Sweden Llama learns Swedish idioms that are in active use today to a large extent.

5.1.2 Translationese

The model's performance on the full Translationese set ($\text{Translationese}_{\text{all}}$) is extremely low — well below chance — indicating a strong bias toward more literal, translationese-like phrasing. However, since the Translationese examples tend to be shorter on average, this preference may reflect structural characteristics rather than genuine linguistic bias. For this reason, we focus our analysis on $\text{Translationese}_{\text{filtered}}$.

Learning Curves. While $\text{Translationese}_{\text{filtered}}$ yields higher scores than $\text{Translationese}_{\text{all}}$, it still performs worst among all tasks. Notably, $\text{Translationese}_{\text{filtered}}$ (and even $\text{Translationese}_{\text{all}}$) reaches 95% of its final performance earlier than any other dataset in training for the SmolLM models, as shown in Table 2. Nonetheless, overall performance remains low, failing to surpass 70%.

Interestingly, for the AI Sweden LLaMA model, 95% of peak performance on the Translationese task is reached last among all tasks. We learn that *if* (and only if) a model has sufficient capacity to learn to prefer idiomatic Swedish over Translationese, this process requires significantly longer adaptation time than learning to prefer grammatically correct Swedish.

Error Analysis. Despite being the strongest model overall, AI Sweden LLaMA frequently assigns lower perplexity to clearly Translationese alternatives that would be dispreferred by native speakers. For example, in the sentence "I enjoy drawing", the model favors the more literal translation *Jag njuter av att rita*, where the verb *njuta* is a direct equivalent of "enjoy". However, the more idiomatic phrasing in Swedish is the alternative, *Jag tycker om att rita*. Similarly, when translating "We only have money for noodles", the model prefers the literal transfer of the expression, *Vi har bara pengar till nudlar*, over the more natural expression *Vi har bara råd med nudlar*.

In some cases, the preferred variants are not only unnatural but also incorrect. For instance, for "hidden costs", the model selects *gömda kostnader*, which is not used idiomatically in Swedish. The correct expression is *dolda kostnader*, as the verb *gömma* typically refers to the physical act of hiding an object, rather than abstract concealment.

5.2 Effect of Instruction Tuning (RQ2)

Across almost all tasks (with the exception of $\text{DaLaJ}_{\text{punctuation}}$, which we will discuss later in this section), we observe a decline in performance following instruction tuning, visible in Figures 1b and 1d, and in Table 1. The largest drop occurs on $\text{Translationese}_{\text{filtered}}$, indicating that the models' previously learned preference for idiomatic phrasings over Translationese is reversed again. For the SmolLM models, performance drops to the level of random guessing (50%), suggesting that the ability to prefer idiomatic over literal translations is entirely lost during instruction tuning.

For the larger AI Sweden LLaMA model, syntactic, morphological, and orthographic accuracy on ScaLA and DaLaJ remains largely unaffected. This suggests that core grammatical competence is robust to instruction tuning with translated data[8], provided the model has sufficient capacity. Lexical phenomena are somewhat more impacted, while the Idioms datasets and $\text{Translationese}_{\text{filtered}}$ degrade substantially.

This finding aligns with prior work showing that instruction tuning can harm general language modeling performance, as measured by perplexity (Jin and Ren, 2025; Holmström and Doostmohammadi, 2023). However, our results also identify which properties are most vulnerable to tuning with machine-translated data: fine-grained idiomatic competence is especially affected, while general linguistic abilities such as syntax and morphology remain relatively stable.

In contrast to our results, Waldis et al. (2024) even report *improvements* in handling syntactic and morphological minimal pairs after instruction tuning. However, crucially, their work is on English, where instruction data is either natively written or generated by LLMs highly competent in English, and never machine-translated. Our findings highlights a key limitation of instruction tuning with machine-translated corpora.

Punctuation stands out in our results: As seen in Figure 1, the accuracy on $\text{DaLaJ}_{\text{punctuation}}$ initially decreases during continued pre-training but *increases* after tuning with translated data. This suggests that the punctuation standard encoded in the DaLaJ benchmark may align more closely with English norms than with common Swedish usage. Since this lies outside the primary scope of the paper, we leave an error analysis for future work.

Implications. The learning (or rather, forgetting) curves in Figure 1 suggest that instruction tuning-data volume may not offer a simple trade-off between preserving linguistic competence and learning instruction-following behavior: The performance on the tasks probing for idiomatic language drops already *early* in the instruction-tuning process. However, alternative regularizing hyperparameters, a lower learning rate, or replay of pretraining data in idiomatic Swedish may offer paths for preserving idiomatic skills, and are promising directions for future work.

