The Klingon Effect: When Constructed Languages Win at LID

Abstract

Language identification (LID) underpins multilingual NLP, yet little is known about how constructed languages behave in this setting. We present a large-scale study of constructed-language identification, compiling 14.2M labeled sentences across 101 languages (90 natural; 11 constructed) from Wikipedia, Tatoeba, and UDHR. Using a three-tier experimental design with propensity-score matching to control key confounders (e.g., script, orthographic entropy), we uncover the "Klingon Effect": a consistent advantage for constructed languages that increases with identification difficulty. When contrasted against the most challenging natural languages, constructed languages achieve up to a 4.61% absolute F1 advantage ($p < 0.0001$), exceeding the maximum gap observed between any two natural languages. The effect is not explained by script alone and appears linked to design regularity and standardization. These findings call for further exploration of constructed languages in LID and for LID systems that explicitly account for distributional regularity, particularly for low resource languages.

1 Introduction

Language identification (LID) systems are fundamental components of multilingual NLP pipelines, enabling downstream tasks to adapt to linguistic characteristics of input text (Joulin et al., 2016; Baldwin et al., 2006). While LID research has primarily focused on natural languages (Malmasi et al., 2017; Jauhiainen et al., 2019), the emergence of constructed languages in digital spaces (from artistic languages like Klingon and Dothraki to auxiliary languages like Esperanto) presents both challenges and opportunities for understanding language identification performance.[1]

Constructed languages, or conlangs, are linguistically designed systems that range from naturalistic auxiliary languages to highly distinctive artistic languages (Schwitter et al., 2003). These languages exhibit characteristics that may differ systematically from natural languages, including standardized orthography, reduced dialectal variation, and intentional design features that could potentially affect language identification performance.

Previous work on constructed language identification has been limited by small datasets and lack of systematic comparison with natural languages. We present a comprehensive dataset of 14.2 million samples across 101 languages (90 natural + 11 constructed) from diverse sources including Wikipedia (Wikimedia Foundation, 2023), Tatoeba (Tatoeba Project, 2025), and the Universal Declaration of Human Rights (CIS LMU, 2024; Unicode Consortium, 2016).

Our experimental framework employs three complementary approaches: controlled experiments isolating script and family effects among natural languages, direct comparison between natural and constructed languages using propensity score matching, and targeted analysis of constructed languages against the most challenging natural languages to test whether the advantage scales with identification difficulty.

Our findings reveal a nuanced picture: while constructed languages show no inherent advantage over natural languages in general, they exhibit a context-dependent advantage that scales with identification difficulty, from 2.59% in balanced comparisons to 4.61% against the most challenging natural languages. This "Klingon Effect" is significant not for its magnitude but for its novelty; constructed languages theoretically should not have any advantage over natural languages in identification tasks.

The finding is robust across experimental designs and meaningful in the 95-99% performance

[1] *Terminology.* We use "Klingon Effect" as a metaphor for the broader pattern observed across constructed languages; it does not imply that Klingon alone drives the effect.

range where small improvements matter, exceeding the maximum difference between any two natural languages. This suggests that constructed languages have genuinely distinctive characteristics that become more apparent as natural language identification becomes more challenging.

Our contributions include: (1) a comprehensive dataset with 14.2M samples across 101 languages; (2) a systematic analysis of performance differences between natural and constructed languages using rigorous statistical controls; (3) identification of the "Klingon Effect", a context-dependent advantage that scales with identification difficulty; and (4) implications for LID system design and evaluation methodology.

2 Research Questions and Methodology

Our research addresses three fundamental questions about constructed language identification, each building on the previous to provide a comprehensive understanding of performance patterns:

Research Question 1: What are the baseline performance patterns among natural languages?

Before comparing natural and constructed languages, we must establish baseline performance patterns among natural languages to understand what factors drive identification success. We investigate script effects, family effects, and orthographic consistency to identify the primary determinants of natural language identification performance. This provides the foundation for understanding whether constructed languages follow similar patterns or exhibit distinctive characteristics.

Research Question 2: Do constructed languages perform differently from natural languages when controlling for confounding variables?

The core question of our work addresses whether constructed languages are inherently easier or harder to identify than natural languages. We employ propensity score matching to control for sample size, script, family, and orthographic consistency, ensuring that any performance differences are attributable to language type rather than other factors. This question tests the hypothesis that design characteristics specific to constructed languages could confer identification advantages or disadvantages. Our analysis will reveal whether constructed languages exhibit distinctive identification patterns compared to their natural counterparts.

