Can Language Models Transfer Syntactic Parameters Across Languages? A Case Study of *Pro*-drop

Anonymous ACL submission

Abstract

Language models typically learn from unannotated corpora, yet their ability to acquire abstract syntactic parameters like *Pro*-drop through cross-lingual transfer remains an open question. Analogous to human second-language acquisition, we examine whether models can leverage annotated data from a source language to induce syntax in a target. We evaluate two multilingual encoder models using an annotated parallel corpus of *The Little Prince* across English, Spanish, Korean, and Chinese, comparing zero-shot baselines with in-language fine-tuning and cross-lingual transfer. While supervised fine-tuning consistently improves performance, cross-lingual transfer yields inconsistent results across pairs. Notably, transfer between Spanish and Chinese results in adverse effects, suggesting difficulty reconciling morphologically-licensed *Pro*-drop with topic-drop. Our findings suggest that language models may learn language-specific licensing strategies rather than a universal syntactic parameter, as cross-lingual exposure does not always facilitate positive transfer.

1 Introduction

Consider the following excerpt from Chapter 21 of *The Little Prince* across four languages, with dropped subjects indicated where applicable:

(1) EN: What must I do to tame you?

(2) ES: ¿Qué Ø tengo que hacer?
What have.1SG to do.INF

(3) KO: Ø ne-lul kiltuli-lye-myen etteh-key
you-ACC tame-PURP-COND how
ha-myen toy-ci?
do-COND become-Q

(4) ZH: Ø yīngdāng zuò xiē shénme ne
should do some what Q

The realization of subjects in these examples illustrate cross-linguistic variation in the *Pro*-drop parameter (Chomsky, 1981). Whereas English requires overt subjects in finite clauses, Spanish, Korean, and Chinese permit phonologically null subjects under distinct licensing conditions: Spanish through rich verbal agreement morphology, and Korean and Chinese through discourse linking and topic prominence (Li and Thompson, 1976; Huang, 1984; Song, 2014). A central puzzle is how human learners acquire these language-specific, abstract syntactic constraints from limited input. While syntactic knowledge has traditionally been associated with innate linguistic knowledge (Chomsky, 1957), recent studies demonstrate language models' ability to acquire grammar through related constructions or languages (Misra and Mahowald, 2024; Varda and Marelli, 2023).

Despite the typological prevalence of *Pro*-drop languages (Dryer, 2013), multilingual models are trained on corpora heavily skewed toward English

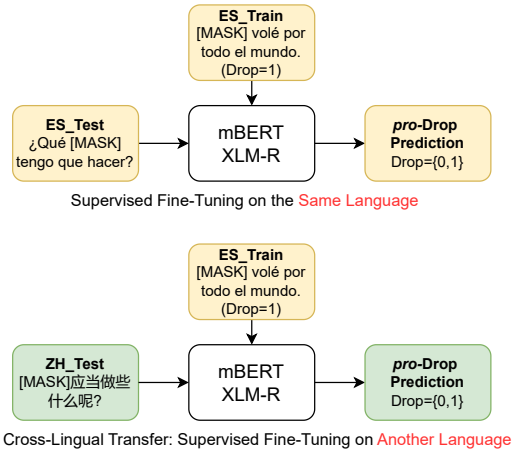

Figure 1: Overview of our supervised experimental framework for *Pro*-drop detection. We fine-tune two multilingual encoder models (mBERT and XLM-R) on masked parallel sentences to predict binary *Pro*-drop labels, comparing in-language performance (top) and cross-lingual transferability (bottom) against baseline models.

(Wendler et al., 2024), a strictly non-*Pro*-drop language. This distributional imbalance raises questions about whether models possess knowledge about *Pro*-drop constraints, and whether such constraints can be learned through supervised learning. Evidence of positive and negative transfer in human bilinguals' learning of *Pro*-drop (Hacohen and Schaeffer, 2007; LaFond, 2001) motivates investigations into comparable cross-lingual transfer effects in language models. Understanding model behavior on *Pro*-drop has both theoretical and practical significance: theoretically, it tests whether neural architectures can capture syntactic categories lacking surface realization; practically, failing to recover *pro* may lead to incomplete or incorrect translations in downstream applications.