5.3 Why is Idiomatic Language Sensitive?

As discussed in Section 5.1.1, many of the idioms in our dataset are relatively rare in the pretraining data. We hypothesize that this is the result of idioms, and idiomatic language in general, being more prevalent in speech than in writing. It has long been established that speech differs fundamentally from text (Woolbert, 1922), with key differences driven by the expected level of formality, the degree of planning, and the number and familiarity of interlocutors (Redeker, 1984). For instance, Allwood (1998) shows that spoken Swedish employs a smaller vocabulary and relies more heavily on pronouns, conjunctions, adverbs,

[8] At least for a language like Swedish, where machine translations produce mostly correct language, and likely even more standardized language than humans.

and interjections, whereas written Swedish makes greater use of prepositions, nouns, and adjectives.

This distinction is relevant not only for the Idioms dataset but also for the Translationese dataset, where certain idiomatic alternative translations occur more frequently in speech, while in many more formal forms of written text, one may prefer styles that differ less across languages.

LLMs have been observed to favor a formal and impersonal style at least in English (Mitrović et al., 2023). Despite the growing role of web-crawled data that includes many domains, encyclopedic and curated text sources remain a substantial part of pre-training. As a result, idiomatic language *is* learned by our models, but more slowly. Instruction tuning then further amplifies this tendency: not only does it shift models toward a Translationese style due to the machine-translated data, but it also encourages more formal phrasing overall due to the structured, technical nature of the instructions and answers in the dataset. We speculate that this readjustment reduces the models' probability to produce idiomatic alternatives.

6 Conclusion

In this work, we introduced two new datasets for measuring how idiomatic the language distribution of LLMs is for Swedish. The first dataset pairs conventionalized idioms with variants where a keyword in each expression is replaced by an alternative, often inspired by idioms in related higher-resourced languages. The second dataset contains sentence pairs contrasting Translationese Swedish with more idiomatic alternatives.

Our first research question (RQ1) asked how idiomatic preferences are acquired during pretraining, compared to preferences for generally acceptable linguistic forms. Our results indicate that "baseline" linguistic skills such as syntactic and morphological preferences are learned early and converge quickly. Idiomatic language, by contrast, is harder to acquire and shows a slow, gradual improvement over the course of pre-training.

We find no evidence that models infer conventionalized idioms from English equivalents: idioms that exist in both English and Swedish are learned at the same pace as Swedish-specific ones. Instead, English pretraining appears to provide a stronger foundation for memorizing idioms when they appear in the data: While models with English pretraining continue to improve over time, our model trained from scratch stops improving early and at a lower performance.

Distinguishing Translationese from natural Swedish is the most challenging of our tasks. Even the strongest model tested often assigns a higher probability to literal, English-like phrasing even over clearly more idiomatic ones.

Our second research question (RQ2) examined the influence of machine-translated instruction-tuning data on idiomatic preferences. We find that this setup substantially weakens models' idiomatic competence already early in the tuning phase, while their performance on general linguistic acceptability probes remains comparatively stable. This suggests that idiomatic language is a particularly vulnerable ability. We therefore conclude that further work is needed to develop methods that preserve nuanced language skills in multilingual models lacking native instruction data.

Limitations

Datasets. The conclusions of this work are based on relatively small, manually curated datasets. As visible in the plots in Figure 1, this leads to some variability in evaluation results. However, we observe that the *trends* remain consistent across models, checkpoints and settings.

Unfortunately, one of our datasets cannot be publicly released, as it is derived from copyrighted material. The source book is currently easily available for purchase, making it possible to reproduce the results, but given its niche nature, its availability may change over time.

Coverage. Our work focuses on Swedish, a mid-resource language closely related to English. Our findings that the acquisition of idiomatic language needs time and data is however particularly consequential for lower-resourced languages, where pre-training data is sparser and translations of instruction data will be worse. Those languages should therefore be the subject of future studies on idiomatic language acquisition.

Finally, we restrict our analysis to small models due to the limited availability of large, open models with Swedish pretraining or continued pretraining and with training checkpoints available for download. Due to resource constraints on our side, we were "only" able to train 135M parameter models ourselves. Future work on larger models may yield insights into the acquisition of idiomatic language at different scales.