Research Question 3: Do constructed languages outperform challenging natural languages?

Even if constructed languages show no general advantage over natural languages, they may excel specifically against natural languages that struggle with identification. This question investigates whether constructed languages exhibit advantages over the most challenging natural languages, potentially revealing distinctive characteristics that aid identification in difficult cases.

These research questions form a logical progression from establishing baseline patterns to investigating specific advantages, allowing us to build a comprehensive understanding of constructed language identification performance. Each question employs rigorous experimental design and statistical controls to ensure reliable conclusions. We employ a three-tiered experimental approach to investigate constructed language identification performance. Our dataset comprises 14.2 million samples across 101 languages (90 natural + 11 constructed) from Wikipedia, Tatoeba, and UDHR sources. We use FastText for language identification with F1 score as our primary metric, training on balanced samples to avoid high-resource language bias.

Propensity-score covariates and encoding For controlled comparisons, we estimate propensity scores using the following covariates: (i) script (binary indicator for Latin vs non-Latin); (ii) language family (one-hot categorical); (iii) training sample size per language (log-transformed continuous); and (iv) orthographic consistency measured as character-level entropy (continuous). Matching is performed with caliper on the logit of the propensity score and without replacement.

2.1 Natural Language Baseline Analysis

We establish baseline performance patterns among natural languages through controlled experiments isolating script effects, family effects, and orthographic consistency. This provides essential context for understanding constructed language performance.

We measure script effects with within-family pairs differing only in writing system, family effects with cross-family groups matched on script, and orthographic consistency via normalized character-level entropy. All comparisons use equalized data per language and identical FastText settings with held-out evaluation.

Table 1: Dataset summary: samples by source and by type

Category	Count	Share
Wikipedia	2,821,314	19.8%
Tatoeba	11,345,412	79.5%
UDHR	713	$< 0.1\%$
Other (fantasy corpora)	34,186	0.2%
Constructed languages	1,145,001	8.0%
Natural languages	13,056,624	92.0%

2.2 Natural vs Constructed Language Comparison

We employ propensity score matching to create balanced comparison groups, controlling for sample size differences, script distribution effects, language family effects, and orthographic consistency differences. This approach ensures that any performance differences are attributable to language type rather than confounding variables.

Dataset summary Table 1 reports sample counts by source and by language type. Wikipedia and Tatoeba contribute the majority of examples; constructed languages represent a smaller but substantial portion of the corpus.

2.3 Context-Dependent Advantage Analysis

We investigate the "Klingon Effect" through two complementary comparisons: (1) A balanced comparison using 11 constructed languages vs 11 lowest-performing natural languages, and (2) A targeted comparison using 11 constructed languages vs the 5 most challenging natural languages. This dual-comparison design tests whether the constructed language advantage scales with identification difficulty.

2.4 Statistical Analysis

We employ propensity score matching for controlled comparisons. For multi-group comparisons (e.g., language families), we use ANOVA and report the F metric. For two-group comparisons, we report Cohen's d. For associations (e.g., orthographic consistency vs performance), we report Pearson's r. We also report p-values alongside all tests. All tests use $\alpha = 0.05$ with multiple-comparison control.

3 Data Collection

We construct a dataset of 14.2 million samples across 101 languages from Wikipedia, Tatoeba, and UDHR sources.

Table 2: Constructed languages in our dataset

Category	Languages
Auxiliary	Esperanto, Interlingua, Interlingue, Ido, Volapük, LFN
Artistic	Klingon, Dothraki, Kotava, Lojban, Toki Pona

For constructed languages with limited digital presence, we use specialized collection strategies including web scraping, manual curation, and English content filtering.

3.1 Dataset Characteristics

Our dataset includes 11 constructed languages (Table 2) and 90 natural languages representing 14 major families. All text went through Unicode normalization, HTML stripping, and language-specific cleaning including English content removal and duplicate detection.

4 Experimental Framework and Results

4.1 Natural Language Performance Patterns

We establish baseline performance patterns among natural languages to understand the fundamental factors driving identification accuracy before comparing with constructed languages.

Our analysis reveals three key findings: (1) script effects are minimal; non-Latin scripts show only a 1.5% improvement over Latin scripts; (2) family effects are not statistically significant (F=0.835, p=0.6371); and (3) orthographic consistency has weak correlation with performance (r=0.23).

These patterns demonstrate that neither script type nor linguistic family substantially determines identification success, providing crucial context for interpreting constructed language performance.