To this end, we evaluate the ability of two multilingual encoder models—mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020)—to learn cross-linguistic *pro* constraints in English, Spanish, Korean, and Chinese, using a binary labeling task derived from a parallel, annotated corpus of *The Little Prince*. First, we test base models in a zero-shot setting to assess their preferences for dropped versus overt subjects across languages. Second, we fine-tune models on a single language and evaluate in-language performance. Third, we test whether cross-lingual transfer is at play by training on one language and evaluating on the three others, examining whether *Pro*-drop behavior learned in one language has positive or negative implications in another.

2 Related Work

2.1 *Pro*-drop in Language Acquisition

Prior work on human language acquisition has modeled *Pro*-drop as a problem of selecting among competing grammatical hypotheses. Yang (2000) proposes a variational model in which children simultaneously entertain multiple grammars compatible with Universal Grammar that probabilistically compete to match the incoming linguistic evidence. In second language acquisition, LaFond (2001) finds that English speakers learning Spanish pass through intermediate stages where they overgeneralize null subjects, after which learners eventually re-rank syntactic constraints requiring overt subjects below discourse constraints licensing null subjects. More broadly, research on cross-linguistic influence in multilingual learners documents that transfer effects from previously acquired languages can be facilitative or inhibitory, depending on the structures of the source and target languages (Odlin, 1989).

2.2 *Pro*-drop in downstream NLP tasks

Computational work on *Pro*-drop has primarily addressed challenges in two task domains: null pronoun resolution and machine translation. In resolution tasks, Yin et al. (2018) introduced self-attention mechanisms with multiple hops to encode *pro* by focusing on the most informative portions of the given context, achieving then state-of-the-art performance on the Chinese OntoNotes dataset. Subsequently, Song et al. (2020) proposed a BERT-based architecture to find the locations of *pro* and recover their referent, in the absence of gold-standard parsing data. In neural machine translation, Hwang et al. (2021) explored data augmentation and contrastive learning, yielding improvements in English-German and English-Korean translation. Tan et al. (2021) employed a two-stage process unifying context modeling and *pro* recovery, which led to improvements in machine translation, question answering, and summarization tasks. While these works propose methods to represent context in detecting and resolving *pro*, the availability of annotated data and the limited coverage of typologically different *Pro*-drop languages remain as challenges.

3 Methods

3.1 Dataset

Our data is adapted from the annotated parallel corpus of *The Little Prince* developed by Song (2014), which provides sentence-aligned translations across English (EN), Spanish (ES), Korean (KO), and Chinese (ZH). This text was selected for several methodological advantages: its manageable length for human annotation (1,755 sentence sets), availability of high-quality translations, and the short and colloquial nature of its sentences, which correlates with higher frequency of *Pro*-drop (Han, 2004). The corpus annotates subjects with detailed linguistic features including categorical type, determiner properties, language-specific morphological markers, and animacy. Importantly, the annotation distinguishes subject realization types: overt subjects, dropped subjects (marked with #), and subjects unrealized in one language but realized in another. Inter-annotator agreement measured by Cohen's κ exceeded 0.8 across all language pairs.

Lang	Overt	Dropped	Total
EN	2,258	50	2,308
ES	844	1,280	2,124
KO	1,633	512	2,145
ZH	1,887	231	2,118
Total	**6,622**	**2,073**	**8,695**

Table 1: Dataset composition by language. English dropped subjects are limited to imperatives. *Pro*-drop is most frequent in Spanish (60.3%), followed by Korean (23.9%) and Chinese (10.9%).

