A Joint Detection Framework for Latvian Loanwords and Calques Using Monolingual Data

Yelingyun Zhang, Atis Kapenieks, Marina Platonova
Institute of Digital Humanities, Riga Technical University
Riga, Latvia
{yelingyun.zhang, atis.kapenieks, marina.platonova}@rtu.lv

Abstract

Lexical borrowing is pervasive across languages with extensive cultural contact, yet its automatic detection remains challenging for low-resource languages, especially regarding calques. Existing methods depend heavily on bilingual resources and focus almost exclusively on phonological loanwords, leaving structural borrowing phenomena like calques largely unaddressed by automated tools. This paper proposes a novel joint binary classification pipeline based solely on monolingual data and mBERT, introducing the first large-scale annotated Latvian borrowing dataset with over 3,000 manually labeled entries across three categories: loanwords, calques, and local words. The pipeline adopts a staged decision process grounded in language contact theory, separating surface-level loanwords before tackling the more ambiguous calque category. Experiments demonstrate that our semi-supervised strategy with pseudo-labeling achieves a macro-F1 of 0.854 on an external test set, outperforming both a direct three-way classifier and a GPT-4o zero-shot baseline. These results establish a performance benchmark for the previously unaddressed task of automatic borrowing detection in Latvian, providing empirical tools for borrowing detection in resource-scarce contexts.

Keywords: Lexical Borrowing Detection, Low-Resource NLP, Calque Identification

1. Introduction

Although multilingual pre-trained models have advanced natural language processing (NLP) for low-resource languages, the automatic detection of lexical borrowing in computational linguistics remains challenging. This issue is particularly acute for languages like Latvian. Latvian has a long history of cultural contact, which has deeply influenced its linguistic structure (Veisbergs, 2017). However, the development of automated tools has been hindered due to the lack of large-scale parallel corpora required for mainstream detection methods (Paikens et al., 2024).

Current approaches to borrowing detection are doubly limited. Firstly, they mainly depend on bilingual resources to compare phonological or orthographic features (Mi et al., 2020; Spektor, 2021; Nath et al., 2022). Secondly, these methods focus almost exclusively on direct loanwords, leaving other phenomena of structural borrowing largely unaddressed. Classic theories of language contact posit that borrowing phenomena are stratified, typically progressing from surface-level phonological and morphological changes to deeper structural and semantic integrations; this hierarchical pattern has been repeatedly confirmed in multilingual contact studies ((Matras, 2009, p. 264)). This signals that the phenomena reflecting deep linguistic influences remain blind spots for existing automated tools.

To address this dual challenge — resource dependency and the calque detection gap — this paper introduces a novel joint binary classification pipeline that identifies both loanwords and calques in Latvian using monolingual data only. Our approach is theoretically grounded in language contact models, employing a phased identification process to distinguish phonetic loanwords from the more ambiguous calque and native vocabulary. Furthermore, our methodology is guided by the principle of parsimony, prioritizing a reproducible and simple solution for low-resource contexts.

The main contributions of this paper include the construction of the first large-scale manually annotated dataset designed specifically for Latvian lexical borrowing detection, which provides a systematic differentiation between loanwords, calques, and local words; proposed and empirically validated the joint detection procedure, which reveals the different challenges of the two types of borrowing detection tasks in low-resource contexts. Our semi-supervised strategy established a new baseline for Latvian borrowing detection, achieving a macro-F1 of 0.854 via a semi-supervised strategy.

All code and reproducibility instructions are publicly available on GitHub[1], and the annotated dataset is archived on Zenodo (Zhang et al., 2025).

2. Related Work

Automatic detection of lexical borrowing and language contact phenomena is an important topic in the current fields of computational linguistics, historical linguistics and natural language processing. With the development of multilingual models and deep learning methods in recent years, the field

[1] https://github.com/Yelingyun-Zhang/Latvian-Loanwords-and-Calques-Detection/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10157–10167
11–16 May 2026.

has achieved progress from manual rules to automated detection. However, List et al. (List, 2019) point out that computational methods for quantitative detection of language contact as a whole are still in their early stages, with theoretical and practical difficulties remaining in the field.

Prior to this study, there was no publicly available research on automatic loanword or calque detection for Latvian, and relevant comparisons can only be made with reference to the experimental intervals of mainstream methods in different languages.

Methodologically, the prevailing paradigm for automatic loanword detection treats the task as a classification problem heavily reliant on bilingual feature comparison and increasing architectural and feature complexity. Mi et al. (Mi, 2023) proposed a Web-based bilingual alignment and pseudo-labeling strategy, which automatically generates candidate sets from multilingual resources, and combines deep learning models such as BiLSTM-CRF and multi-feature fusion to improve the performance of cross-language loanword recognition. Zhang et al. (Zhang et al., 2021) compared various pronunciation similarity algorithms such as edit distance, pointwise mutual information (PMI) with sound class alignment (SCA), relying on multilingual parallel data and complex sequence comparison processes. While Nath et al. (Nath et al., 2022) proposed a multimodal feature integration approach that incorporates features such as textual Levenshtein distance, six types of articulatory distances (PanPhon), CLS vectors semantic similarity from multilingual pre-trained models such as mBERT and XLM, as well as an articulatory alignment deep neural network (DNN) within their detection framework, supporting multi-lingual pair/single-pair modelling. More recently, Ali et al. (Ali et al., 2024) pursued a similar goal for the extremely low-resource language Emakhuwa, employing a complex pipeline to generate a Donor candidate via a sequence-to-sequence model and dictionary matching, which was then used as input for a fine-tuned CANINE model. While achieving excellent performance, their state-of-the-art method still relies on bilingual resources (i.e., bilingual lexicons and a donor-language dictionary) in its intermediate steps.