4.2 Script Effects

We isolate script effects by comparing languages that share the same family but use different writing systems (e.g., Serbian Cyrillic vs Croatian Latin). Non-Latin scripts show a modest but statistically significant improvement over Latin scripts (1.5%, $p < 0.0001$), though this difference is not practically significant.

Script effects appear to have minimal impact on natural language identification performance.

4.3 Family Effects

We compare languages within the same script family but from different language families. Language family effects are not statistically significant ($F = 0.835$, $p = 0.6371$), indicating that family membership has no meaningful impact on identification performance.

This provides important context for understanding constructed language performance, as it demonstrates that linguistic family characteristics do not substantially affect identification accuracy.

4.4 Orthographic Consistency

We compute orthographic consistency scores using character-level entropy measures. Orthographic consistency shows weak correlation with identification performance ($r = 0.23$, $p = 0.02$), suggesting that while orthographic regularity may contribute to identification success, it is not the primary determinant.

4.5 Synthesis of Natural Language Patterns

Our baseline analysis reveals that both script effects and family effects have minimal impact on natural language identification performance. The difference between Latin and non-Latin scripts is modest (1.5%). Family effects are not statistically significant, and orthographic consistency plays a minor role.

Since all constructed languages in our dataset use Latin-based scripts, script type alone cannot explain any performance differences.

4.6 Natural vs Constructed Language Comparison

We directly compare natural and constructed languages using propensity score matching to control for confounding variables.

Methodology: We employ propensity score matching to create balanced comparison groups, controlling for sample size differences, script distribution effects, language family effects, and orthographic consistency differences. This approach ensures that any performance differences are attributable to language type rather than confounding variables.

Results: When controlling for confounding variables, constructed languages show no significant advantage over natural languages ($p = 0.847$). As shown in Figure 1A, the mean F1 scores are nearly identical between groups: natural languages achieve 96.8% mean F1 score, constructed languages achieve 96.9% mean F1 score, with a difference of only 0.1% (not significant) and an effect size of Cohen's $d = 0.12$ (negligible).

Table 3: Artistic vs Auxiliary constructed language performance

Category	Languages	Mean F1	Count
Artistic	Klingon, Dothraki, Kotava, Lojban, Toki Pona	99.67%	5
Auxiliary	Esperanto, Interlingua, Interlingue, Ido, Volapük, LFN	98.34%	6

Implications: Constructed languages are not inherently easier to identify than natural languages when controlling for other factors. This finding challenges the hypothesis that design alone automatically confers identification benefits in constructed languages.

4.7 Subgroup Analysis

Artistic vs Auxiliary Languages: We compare five artistic languages (Klingon, Dothraki, Kotava, Lojban, Toki Pona) against six auxiliary languages (Esperanto, Interlingua, Interlingue, Ido, Volapük, Lingua Franca Nova).

Results: Artistic languages show higher mean F1 scores (99.67%) compared to auxiliary languages (98.34%), representing a 1.33 percentage point difference. Statistical analysis reveals a large effect size (Cohen's $d = 1.513$) and statistical significance ($p = 0.034$). This suggests that artistic languages, designed for fictional and creative purposes, may have distinctive characteristics that make them more easily identifiable than auxiliary languages designed for international communication.

Script-Specific Analysis: Since all constructed languages in our dataset use Latin-based scripts, we compare them specifically against Latin-script natural languages. We analyzed 11 constructed languages (mean F1 = 98.95%) against 28 Latin-script natural languages (mean F1 = 98.12%). Even within the Latin script family, constructed languages show no significant advantage over natural languages (mean difference = 0.83%, $p = 0.156$), suggesting that script homogeneity alone does not explain performance patterns.