We transformed these annotations into a binary classification dataset. For each sentence containing n subjects, we generate n separate instances. Within each instance, the target subject position is replaced with a placeholder token [M], while all other overt subjects are preserved. For dropped subjects, the mask token is inserted at the syntactically appropriate position. Each instance receives a binary label: 1 for dropped subjects, 0 for overt subjects. For Korean, the masked span includes case-marking particles following the subject to prevent morphological cue leakage. Table 1 shows the corpus breakdown of 8,695 subject instances across four languages.

We partitioned each language independently into train/validation/test splits (70:10:20) with stratified sampling based on the binary label and grammatical property. See Appendix A for details.

3.2 Models

We employ two multilingual encoder architectures: XLM-RoBERTa Base (XLM-R; Conneau et al. 2020) and multilingual BERT (mBERT; Devlin et al. 2019). mBERT was pretrained on Wikipedia across 104 languages; XLM-R was trained on 2.5TB of CommonCrawl data covering 100 languages. Both models provide robust multilingual competence and have demonstrated strong cross-lingual transfer capabilities (Janeiro et al., 2025).

For each language, we fine-tuned both base models on the training set, yielding 8 language-specific classifiers (4 languages $\times$ 2 architectures). Fine-tuning added a linear classification head on the [CLS] token. Input sentences were tokenized with [M] replaced by the model-specific mask token. Models were evaluated using macro-averaged F1 as the primary metric. See Appendix B for implementation details.

Exp	Model	Train	Eval
1	Base (MLM)	None	All 4 langs
2	Fine-tuned	Same lang	Same lang
3	Fine-tuned	Source lang	Target lang

Table 2: Experimental design. Exp 1: zero-shot (8 conditions), Exp 2: in-language (8 conditions), Exp 3: cross-lingual transfer (24 conditions).

3.3 Experiments

We conducted three experiments to assess models' understanding of *Pro*-drop and cross-lingual transferability (Table 2). Experiment 1 probes whether pretrained models exhibit biases toward overt versus dropped subjects without fine-tuning by extracting the top-1 predicted token at the mask position; if the token is punctuation (. , ; ! ?), we label it as dropped (1), otherwise as overt (0). Experiment 2 evaluates each fine-tuned model on its corresponding test set to establish supervised learning baselines. Experiment 3 tests all cross-lingual pairs by evaluating each fine-tuned model on all four test sets, yielding 12 pairs per architecture (24 total conditions), examining whether training on Spanish (morphological *Pro*-drop) improves performance on Korean/Chinese (discourse-based drop), whether English training harms *Pro*-drop language performance, and whether transfer effects are symmetric.

4 Results

Table 3 presents complete results across all experiments. We report macro-averaged F1, accuracy, precision, and recall.

4.1 Baseline and In-Language Performance

Base models achieved high accuracy on English (mBERT: .970, XLM-R: .978) but low F1 scores (.642, .495), reflecting trivial majority-class prediction on a dataset that is 97.8% overt subjects. On *Pro*-drop languages, both models failed systematically with F1 scores below .48, indicating they lack coherent representations of *Pro*-drop licensing without supervision.

Supervised fine-tuning dramatically improved performance. English models achieved near-perfect results (mBERT F1: .943, XLM-R F1: .860), successfully learning to identify rare imperative contexts. Spanish fine-tuning yielded strong performance (F1 > .84 for both models), demonstrating successful acquisition of morphological licensing cues. Korean showed moderate perfor-