While these methods have achieved promising results in multilingual, resource-rich environments, typically reporting F1 scores in the range of 70% to 93%, they share fundamental limitations. These methods focus on loanwords formed on the basis of phonological similarities. Under the prevailing paradigm of identifying borrowing by comparing morphological and phonological features, most current automatic loanword detection methods rely heavily on large-scale bilingual or multilingual data with complex feature engineering, (Mi et al., 2020; Mi, 2023; Zhang et al., 2021; Spektor, 2021; Nath et al., 2022; Ali et al., 2024). When applied to data-scarce linguistic environments lacking high-quality alignment resources, it is costly to apply a similar approach. It is also worth noting that fine-tuned models tend to be limited in cross-domain generalisation capabilities and flexibility is reduced through adaptation (Matthews and Lillis, 2022), which further complicates the implementation of similar approaches across different low-resource languages.

At the same time, borrowing phenomena in language is not limited to loanwords formed on the basis of phonetic and spelling. Intercultural contact not only affects the text itself, but also profoundly influences the target culture and thinking, even changing the lexical composition, norms, and structure of the language (Veisbergs, 2017). The role played by structurally-induced borrowings, such as calques, cannot be ignored in the text. This phenomenon has been well studied and described in the field of linguistics, however, it remains unexplored in the field of automatic detection.

For low-resource languages, this gap occurs for the dual reason: there is typically little or no availability of annotated corpora for calque detection, and more importantly, current automatic detection methods relying on bilingual comparisons have natural limitations in calque recognition. The fact that the components of the calque contain local morphemes or words leads to a closer proximity to the native word, making it difficult for alignment strategies based on bilingual comparisons and quantified distances to distinguish them from true local words. Therefore, monolingual annotation and automatic modeling strategies to uncover structural borrowing features within the target language itself becomes a potential breakthrough.

3. Task-definition

In the field of language contact and lexical borrowing studies, a **loanword** usually refers to an equivalent word borrowed directly from an external language (donor language) to a local language (recipient language) based on phonology or orthography (Weinreich, 2010, p. 53). A **calque** refers to a word or expression that is clearly imitates the foreign expression in structure or semantics by translating the structure or meaning of the external expression part by part into the local lexeme or phrase (Sewell, 2001). A **local word** refers to a purely local vocabulary whose word form, structure, and semantics have not been directly affected by foreign influence (Veisbergs, 2018).

To achieve automated recognition, this study combines linguistic theory with practical engineering needs and adopts the following three-classification system (see Table 1), Here

Phono/Ortho stands for Phonological / Orthographic, and 'localization level' indicates the degree to which the lexicon has been integrated into the Latvian lexicon.

Words and phrases consisting entirely of borrowed elements are classified as loanwords (e.g., LV: *meteoroloģijas*, EN: *meteorology*; LV: *arktisko aktivitāšu*, EN: *Arctic activities*). Compound words or phrases composed with native morphemes but structurally or semantically modeled after a foreign expression, as well as compounds or phrases containing both loanword and native elements, are classified as calques (e.g.LV: *naudas mazgāšana* [local+local], EN: *money laundering*; LV: *mākslīgais intelekts* [local+loan], EN: *artificial intelligence*). Local word covers both fully local words and long-term historical borrowings, the latter being difficult to trace due to deep localisation.

Borrowing Type	Category	Pattern	Loc.
Phono/Ortho	loanword	loan/ loan+loan	- -
Structural	calque	local+local	-
Hybrid	calque	loan+local	-
Historical	localword	loan	+
Native	localword	local	+ +

Table 1: Classification scheme for Latvian words and phrases used in this study. "Loc." indicates localization level from low (- -) to high (+ +).

It is important to note that our use of the calque category is extended beyond its traditional linguistic definition for the purposes of this computational task. We broaden the class to include hybrid compounds containing both loanword and native elements, which are conventionally treated as a separate typology.

Linguistically, lexical borrowing represents a continuum of integration, and such hybrid forms are widespread and significant in Latvian (Veisbergs, 2018). From a computational perspective, both pure calques and hybrid forms face a similar challenge, the influence of foreign elements manifests mainly in structure and semantics rather than explicit orthographic signals. Through this grouping approach, the model can distinguish words with apparent formal foreignness from those being affected at deeper structural levels, thereby achieving more robust performance. This classification scheme is a pragmatic operationalization designed for the computational task and is not intended to revise or replace established linguistic typologies.

Annotation and criteria development were carried out with Latvian linguistics experts and native speakers, taking into account the etymological dictionaries, phonological and orthographic features as well as structural cross-references (see the Data & Annotation section for more details).

4. Data and Annotation

All data in this study were drawn from the Latvian Wikipedia corpus (Latvian Wikipedia dump downloaded 02-Feb-2025 (Wikimedia Foundation, 2025)). The initial processing involved removing Wiki markup, redirections, category tags, non-Latvian content, and disambiguation pages, retaining only actual Latvian text. Candidate words and phrases were then extracted by sentence segmentation, length and character filtering, and random sampling to ensure adequate representation across linguistic variables. In order to examine the performance of the model under different data sizes, we partitioned the main training set into four stratified subsets: 25%, 50%, 75%, and 100%. The stratification object covers categories (Local word, Loanword, Calque) and types (words/phrases) to ensure comparability across different subsets. Detailed distributions are provided in Table 2.