Acknowledgments

I thank the action editor and the anonymous reviewers for their careful reading, constructive feedback, and insightful suggestions, which substantially improved this paper. I am grateful to student assistants Anja Jarochenko and Salome Kasendu for their help with annotating and revising the datasets in this study. Finally, I thank all colleagues who provided valuable feedback on edge-case idioms, in particular Karin Baardsen, Olle Torstensson, Oskar Holmström, Marco Kuhlmann, and Lukas Borggren. This research was supported by TrustLLM funded by Horizon Europe GA 101135671. The computations were enabled by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model.

Jens Allwood. 1998. Some Frequency based Differences between Spoken and Written Swedish. In *Papers from the 16th Scandinavian Conference of Linguistics. Turku University, Dept of Finnish and General Linguistics*, pages 18–29.

Anthropic. 2024. Introducing the next generation of Claude.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Mona Baker. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In *Text and Technology*. John Benjamins.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic Evaluation and Analysis of Idioms in Neural Machine Translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative Language in Recognizing Textual Entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2024. Characterizing Learning Curves During Language Model Pre-Training: Learning, Forgetting, and Stability. *Transactions of the Association for Computational Linguistics*, 12:1346–1362.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or Multilingual Instruction Tuning: Which Makes a Better Alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The Grammar-Learning Trajectories of Neural Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Marta Ruiz Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Ken-591 neth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam,

Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, C. Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alex Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841 – 846.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the Tip of the Iceberg: A Data Set for Idiom Translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation studies in Scandinavia: Poceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, number 75 in Lund Studies in English, page 88–95. CWK Gleerup, Lund.

Aaron Gluck, Katharina von der Wense, and Maria Leonor Pacheco. 2025. CLIX: Cross-Lingual Explanations of Idiomatic Expressions.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Oskar Holmström and Ehsan Doostmohammadi. 2023. Making Instruction Finetuning Accessible to Non-English Languages: A Case Study on Swedish Models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 634–642, Tórshavn, Faroe Islands. University of Tartu Library.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Dieuwke Hupkes and Willem Zuidema. 2018. Visualisation and 'Diagnostic Classifiers' Reveal how Recurrent and Recursive Neural Networks Process Hierarchical Structure (Extended Abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5617–5621. International Joint Conferences on Artificial Intelligence Organization.

Xisen Jin and Xiang Ren. 2025. Demystifying Language Model Forgetting with Low-rank Example Associations.

Alexander Katourgi. 2022. *Svenskan går bananer: En bok om översättningar som syns*. Lys förlag.

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of Potentially Idiomatic Expressions with Contextual Embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94, online. Association for Computational Linguistics.

Hele-Andra Kuulmets and Mark Fishel. 2023. Translated Benchmarks Can Be Misleading: the Case of Estonian Question Answering. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 710–716, Tórshavn, Faroe Islands. University of Tartu Library.

Batia Laufer. 2000. Avoidance of idioms in a second language: The effect of L1-L2 degree of similarity. *Studia linguistica*, 54(2):186–196.

Vladimir I Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in Literalism: How Supervised Training Shapes Translationese in LLMs.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the Threshold: Idiomatic Machine Translation through Retrieval Augmentation and Loss Weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

Li Liu and Francois Lareau. 2024. Assessing BERT's sensitivity to idiomaticity. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 14–23, Torino, Italia. ELRA and ICCL.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan

Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology.

Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Vladislav Mikhailov, Tita Enstad, David Samuel, Hans Christian Farsethås, Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2025. NorEval: A Norwegian Language Understanding and Generation Evaluation Benchmark. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3495–3541, Vienna, Austria. Association for Computational Linguistics.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Dan Nielsen. 2023. ScandEval: A Benchmark for Scandinavian Natural Language Processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Bolette S. Pedersen, Nathalie Sørensen, Sanni Nimb, Dorte Haltrup Hansen, Sussi Olsen, and Ali Al-Laith. 2025. Evaluating LLM-Generated Explanations of Metaphors – A Culture-Sensitive Study of Danish. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 470–479, Tallinn, Estonia. University of Tartu Library.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language.

Laura Pinnavaia et al. 2002. The grammaticalization of English idioms: a hypothesis for teaching purposes. *Mots Palabras Words*, 1(2002):53–60.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs Produce Less Literal Translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Gisela Redeker. 1984. On differences between spoken and written language. *Discourse Processes*, 7(1):43–55.

David Samuel, Vladislav Mikhailov, Erik Velldal, Lilja Øvrelid, Lucas Georges Gabriel Charpentier, Andrey Kutuzov, and Stephan Oepen. 2025. Small Languages, Big Models: A Study of Continual Training on Languages of Norway.

In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 573–608, Tallinn, Estonia. University of Tartu Library.