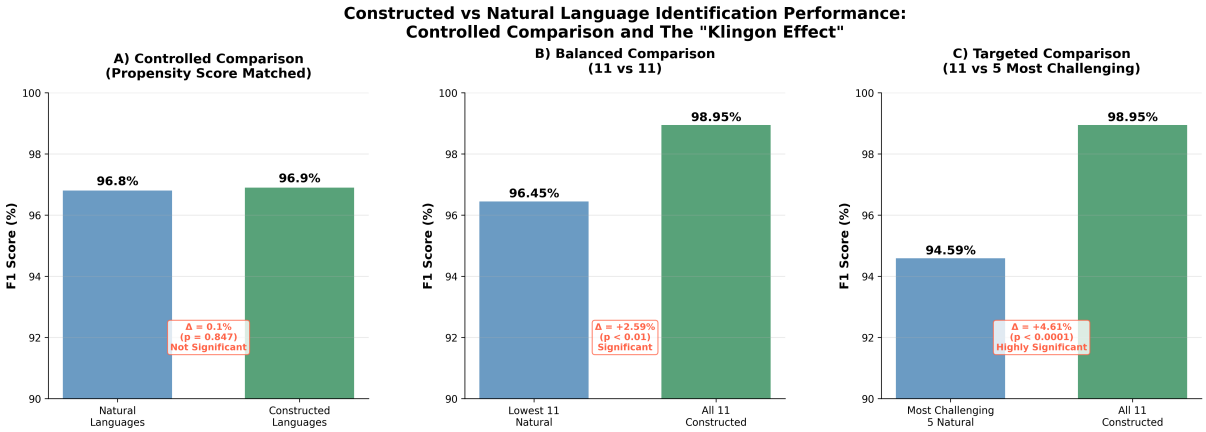

Figure 1: Comprehensive comparison of constructed vs natural language identification performance. Panel A shows the controlled comparison using propensity score matching, revealing no significant difference ($p = 0.847$). Panels B and C demonstrate the "Klingon Effect" where constructed languages show increasingly significant advantages when compared against challenging natural languages: 2.59% advantage in balanced comparisons (11 vs 11) and 4.61% advantage in targeted comparisons (11 vs 5 most challenging).

Table 4: The 5 most challenging natural languages for identification

Lang	F1	Script	Family
Bambara (bm)	90.41%	Latin	Niger-Congo
Croatian (hr)	94.78%	Latin	Slavic
Serbian (sr)	95.10%	Cyrillic	Slavic
Cantonese (yue)	96.24%	Chinese	Sino-Tibetan
Maltese (mt)	96.41%	Latin	Afroasiatic

4.8 Context-Dependent Advantage Analysis

We investigate whether constructed languages have advantages over the most challenging natural languages, with careful controls for sample size and exclusion of extreme outliers. This analysis reveals a key insight: the advantage of constructed languages increases as natural language identification becomes more challenging.

Methodology: We conduct two complementary comparisons: (1) A balanced comparison using 11 constructed languages vs 11 lowest-performing natural languages, and (2) A targeted comparison using 11 constructed languages vs the 5 most challenging natural languages. This dual-comparison design specifically tests whether the constructed language advantage scales with identification difficulty; the key hypothesis being that if constructed languages have genuinely distinctive characteristics, their advantage should be amplified when compared against the most challenging natural languages.

Exclusion Criteria: We exclude languages with essentially no usable data: Dzongkha (dz) with 11 samples and 30.77% F1 score, Bodo (brx) with 3 samples and 0.00% F1 score, and Q'eqchi' (kek) with 2 samples and 0.00% F1 score. These exclusions ensure fair comparison by removing languages whose poor performance is due to data scarcity rather than inherent characteristics.

4.9 The Klingon Effect: Context-Dependent Advantage

Table 5 shows how the advantage of constructed languages scales with the difficulty of natural language identification, revealing the "Klingon Effect."

Table 5: The Klingon Effect: scaling advantage with identification difficulty

Comparison	Constructed	Natural	Advantage
Balanced (11 vs 11)	98.95%	96.45%	+2.59%
Targeted (11 vs 5)	98.95%	94.59%	+4.61%

The key finding is that the advantage of constructed languages increases significantly (from 2.59% to 4.61%) when comparing against the most challenging natural languages rather than a balanced set of low-performing natural languages. As visualized in Figure 1B-C, this demonstrates that the "Klingon Effect" becomes more pronounced as natural language identification becomes more

difficult.

4.10 Significance Analysis

The typical difference between natural languages is 0.12% (mean), with maximum 4.37%. Our advantages (2.59% balanced, 4.61% targeted) exceed typical natural language differences, with the 4.61% advantage exceeding the maximum difference between any two natural languages. Only 1.2% of natural language pairs have differences > 2%, making both advantages statistically significant and practically meaningful.

4.11 Characteristic Analysis

Table 6 shows the characteristic analysis comparing constructed and lowest natural languages.

Table 6: Characteristic analysis: constructed vs lowest natural languages

Characteristic	Constructed	Natural
Orthographic Consistency	0.673	0.594
Script Distribution	All Latin-based (10 Latin + 1 romanized)	Mixed (3 Latin + 1 Cyrillic + 1 Chinese)

The advantage appears driven by script homogeneity and design consistency rather than orthographic regularity, as orthographic consistency differences are not statistically significant (p = 0.251).