The external generalisation test set was constructed separately. It was also sampled from the Latvian Wikipedia using different filtering criteria and overlap checking to ensure that there was no significant duplication with the main training set. This test set covers a broader range of topics and linguistic styles, providing a realistic evaluation of model generalization. Its distribution is shown in Table 3.

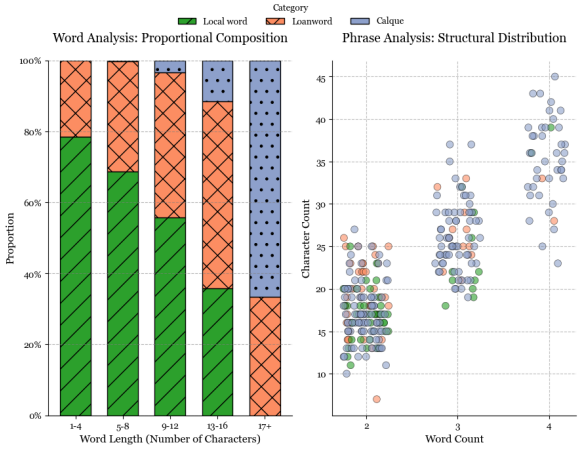

Figure 1: Class distribution of the full training set. (Left) Proportional composition of lexical categories by word length. (Right) Distribution of phrases by word count and total character count.

All samples were manually annotated by the authors. Annotation guidelines were developed and continuously refined under the guidance of domain experts. The annotation process involved checking multilingual dictionaries (Tilde, 2025; Latvijas Zinātņu akadēmijas Terminoloģijas komisija, 2025), comparing receptor and donor languages, and referencing traditional text corpora (Latvijas Univer-

Dataset Version	Total	Local Word	Loanword	Calque	Word	Phrase
25%	785	429	225	131	645	140
50%	1568	858	450	260	1291	277
75%	2350	1285	676	389	1932	418
100%	3130	1714	901	515	2578	552

Table 2: Distribution of the final annotated training datasets

Dataset	Total	Local Word	Loanword	Calque	Word	Phrase
External Test Set	1242	471	465	306	744	498

Table 3: Distribution of the external generalization test set

sitātes Literatūras, folkloras un mākslas institūts, 2025) for ambiguous cases. These bilingual resources are used only during the manual annotation process. The computational pipeline itself follows the monolingual approach, with the mBERT model processing only Latvian-language text during both training and inference phases, without incorporating any external bilingual alignments or lexical features. The overall annotation system and classification criteria are detailed in Section 3. The labelled data were not artificially balanced in terms of category proportions, but rather directly present the natural distribution in the Latvian Wiki corpus. This design choice allows the model to stay close to the actual linguistic environment and accurately learn the dynamics of lexical borrowing in Latvian, thereby improving its generalization capacity and practical applicability.

Figure 1 illustrates the distribution patterns of the full training dataset, revealing a correlation between lexical categories and their physical properties. Shorter words are mainly Local words, accounting for 78.55% of items in the 1-4 character bin. The proportion of Loanwords increases systematically with character length, while Calques are distinguished by their higher structural complexity, exhibiting greater word and character counts.

Comparison Pair	Cohen's Kappa (κ)
Annotator A vs. B	0.8524
Annotator A vs. Gold	0.8162
Annotator B vs. Gold	0.8592

Table 4: Inter-Annotator Agreement (IAA) results.'Gold' refers to the author's original labels

To validate the reliability of our annotation scheme, we conducted an inter-annotator agreement (IAA) study. Two independent annotators (one a native Latvian speaker, the other a linguistics professional with a background in translation studies) were asked to perform a back-to-back annotation of 200 randomly sampled and stratified items, using the classification criteria outlined in Section 3. A high degree of agreement was observed across all pairings, with an average Cohen's Kappa (κ) coefficient exceeding 0.83 (see table 4] for detailed results). This score confirms the stability of the dataset as a reliable linguistic resource for future research. Furthermore, this result highlights the cognitive salience of these linguistic categories; the distinction between loanwords, calques, and local words appears to be a cognitively stable phenomenon for proficient speakers. This finding also provides theoretical support for our computational model, which relies on learning the intrinsic patterns within monolingual data.

5. Method

This study is based on the multilingual pretraining model mBERT, with supervised and semi-supervised strategies for fine-tuning the model to achieve automatic classification of loanwords, calques with local words in Latvian. Compared to traditional recognition methods that rely on rules and feature engineering, mBERT has extensive cross-language knowledge transfer capabilities. Therefore, under its word embedding and subword modelling mechanism, the distributional patterns of loanwords and calques can be effectively captured and generalised from the Latvian annotated data alone without relying on additional external comparison corpus. This advantage makes mBERT particularly suitable for work on borrowing detection in low-resource languages such as Latvian that are subject to multiple linguistic influences.

Since the training data retains the natural category proportions of Latvian, we introduce category-weighted cross-entropy in the loss function, considering that deep models are prone to bias under highly unbalanced data. Through the adaptive adjustment of category weights according to the occurrence frequency of each category in the training set, the minority category samples can obtain higher weights in the overall loss, thus effectively reducing the category imbalance problem. This approach has been shown to be the mainstream deep learning solution in unbalanced data scenarios (Buda et al., 2018). The specific loss function

is as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log p(y_i|x_i) \quad (1)$$

where w_{y_i} denotes the class weight for label y_i, and $p(y_i|x_i)$ is the predicted probability for the true class.