Naomi Saphra and Adam Lopez. 2019. Understanding Learning Dynamics Of Language Models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

Ekaterina V Shutova. 2011. Computational approaches to figurative language. Technical report, University of Cambridge, Computer Laboratory.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI Models' Performance on Figurative Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-On What Language Model Pre-training Captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Irina Tarabrina. 2019. *Addressing idiom avoidance in culturally and linguistically diverse students through biography-driven instruction*. Kansas State University.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Teknium. 2023. OpenHermes 2.5: An Open Dataset of Synthetic Data for Generalist LLM Assistants.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. LLM Circuit Analyses Are Consistent Across Training and Scale. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Volodina, Yousuf Ali Mohammed, Aleksandrs Berdicevskis, Gerlof Bouma, and Joey Öhman. 2023. DaLAJ-GED - a dataset for grammatical error detection tasks on Swedish. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 94–101, Tórshavn, Faroe Islands. LiU Electronic Press.

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. 2025. PolyPythias: Stability and Outliers across Fifty Language Model Pre-Training Runs. In *The Thirteenth International Conference on Learning Representations*.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes Σ A Benchmark to Assess the Linguistic Competence of Language Models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Charles H Woolbert. 1922. Speaking and writing — a study of differences. *Quarterly Journal of Speech*, 8(3):271–285.

Arnav Yayavaram, Siddharth Yayavaram, Prajna Devi Upadhyay, and Apurba Das. 2024. BERT-based Idiom Identification using Language Translation and Word Cohesion. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies*

(MWE-UD) @ LREC-COLING 2024, pages 220–230, Torino, Italia. ELRA and ICCL.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. The Nordic Pile: A 1.2TB Nordic Dataset for Language Modeling.

A Details about the Construction of the *Idioms* Dataset

This appendix provides a detailed description of how the idiomatic minimal pairs were constructed, complementing the briefer description in Section 3.2.1. Its goal is to clarify how idiomatic keywords were identified (§A.1), how plausible but non-idiomatic alternatives were selected (§A.2), and how these elements were embedded into full minimal-pair sentences (§A.3).

A.1 Keyword Identification

For each idiom, our annotators first identified a single idiomatic keyword that is essential to the conventionalized expression and that distinguishes it from plausible but non-idiomatic variants. This keyword is a content word that is required for the idiom.[9] Idioms that did not have a suitable content word were discarded from the dataset. Annotators then formulated a prompt sentence that explicitly conveys the idiom's meaning or definition and is designed to elicit this keyword. These prompts are not intended to be an everyday usage of the idiom but rather to provide a clear semantic cue for the idiomatic expression.

Consider this example from Section 3.2.1, for the idiom *rik som ett troll*:[10]

> *Någon som är väldigt rik är rik som ett troll.*
> ("Someone who is very rich is rich like a troll.")

Here, the idiomatic keyword is *troll*, which is the central word of the conventionalized idiom *rik som ett troll* ("rich like a troll"; "very rich").

A.2 Construction of Plausible Alternatives

For each idiomatic keyword, we manually selected a distractor keyword that is semantically plausible in the same context but does not yield the idiomatic expression. Whenever possible, distractors were inspired by idioms in closely related high-resource languages (English or German). When no such counterpart existed, we selected a semantically compatible alternative from the same lexical category.

For the example above, the idiomatic keyword *troll* was replaced by the distractor keyword *kungabarn* ("prince" or "princess"), which is semantically equally plausible in context as both words convey the meaning "very rich", but only *rik som ett troll* forms a conventionalized idiom.

Another example (also from Section 3.2.1 is:[11]

> *Om man jämför saker som inte är jämförbara, då jämför man äpplen och . . .*
> ("If one compares things that are not comparable, then one compares apples and . . . ")
> Keyword: *Päron* ("Pears")
> Distractor: *Apelsiner* ("Oranges")

Here, the distractor *apelsiner* ("oranges") is chosen because there is a similar English idiom that goes "comparing apples and oranges".

A.3 Rewriting into Minimal Pair Sentences

To construct minimal pairs, we used the chatbot assistant Claude (Anthropic, 2024) to rewrite the original prompt into two complete sentences that are identical except for the idiomatic keyword versus the distractor. The assistant was instructed to ensure that both variants were well-formed Swedish sentences and equal except for the keyword. After this rewriting step, the idiomatic keyword (and its distractor) may appear in any position in the sentence, depending on what yields the most natural construction. This step results in minimal pairs suitable for full-sentence perplexity comparison. To give an example, the annotators, for the idiom *ana ugglor i mossen* ("to sense owls in the bog", "to sense that something is wrong"),[12] wrote the following initial question and keyword:

> *Vilka djur anar man i mossen när man misstänker att något är fel?*
> ("Which animals does one sense in the bog when one suspects that something

[9] The vast majority of the keywords are nouns, but even verbs, adjectives and adverbs occur.