			mBERT				XLM-RoBERTa			
Exp	**Train**	**Test**	**F1**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**
1	–	EN	**.642**	.970	.642	.642	**.495**	.978	.489	.500
	–	ES	**.428**	.449	.516	.511	**.285**	.398	.199	.500
	–	KO	**.478**	.653	.478	.482	**.433**	.762	.381	.500
	–	ZH	**.471**	.889	.446	.499	**.471**	.892	.446	.500
2	EN	EN	**.943**	.996	.998	.900	**.860**	.987	.831	.896
	ES	ES	**.841**	.852	.854	.834	**.842**	.847	.839	.845
	KO	KO	**.647**	.741	.645	.648	**.670**	.781	.692	.657
	ZH	ZH	**.572**	.854	.585	.565	**.471**	.892	.446	.500
3	ES	EN	**.403**	.567	.524	.779	**.489**	.729	.537	.862
	KO	EN	**.396**	.561	.519	.727	**.592**	.913	.567	.760
	ZH	EN	**.491**	.965	.489	.493	**.495**	.978	.489	.500
	EN	ES	**.623**	.633	.760	.695	**.362**	.440	.682	.534
	KO	ES	**.619**	.621	.680	.664	**.369**	.440	.630	.531
	ZH	ES	**.302**	.407	.701	.508	**.285**	.398	.199	.500
	EN	KO	**.443**	.765	.882	.505	**.506**	.774	.751	.535
	ES	KO	**.431**	.431	.588	.590	**.525**	.538	.590	.620
	ZH	KO	**.445**	.748	.481	.498	**.433**	.762	.381	.500
	EN	ZH	**.493**	.894	.947	.511	**.597**	.889	.689	.575
	ES	ZH	**.429**	.483	.549	.624	**.460**	.521	.564	.665
	KO	ZH	**.472**	.580	.533	.583	**.557**	.764	.553	.591

Table 3: Complete experimental results: base models, in-language fine-tuning and cross-lingual transfer.

mance (F1: .647-.670), likely reflecting the complexity of discourse-based licensing. Chinese presented a critical failure for XLM-R (F1: .471), where severe class imbalance (89% overt subjects) caused early stopping after one epoch, resulting in a majority-class baseline. mBERT performed better (F1: .572) but still struggled relative to Spanish.

4.2 Cross-Lingual Transfer

Cross-lingual results reveal complex patterns depending on typological alignment. Transfer to English showed asymmetric effects: Korean to English was surprisingly effective for XLM-R (F1: .592), substantially outperforming the zero-shot baseline, while Spanish to English yielded near-chance performance for both models (F1: .403-.489). This suggests discourse-based licensing provides more generalizable knowledge than morphological cues for identifying English imperatives. English training proved beneficial for Spanish (mBERT: .623), suggesting it helps models learn a general distinction between obligatory versus optional subject positions. However, transfer to Korean was weaker (F1: .443-.506).

The most striking finding is mutual negative transfer between Spanish and Chinese. Spanish to Chinese transfer failed to improve over baseline (mBERT: .429 vs. .471 baseline), while Chinese to Spanish actually degraded performance (mBERT: .302 vs. .428 baseline; XLM-R: .285 vs. .285 baseline). This mutual interference provides strong evidence that models learn language-specific licensing strategies rather than a unified *Pro*-drop parameter. Spanish training teaches reliance on verbal inflection, which is absent in Chinese; Chinese training emphasizes topic continuity, which does not predict Spanish null subjects.

Korean showed intermediate transfer patterns (F1: .431-.525), potentially reflecting its hybrid system with both morphological markers and discourse-based licensing.

5 Conclusion

This study demonstrates that multilingual encoders do not inherently represent *Pro*-drop as a unified cross-lingual feature. While supervised fine-tuning successfully teaches language-specific *Pro*-drop distributions, cross-lingual transfer is highly sensitive to typological alignment. The adverse transfer between Spanish and Chinese *Pro*-drop indicates models learn surface-level licensing strategies rather than abstract syntactic principles. These findings suggest that transformer architectures represent *Pro*-drop as collections of language-specific patterns rather than as a parameterized universal. Future work should investigate whether explicit typological encoding or multi-task learning across diverse *Pro*-drop languages can better capture these varying licensing mechanisms.

Limitations

This study has several limitations. First, the parallel corpus is derived from a single literary text, which may exhibit distinct *Pro*-drop distributions compared to spoken or formal written registers. Second, we focus exclusively on subject-drop; object-drop and topic-drop in their full complexity remain subjects for future research.