For the reality that annotated data is limited for low-resource scenarios, we used high-confidence pseudo-labeling for semi-supervised expansion of the dataset. This process uses the initial models fine-tuned from various volumes of manually annotated data as a basis for predicting large-scale unannotated corpora. Then filter the pseudo-labelled samples with confidence higher than specified thresholds, mix them with real labelled data with different ratios (e.g., 1:1; 2:1; 3:1), and achieve the expansion of the data set size. The sample proportions and confidence thresholds for pseudo-labeling are set dynamically based on model performance and category distribution. Detailed statistics on the generated pseudo-labels for each data split are provided in the appendix (see table 6).

For model architecture, we adopt a joint binary classification pipeline rather than a direct three-way classifier. Preliminary experiments and data analysis show that loanwords exhibit clear, separable features, whereas calques and local words often overlap and are easily confused in a single-stage classifier. A direct three-way approach is thus more vulnerable to error propagation and class imbalance, especially for low-resource and complex borrowing phenomena.

The final pipeline first applies a binary loanword classifier, followed, if necessary, by a calque classifier, with decision stages tailored for both single words and multi-word expressions (see Fig. 2). In the phrase path, we use a confidence differential threshold (CalqueConf - LoanConf > 0.2). Diagnostics on the internal validation set revealed that the decision branch affected by this parameter is extremely sparse (n=10). Therefore, to prioritize preventing calque false positives in practice, we adopted the conservative heuristic threshold of $\tau = 0.2$. This multi-stage design not only leverages the distinctive features of loanwords for robust early separation, but also systematically reduces misclassification rates for the more ambiguous calque category under real-world class imbalance.

Beyond practical engineering benefits, this approach is theoretically grounded in classic models of language contact, which posit a hierarchical progression of borrowing: surface-level phonological and morphological adaptations typically precede deeper structural and semantic integration (Matras, 2009, p. 264). Our 'easy-to-difficult' staged decision process thus closely mimics the natural stratification of borrowing observed in linguistic contact

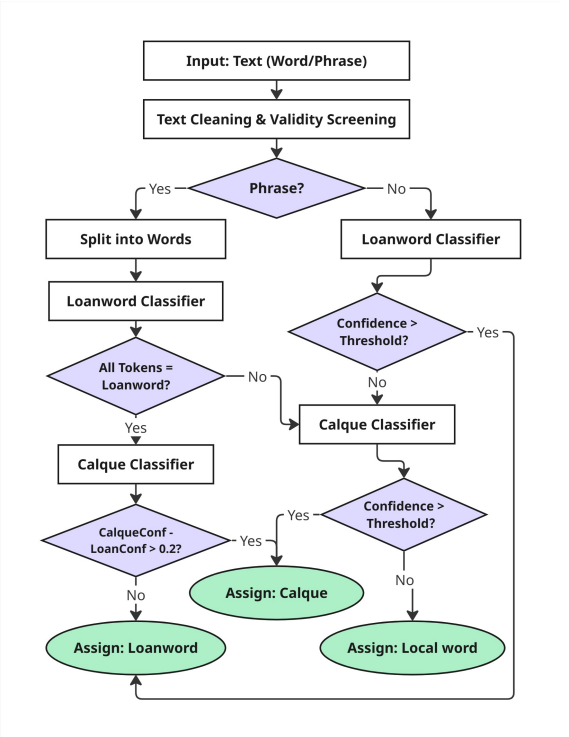

Figure 2: Overall pipeline for joint classification.

zones, supporting improved generalization and robustness in low-resource scenarios.

6. Experiment

The experiments used multiple versions of the previously constructed training dataset (25%, 50%, 75%, 100%) in order to evaluate the performance of the model under different data volumes and settings. From each version of the dataset 10% of the samples are randomly selected as the test set, then about 11.1% of the remaining 90% is divided for the validation set and the rest as the training set. All partitions were stratified by both class (Local word, Loanword, Calque) and form (single word or phrase), ensuring consistent distribution across subsets. In addition, a separate external test set was assembled under different sampling criteria and with a distinct topical and structural profile, to provide an independent assessment of the models' generalizability. The main metrics for evaluation included macro-averaged F1 score, accuracy, precision, and recall. To compare model performance differences, we used a bootstrap method to estimate 95% confidence intervals and assessed the statistical significance of key performance differences by McNemar's test. The main text presents results for core models—namely, the 100% annotated model, the 100% semi-supervised model, the 100% three-way classification model, and the 100% semi-supervised three-way classifi-

Model	Acc	F1-mac	F1-Lw	F1-Cq	F1-Loc	Prec-mac	Rec-mac	CI@F1-mac
100p	0.830	0.828	0.877	0.799	0.807	0.827	0.833	[0.806, 0.849]
100psemi	0.857	0.854	0.899	0.822	0.840	0.852	0.858	[0.835, 0.873]
100p_threeway	0.752	0.753	0.800	0.739	0.720	0.772	0.774	[0.729, 0.775]
100psemi_threeway	0.795	0.792	0.878	0.757	0.742	0.807	0.815	[0.792, 0.815]

Table 5: Performance of main models on 100% annotated data. Model abbreviations: 100p stands for the joint binary pipeline trained on 100% labelleddata; 100psemi incorporates the semi-supervised pseudo-labeling strategy; threeway refers to the three-way classification baseline. Metrics: Acc: accuracy, F1-m: macro-F1, Lw: loanword, Cq: calque, Loc: local word, Prec-m: macro precision, Rec-m: macro recall, CI: 95% confidence interval for macro-F1.

cation model—on the external test set (see Table 5). Full, detailed results for all models and categories, including per-class metrics across all data subsets, are available as supplementary material in the anonymous repository.