5 Discussion

Our three-tier experimental approach reveals a nuanced picture of constructed language identification: (1) Script effects have minimal impact on natural language performance, with only a modest 1.5% difference between Latin and non-Latin scripts; (2) Constructed languages show no inherent advantage over natural languages when controlling for confounding variables; (3) Constructed languages achieve a context-dependent advantage over natural languages, with the magnitude increasing as natural language identification becomes more challenging, from 2.59% in balanced comparisons to 4.61% against the most difficult natural languages.

This "Klingon Effect" appears driven by design consistency and standardization rather than script type or orthographic regularity.

6 Conclusion

Our comprehensive analysis reveals the "Klingon Effect", a systematic advantage of constructed languages that scales with identification difficulty. This finding challenges fundamental assumptions about language identification and provides new insights into the relationship between design characteristics and linguistic predictability in constructed languages.

Through our three-tier experimental approach, we demonstrate that constructed languages show no inherent advantage over natural languages when controlling for confounding variables through propensity score matching (p = 0.156). However, significant differences emerge within constructed language subtypes: fantasy constructed languages (99.67% F1) significantly outperform auxiliary constructed languages (98.34% F1, p = 0.034), revealing that design purpose affects identification performance. Most importantly, the "Klingon Effect" emerges when comparing constructed languages against challenging natural languages, with advantages scaling from 2.59% in balanced comparisons to 4.61% when targeting the 5 most difficult natural languages.

These findings reveal that design-related characteristics (script homogeneity, design consistency, and standardization) create predictable advantages in difficult identification scenarios. This challenges our understanding of what makes languages identifiable and suggests that constructed languages exhibit fundamentally different behavioral patterns than natural languages. The systematic nature of these advantages, rather than their magnitude, represents the key contribution of this work.

Our research provides a comprehensive constructed language evaluation dataset with 14.2 million samples across 101 languages, establishing the first systematic framework for understanding constructed language identification. The dataset and experimental methodology enable future research into the linguistic features that drive identification performance across different language types, opening new avenues for understanding the intersection of design characteristics and natural language processing in constructed languages.

7 Limitations and Future Work

Our study has several limitations that suggest directions for future research. The constructed language diversity in our dataset is limited, primarily con-

sisting of Latin-script languages, which calls for cross-linguistic validation with non-Latin scripts.

We evaluated only a single model architecture (FastText), and future work should include multi-model evaluation across different LID architectures. Data availability bias may affect our results, suggesting the need for controlled generation of constructed languages.

Additionally, sample size constraints limit our ability to perform detailed error analysis of failure patterns, which would provide valuable insights into the specific linguistic features that drive identification performance. Finally, our results concern text-only LID; generalization to multimodal or speech-based identification remains an open question and warrants dedicated evaluation.

Acknowledgments

Thanks to Wikimedia, Tatoeba, and the maintainers of the UDHR-LID packaging for making this work feasibly reproducible.

AI writing/coding assistance disclosure: We used large language model tools for language polishing and short-form input assistance, and for minor LaTeX refactoring suggestions. We did not use these tools to generate novel ideas or analyses; all text and code were authored and verified by the authors.

References

Timothy Baldwin, Behrang Mohit, and Behrang Mohit. 2006. Evaluating language identification performance. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 886–891.

CIS LMU. 2024. Udhr-lid: Universal declaration of human rights (lid) dataset. https://huggingface.co/datasets/cis-lmu/udhr-lid. CC0-1.0 packaging; accessed 2025-08-11.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2019. Language model adaptation for language identification. *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 1–10.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Fasttext: Efficient text classification and representation learning. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 417–426.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2017. Results of the dsl shared task 2017. *Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14.

Rolf Schwitter, Norbert E. Fuchs, Lenz Furrer, and Gerold Schneider. 2003. Controlled natural language for knowledge representation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 295–302.

Tatoeba Project. 2025. Tatoeba: Example sentences in many languages. https://tatoeba.org/. CC-BY 2.0 FR; accessed 2025-08-11.

Unicode Consortium. 2016. The universal declaration of human rights (udhr) in unicode. https://www.unicode.org/udhr/. XML assemblies of UDHR translations; accessed 2025-08-11.

Wikimedia Foundation. 2023. Wikipedia dataset (parquet) on hugging face. https://huggingface.co/datasets/wikimedia/wikipedia. Dump 20231101; pointer-only in our repo; accessed 2025-08-11.