A critical limitation concerns XLM-R's performance on Chinese in-language fine-tuning. Due to extreme class imbalance in the Chinese training data (89% overt subjects), XLM-R's training triggered early stopping after one epoch, yielding a degenerate majority-class baseline (F1: .471, accuracy: .892). The model failed to learn meaningful *Pro*-drop patterns and instead predicted "overt" for all instances. Consequently, XLM-R's cross-lingual transfer results involving Chinese as source or target should be interpreted with caution. This highlights a broader challenge: multilingual models may struggle to learn minority-class patterns under severe class imbalance, even with stratified sampling and standard regularization. Future work should explore class-balanced sampling, focal loss, or other techniques to address this issue.

References

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Noam Chomsky. 1981. *Lectures on Government and Binding: The Pisa Lectures*, volume 9 of *Studies in Generative Grammar*. Foris Publications, Dordrecht.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Matthew S Dryer. 2013. Expression of pronominal subjects. *The world atlas of language structures online*.

Aviya Hacohen and Jeannette Schaeffer. 2007. Subject realization in early hebrew/english bilingual acquisition: The role of crosslinguistic influence. *Bilingualism: Language and Cognition*, 10(3):333–344.

Na-Rae Han. 2004. Korean null pronouns: Classification and annotation. In *Proceedings of the Workshop on Discourse Annotation*, pages 33–40, Barcelona, Spain. Association for Computational Linguistics.

C-T James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry*, pages 531–574.

Yongkeun Hwang, Hyeongu Yun, and Kyomin Jung. 2021. Contrastive learning for context-aware neural machine translation using coreference information. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1135–1144, Online. Association for Computational Linguistics.

João Maria Janeiro, Belen Alastruey, Francisco Massa, Maha Elbayad, Benjamin Piwowarski, Patrick Gallinari, and Loic Barrault. 2025. Mixture of languages: Improved multilingual encoders through language grouping. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29707–29722, Suzhou, China. Association for Computational Linguistics.

Larry L. LaFond. 2001. *The Pro-drop Parameter in Second Language Acquisition Revisited: A Developmental Account*. Doctoral dissertation, University of South Carolina.

Charles N. Li and Sandra A. Thompson. 1976. Subject and topic: A new typology of language. In Charles N. Li, editor, *Subject and Topic*, pages 457–489. Academic Press, New York.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press, Cambridge.

Linfeng Song, Kun Xu, Yue Zhang, Jianshu Chen, and Dong Yu. 2020. ZPR2: Joint zero pronoun recovery and resolution using multi-task learning and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5429–5434, Online. Association for Computational Linguistics.

Sanghoun Song. 2014. Subject-drop vs. topic-drop: Evidence from a multilingual text. *Language and Linguistics*, 64:149–182.

Xin Tan, Longyin Zhang, and Guodong Zhou. 2021. Coupling context modeling with zero pronoun recovering for document-level natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrea Gregor de Varda and Marco Marelli. 2023. Data-driven cross-lingual syntax: An agreement study with massively multilingual models. *Computational Linguistics*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Charles D Yang. 2000. *Knowledge and learning in natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018. Zero pronoun resolution with attention-based neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 13–23, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Data Partitioning Details

We partitioned each language independently into train/validation/test splits using a 70:10:20 ratio with stratified sampling. The stratification key combines the binary label with the grammatical property tag from Song (2014)'s annotation (e.g., pronoun, noun phrase), ensuring balanced representation of both subject realization and grammatical category across splits. For rare categories with fewer than 5 instances (insufficient for stratification), we applied random splitting and allocated instances primarily to the training set to maximize training data for low-frequency patterns.

B Training Implementation

Fine-tuning used the following hyperparameters: learning rate 2×10^{-5} with linear warmup, batch size 16 (training) and 32 (evaluation), 5 epochs with early stopping, weight decay 0.01, AdamW optimizer, maximum sequence length 128 tokens with dynamic padding. Models were evaluated on the validation set after each epoch using macro-averaged F1 score. The best-performing checkpoint based on validation F1 was retained for testing. All models were trained on a single NVIDIA T4 GPU with mixed-precision training enabled.