All models were built using the HuggingFace Transformers library, with fine-tuning based on the pretrained mBERT model. Hyperparameter choices followed the widely adopted settings in Devlin et al. ((Devlin et al., 2019)); also, AdamW was used for optimization, and class-weighted cross-entropy loss addressed imbalances among classes. Early stopping was based on validation loss, and further details of the implementation, including all parameter settings, are available in the project's code repository.

To gain a comprehensive understanding of how different methods and data strategies impact performance, several sets of controlled comparisons and ablation studies were conducted:

- **Comparison of supervised and semi-supervised training:** Performance was measured both for models trained on annotated data and for models that incorporated pseudo-labelleddata, in order to evaluate the added value of pseudo-labeling for each class.
- **Variation in pseudo-labeling ratio:** For the 75% training set version, additional experiments for the Calque classification task compared pseudo-label to gold label ratios of 2:1 and 3:1. This allowed closer inspection of how the scale of pseudo-labelleddata affected Calque recognition. Since the Loanword category already exhibits high performance at lower pseudo-labeling scales, additional comparisons of this type of scale were not performed.
- **Model Structure Comparison Experiment:** the performance of the three-classification mBERT model is compared with our proposed joint classification pipeline based on the combination of two binary classification models, so as to clarify the impact of classification structure design on the identification performance of each category.

Additionally, to provide a broader reference point for model performance, we evaluated the capabilities of a current large language model (gpt-4o) on this task. For this task, the classification guidelines and typical examples used in our study were presented to gpt-4o at the outset of the session, and the model was then asked to categorize each instance from the external test set. It should be noted that this is not a strict zero-shot evaluation, but rather a preliminary reference, offering an initial benchmark of advanced language model capacity in the absence of established results for Latvian lexical borrowing detection. This baseline serve as a useful point of comparison for subsequent research.

7. Results & Discussion

7.1. Model Performance and Ablation Analysis

The experimental results demonstrate the strong connection between labelled data size, semi-supervised strategy, model architecture and overall performance (see Fig. 3). Macro-F1 increases steadily with more annotated data (0.734 at 25% to 0.828 at 100%), but marginal gains diminish as the dataset approaches saturation—most notably, the largest improvement (0.065) occurs from 25% to 50% annotation, with much smaller gains (0.007) from 75% to 100%.

On the basis of the annotated data, we further validate the performance improvement of the semi-supervised pseudo-labeling strategy. Incorporating high-confidence pseudo-labelleddata yields consistent performance improvements—most notably at 100% annotation, where semi-supervised macro-F1 rises to 0.854 (a 0.026 increase over pure supervision, 95% CI: [0.0948, 0.1440]). Notably, the benefit of pseudo-labeling is especially pronounced as annotation coverage increases. This pattern suggests that despite the diminishing gain effect presented in supervised learning, the model's ability of using weakly supervised signals such as pseudo-labels is further enhanced when sufficient annotation data is available. The above findings highlight

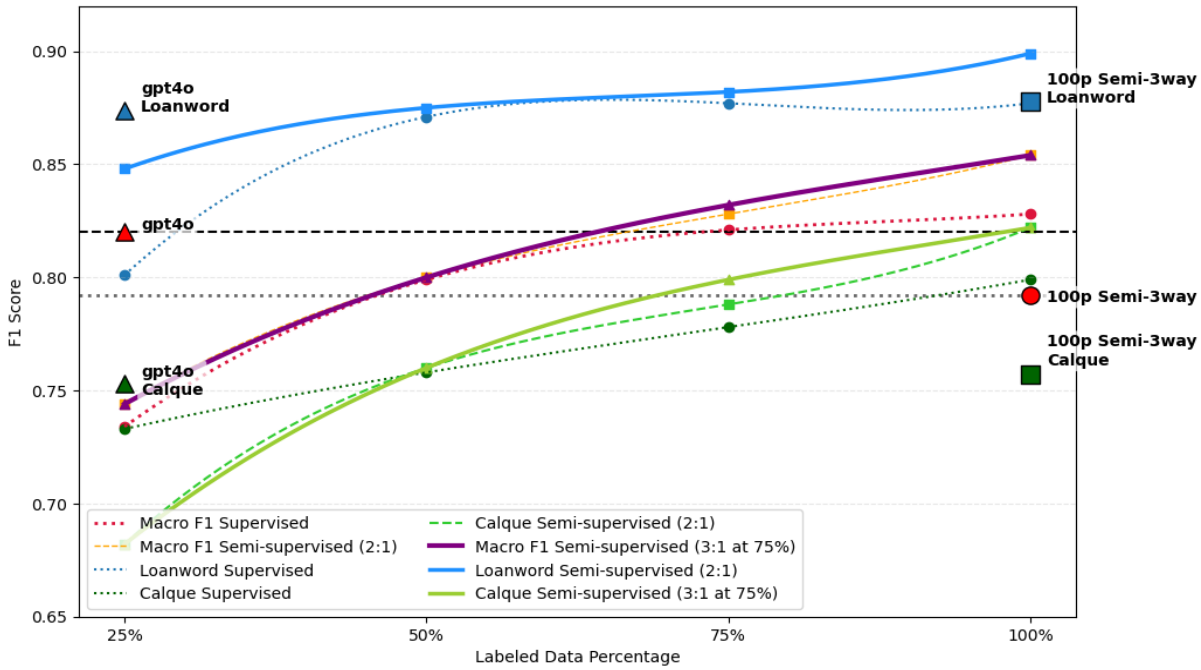

Figure 3: Macro and Class-wise F1 scores of the model under different labelleddata proportions (25% to 100%).The chart compares the training trajectories for supervised training and semi-supervised training using pseudo-labelling strategy. Each individual data point represents a key baseline: the performance of GPT-4o, and three-way classification models (3way) trained on full data.

the value of continuously expanding the annotated dataset and provide an empirical basis for improving the performance of the model in real-world application scenarios.