[10] https://sv.wiktionary.org/wiki/rik_som_ett_troll

[11] https://sv.wiktionary.org/wiki/Ãd'pplen_och_pÃd'ron)

[12] https://sv.wiktionary.org/wiki/ana_ugglor_i_mossen

is wrong?")
Keyword: *Ugglor* ("Owls")
Distractor: *Råttor* ("Rats")

Then, the Claude model rewrote the example into the following minimal pair:

> *När man misstänker att något är fel, anar man ugglor i mossen.*
> ("When one suspects that something is wrong, one senses owls in the bog."")
>
> *När man misstänker att något är fel, anar man råttor i mossen.*
> ("When one suspects that something is wrong, one senses rats in the bog.")

All generated sentence pairs were manually reviewed and, if necessary, edited by the annotators to ensure grammatical correctness, semantic equivalence between the two variants, and a clear preference for the idiomatic variant according to native-speaker judgments.

Prompt We used the following prompt to generate the minimal pairs:

> *Construct two well-formed and coherent Swedish sentence alternatives from the following sentence by filling in the two completion words. The resulting sentences should be identical except for the completion word.*
>
> - *Sentence:* [original sentence]
> - *Completion 1:* [keyword]
> - *Completion 2:* [distractor]

B Ablations: Filtering of the *Translationese* Subset

The filtered Translationese dataset was constructed using two criteria: (i) sentence pairs must have the same length, and (ii) human annotators must agree that the Translationese sample is less idiomatic than its corrected counterpart. In this section, we ablate the two criteria to assess their individual impact.

While human annotators excluded only 263 instances as not clearly Translationese, enforcing the strict length criterion eliminated 829 samples for SmolLM and 782 for Llama. For the best-performing AI Sweden Llama model, human filtering increased accuracy by 5.49 points, whereas the length criterion increased accuracy by 24.06 points (using the model tokenizer, as per our standard setup).

B.1 Effects of the Human Filtering

As noted in the preceding paragraph, the human filtering leads to slightly higher scores, but does not do the heavy lifting in the subset. But although the length-matching constraint has a substantially larger practical influence, we hold on to the filtering because it impacts the quality of the dataset if some samples are not clearly identifiable as Translationese without more context.

B.2 Effects of the Tokenizer

As noted in Section 3.2.2, a limitation of basing the length restriction on the model tokenizer is that results cannot be directly compared across models with different tokenizers. While our focus in this paper is on intra-model comparisons (across model training checkpoints and before vs. after instruction tuning), this choice constrains the applicability of the dataset in comparisons of models with different tokenizers. In addition, the strict criterion substantially reduces the number of usable samples. To address this, we experiment with an alternative approach in which the length restriction is applied to whitespace-tokenized text instead.

Table 3 shows that relaxing the length criterion in this way yields a substantially larger dataset, with 285 samples retained compared to only 98 or 134 when using model-specific tokenizers. However, the bias towards the longer Translationese examples also remains stronger with the whitespace tokenizer: For example, for the AI Sweden Llama model, accuracy on the final subset is 82.08 when using the model tokenizer, but 61.40 when using the whitespace-based tokenizer.

Finally, we would like to emphasize that both tokenizers for the models included in this study are English-centric. Tokenizers that are better adapted to Swedish, and thereby have a lower fertility for our dataset, will likely result in larger subsets.

Subset	Size of Subset	$\text{SmolLM}_{\text{scratch}}$	$\text{SmolLM}_{\text{CPT}}$	AI Sweden Llama
All samples	967	46.94	47.36	48.91
All filters, model tokenizer	98 (SmolLM) / 134 (Llama)	58.20	69.40	82.08
Manual only	704	49.57	50.85	54.40
Length only, model tokenizer	138 (SmolLM) / 185 (Llama)	65.21	70.28	72.97
All filters, whitespace tokenizer	285	56.84	62.10	61.40
Length only, whitespace tokenizer	395	53.67	55.18	55.44

Table 3: **Ablations on the Translationese subset filtering**. *Manual only* includes examples where the idiomatic variant is clearly preferable according to general Swedish language norms, independent of additional context (see Section 5.1.2 for details). *Length only* includes examples where the Translationese sample and the idiomatic alternative have the same length, determined either by whitespace tokenization or by the model tokenizer, as specified in the table. *All filters* applies both criteria simultaneously.