Further analysis across different binary models reveals variations in the sensitivity of different categories to the pseudo-labeling strategy. Among them, the Loanword category shows higher sensitivity to pseudo-labeling, with an F1 improvement of 0.047 at 25% annotation size, and also a significant improvement of 0.022 at 100% data size. This suggests that Loanword has clear orthographic and morphological features, and that the pseudo-labeling strategy is effective in expanding the diversity of the training samples.

In contrast, for the calque class, due to its complex structure, proximity to the local word morphology, and smaller sample size, performance decreases at low annotation scales (i.e., semi-supervised F1 decreases by 0.051 in the 25% data). The reason for this is speculated to be the limited capacity of the early classifiers, with insufficient accuracy of pseudo-labeling in highly ambiguous samples, and mislabeling instead introducing noise, which affects the overall performance. At 75% and 100% labeling conditions, the introduction of pseudo-labels at a 3:1 ratio resulted in F1 enhancements of 0.020 and 0.023, with no significant diminishing marginal effects. This phenomenon suggests that for the calque recognition task, the current data size has not yet reached the saturation point of performance improvement, and subsequent experiments with further expansion of the dataset could still be expected to achieve sustained gains.

Considering the lack of relevant research in Latvian, this study explores the large language model GPT-4o as a baseline for performance in an almost zero-sample scenario. The Macro-F1 indicator of GPT-4o was found to reach 0.820, approaching the 75% supervised model (0.821). In particular, GPT-4o performs more prominently on the Loanword class (F1 = 0.874), but lags behind on the Calque class (with an F1 of only 0.753), which is lower than any specialised model trained under a high annotation or semi-supervised strategy (only better than supervised and semi-supervised models with 25% data volume), reflecting the instability of large language models in identifying localised expressions.

In addition to data size and pseudo-labeling strategies, this study also examines the impact of different model architectures on recognition effectiveness. Comparative experimental results show that the joint binary classification pipeline model proposed in the method section of this paper outperforms the direct three classification model. In the 100% annotation scenario, the joint binary classifi-

cation model achieves a Macro-F1 of 0.828, while the three classification model reaches only 0.753, which performance is closest to the joint binary classification model obtained by training with 25% of the data volume (with a Macro-F1 of 0.734). After introducing the semi-supervised strategy optimisation (pseudo-labeling ratio of 3:1), the performance of the three classification model improves to 0.792, which is close to the performance of the model obtained by supervised training on 50% of the data (Macro-F1 of 0.799).

This gap results from the fact that three classification models are prone to an accumulation of category boundary misclassification in one-time decision making, and thus such models require a larger amount of data to be invested in order to further enhance the model performance and obtain satisfactory results, and are not an optimal choice for low-resource scenarios. Unlike end-to-end multiclass models, our layered binary approach better reflects the cognitive process of human borrowing detection—enabling the model to learn discriminative features for each borrowing type and reducing category confusion caused by data scarcity. This framework offers a broadly applicable solution for complex, multi-stage classification tasks in resource-constrained settings, with methodologically generalizable value.

Importantly, our findings show that with a properly structured, layered pipeline, monolingual data alone suffices for accurate and robust borrowing type identification, maximizing generalization and adaptability in true low-resource scenarios.

7.2. Error Analysis

To better understand the performance boundaries of the model, we conducted a systematic analysis of the errors made by the primary model (100psemi) on the test set. The analysis revealed that the model's error patterns correlate with the inherent linguistic features present in the data it learned from. As indicated by the data analysis (see Figure 1), the proportion of loanwords increases with character length. The model's reliance on such statistical shortcuts results in its predictable bimodal error pattern. The model error rate reached its height in two different zones (detailed error patterns are shown in Figure 4): At the short-end(1–4 characters), the error rate reached its peak (23.9%). We attribute this to information scarcity. These ultra-short words typically lack sufficient morphological features, leading the model to predict short loanwords as the majority class—native words—which account for 78.6% of this interval.

At the long-end , the error rate rose again (21.7% for 13–16 characters and 17.7% for 17+ characters). The model overgeneralises relationships between length, structural complexity, and foreign influences. Particularly for character counts above 17, when facing counterexamples— i.e., structurally complex native words—shortcuts fail, causing the model to frequently misclassify them into the more structurally complex Calque category.

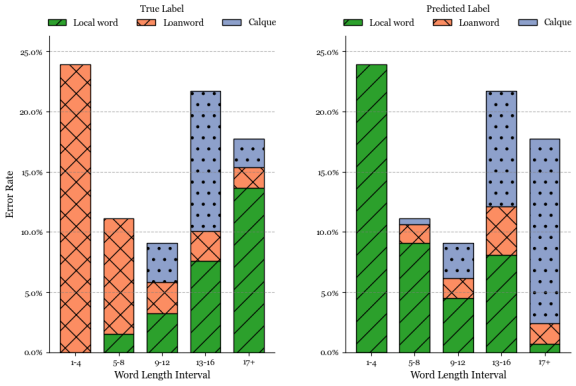

Figure 4: Model Error Rate by Word Length, Error Source (Left) and Misclassification Type (Right).

Although class-weighted cross-entropy was employed as a mitigation strategy, this bidirectional challenge remains. This result indicates that adjusting the loss function at the output level is partially effective, yet insufficient to fundamentally prevent the model from learning and leveraging such a strong statistical bias at the input end. This points the way towards future work. A more principled approach requires intervention at the level of feature representation itself. The goal is to achieve feature disentanglement, encouraging the model to learn representations that capture core linguistic properties while remaining unaffected by surface-level statistical cues such as word length. This goal might be achieved by employing adversarial training schemes that penalize representations that encode information about confounding features.

8. Conclusion

In this paper, for the task of automatic recognition of different types of lexical borrowings in the low-resource language Latvian, a mBERT-based joint binary classification pipeline is proposed and a large-scale, authentically annotated dataset of Latvian lexical borrowings is constructed for the first time.

Experiments show that the method not only achieves significant gains in loanword and calque distinction, but also the semi-supervised pseudo-labeling strategy further expands the model's generalisation ability to unseen samples, providing empirical foundations and technological paths for automated borrowing detection in low-resource environments.

Meanwhile, this paper also verifies that monolingual annotated data together with pre-trained models can achieve effective differentiation of complex lexical borrowing types, which could be expected to provide a powerful tool for related language policy making and educational resources development. This study demonstrates that a structured, layered pipeline is not merely an engineering compromise, but a theoretically driven approach to complex, low-resource tasks—highlighting the value of explicit task decomposition and progressive decision-making for generalization and robustness in NLP.

In future research, we expect to continue expanding the multidomain and multigenre corpus to improve the model's adaptability in diverse scenarios. A primary future direction will be to investigate methods for feature disentanglement, such as adversarial training, to enhance overall model robustness. Further explorations will also include more structurally flexible joint modelling schemes and dedicated toponymic identification mechanisms to better handle boundary cases.

9. Limitations

This study relies on the single-domain data source of the Latvian Wikipedia for modelling and evaluation. Although Wikipedia covers a wide range of content, its language style and vocabulary usage patterns are relatively uniform, which might lead to some limitations in the generalisation ability of the model. Therefore, the effectiveness of the application in other domains, e.g., news, social media or literary texts, has to be further verified, and future research may consider extending to more diverse corpus sources to continue enhancing the model's adaptability and robustness in real-world application scenarios.

In addition to the above limitations, the model's performance on special lexical categories, particularly named entities (especially toponyms), reveals a remaining area for improvement. While semi-supervised learning helps the model recognize more proper nouns, it often fails to identify their morphological variations. This reveals the model's difficulty in generalizing from known entity forms to their unseen morphological variations. Future research could introduce specialised named entity recognition tools to deal with these edge cases in a more refined manner and to continuously improve the reliability of the system.

Furthermore, this study adopted an extended definition of calques, which included hybrid compounds. While effective for capturing structural influence, another important future direction is to separate the identification task of 'pure calques' from that of 'hybrid forms' to achieve a more fine-grained linguistic classification.

10. Ethics Statement

The data used in this study was sourced from publicly available open-source repositories, primarily the Latvian Wikipedia, and linguistic reference dictionaries. As this classification process does not generate language, the risk of producing harmful content is minimal. However, we acknowledge that the underlying pre-trained language model (mBERT) may carry potential biases due to its original pre-training corpus. The datasets, annotation guidelines, code and fine-tuned model weights used in this study have been released as open source to promote transparency, equitable access and reproducibility in low-resource natural language processing research.

11. Acknowledgments

This research was supported by the Latvian State Research Programme "Education" under the project "AI4EDULAB: Advancing AI-Driven, Ethical and Labour Market-Aligned Personalised Learning Content" (Project No. VPP-IZGLĪTĪBA-2025/1-0007). We also acknowledge the support provided by the Institute of Digital Humanities, Faculty of Computer Science, Information Technology and Energy at Riga Technical University.

12. Bibliographical References

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Detecting Loanwords in Emakhuwa: An Extremely Low-Resource Bantu Language Exhibiting Significant Borrowing from Portuguese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4750–4759, Torino, Italia. ELRA and ICCL.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

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.

Johann-Mattis List. 2019. Automated methods for the investigation of language contact, with a focus on lexical borrowing. *Language and Linguistics Compass*, 13(10):e12355. E12355 LNCO-0781.R2.

Yaron Matras. 2009. *Language Contact*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Tamara Matthews and David Lillis. 2022. Experimenting with ensembles of pre-trained language models for classification of custom legal datasets. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 68–78, Trento, Italy. Association for Computational Linguistics.

Chenggang Mi. 2023. Loanword identification based on web resources: A case study on wikipedia. *Computer Speech & Language*, 81:101517.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2020. Loanword Identification in Low-Resource Languages with Minimal Supervision. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(3):43:1–43:22.

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. 2022. A Generalized Method for Automated Multilingual Loanword Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Peteris Paikens, Lauma Pretkalniņa, and Laura Rituma. 2024. A Computational Model of Latvian Morphology. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 221–232, Torino, Italia. ELRA and ICCL.

Penelope Sewell. 2001. The occurrence of calque in translation scripts. *Meta*, 46(3):607–615.

Yulia Spektor. 2021. *Detection and Morphological Analysis of Novel Russian Loanwords*. Thesis, City University of New York.

Andrejs Veisbergs. 2017. Translation language: The major force in shaping modern latvian. *Vertimo studijos*, 2:54.

Andrejs Veisbergs. 2018. Word-Formation Pattern Borrowing in Latvian. *Baltic Journal of English Language, Literature and Culture*, 8:129–146.

Uriel Weinreich. 2010. *Languages in Contact: Findings and Problems*. De Gruyter Mouton.

Liqin Zhang, Ray Fabri, John Nerbonne, and John Nerbonne. 2021. Detecting loan words computationally. In Enoch O. Aboh and Cécile B. Vigouroux, editors, *Variation Rolls the Dice*, pages 269–288. John Benjamins Publishing Company.

13. Language Resource References

Latvijas Universitātes Literatūras, folkloras un mākslas institūts. 2025. Dainu skapis. https://www.dainuskapis.lv/. Accessed: 2025-10-17.

Latvijas Zinātņu akadēmijas Terminoloģijas komisija. 2025. Latvijas nacionālais terminoloģijas portāls. https://termini.gov.lv/. Accessed: 2025-10-17.

Tilde. 2025. Letonika.lv. https://www.letonika.lv/. Accessed: 2025-10-17.

Wikimedia Foundation. 2025. Latvian wikipedia dump (02-feb-2025). https://dumps.wikimedia.org/lvwiki/20250202/. Data dump generated 02-Feb-2025. The live server index for this date has since been rotated.

Yelingyun Zhang, Atis Kapenieks, and Marina Platonova. 2025. Annotated latvian borrowing dataset. https://doi.org/10.5281/zenodo.17380901. Dataset created for this study.

Appendix A. Detailed Experimental Results

Data	Task	Total	Loc.	Lw.	Cq.	P:R
25%	Lw	1308	858	450	-	1:1
25%	Cq	1012	896	-	116	1:1
50%	Lw	3924	2574	1350	-	2:1
50%	Cq	3036	2688	-	348	2:1
75%	Lw	5883	3855	2028	-	2:1
75%	Cq	4628	4054	-	574	2:1
75%	Cq*	6152	5399	-	753	3:1
100%	Lw	10460	6856	3604	-	3:1
100%	Cq	8139	7112	-	1027	3:1

Table 6: Distribution of samples for semi-supervised training. Abbreviations: **Task**: Targeted classification task (**Lw**: Loanword, **Cq**: Calque, **Cq***: Calque with enhanced 3:1 pseudo-label ratio); **Loc.**: Local word; **P:R**: Pseudo-to-real label ratio.

Data %	Strategy	Accuracy	F1_Mac	Precision_Mac	Recall_Mac	95% CI (Acc)
25%	Supervised	0.734	0.734	0.744	0.756	[0.710, 0.758]
	Semi-sup.	0.756	0.745	0.751	0.755	[0.718, 0.768]
50%	Supervised	0.804	0.799	0.797	0.805	[0.776, 0.820]
	Semi-sup.	0.804	0.800	0.799	0.804	[0.777, 0.821]
75%	Supervised	0.825	0.821	0.829	0.818	[0.798, 0.841]
	Semi-sup. (2:1)	0.833	0.829	0.830	0.829	[0.807, 0.849]
	Semi-sup. (3:1)	0.836	0.832	0.833	0.832	[0.811, 0.853]
100%	3-way Sup.	0.752	0.753	0.772	0.774	[0.729, 0.775]
	3-way Semi.	0.795	0.792	0.807	0.815	[0.769, 0.815]
	Binary Sup.	0.830	0.828	0.827	0.833	[0.806, 0.849]
	Binary Semi.	0.857	0.854	0.852	0.858	[0.835, 0.873]
GPT-4o (Zero-shot ref.)		0.828	0.820	0.821	0.826	-

Table 7: Full macro-evaluation metrics of all models evaluated on the external test set. Sorted by annotated data volume. Model strategy abbreviations: **Sup.** refers to purely supervised training; **Semi-sup.** refers to semi-supervised pseudo-labeling (with pseudo-to-real ratios indicated in parentheses); **3-way** refers to the three-class classification baseline; **Binary** refers to our proposed joint binary pipeline.

Model Configuration	Class	Precision	Recall	F1 Score
GPT-4o (Zero-shot)	Loanword	0.859	0.890	0.874
	Calque	0.698	0.817	0.753
	Local word	0.907	0.772	0.834
25% Supervised	Loanword	0.893	0.726	0.801
	Calque	0.607	0.925	0.733
25% Semi-supervised	Loanword	0.800	0.902	0.848
	Calque	0.625	0.752	0.682
	Local word	0.828	0.611	0.703
50% Supervised	Loanword	0.867	0.875	0.871
	Calque	0.706	0.817	0.758
	Local word	0.816	0.725	0.768
50% Semi-supervised	Loanword	0.916	0.837	0.875
	Calque	0.718	0.807	0.760
	Local word	0.763	0.768	0.765
75% Supervised	Loanword	0.931	0.828	0.876
	Calque	0.800	0.758	0.779
	Local word	0.756	0.867	0.808
75% Semi-supervised (2:1)	Loanword	0.910	0.856	0.882
	Calque	0.784	0.794	0.789
	Local word	0.796	0.837	0.816
75% Semi-supervised (3:1)	Loanword	0.910	0.856	0.882
	Calque	0.792	0.807	0.799
	Local word	0.797	0.834	0.815
100% Joint Binary Sup.	Loanword	0.933	0.828	0.877
	Calque	0.752	0.853	0.799
	Local word	0.797	0.817	0.807
100% Joint Binary Semi.	Loanword	0.944	0.858	0.899
	Calque	0.780	0.869	0.822
	Local word	0.833	0.847	0.840
100% Direct 3-way Sup.	Loanword	0.956	0.688	0.800
	Calque	0.607	0.944	0.739
	Local word	0.752	0.690	0.720
100% Direct 3-way Semi.	Loanword	0.935	0.828	0.878
	Calque	0.622	0.967	0.757
	Local word	0.865	0.649	0.742

Table 8: Detailed class-wise evaluation metrics corresponding to the performance trends. The support sizes for the external test set remain constant across all evaluations: Loanword ($n = 471$), Calque ($n = 306$), and Local word ($n = 465$).