LuxBorrow: From Pompier to Pompjee, Tracing Borrowing in Luxembourgish

Nina Hosseini-Kivanani[*1,2], **Fred Philippy**[*3]
[1] University of Luxembourg, Luxembourg
[2] Radio Télévision Luxembourg (RTL), Luxembourg
[3] SnT, University of Luxembourg, Luxembourg
nina.hosseinikivanani@ext.uni.lu, fred.philippy@uni.lu

Abstract

We present LuxBorrow, a borrowing-first analysis of Luxembourgish (LB) news, from 1999 to 2025, based on 259,305 RTL articles and 43.7 million tokens. Our pipeline combines sentence-level language identification (LB/DE/FR/EN) with token-level borrowing resolution in LB sentences using lemmatization, a collected loanword registry, and compiled morphological and orthographic rules. LB remains the matrix language throughout the corpus, while multilingual practice is pervasive: 77.1% of articles contain at least one donor language and 65.4% contain three or four. However, this breadth remains low in intensity, as median code-mixing index (CMI) rises only from 3.90 in LB+1 to 7.00 in LB+3, indicating localized insertions rather than balanced bilingual text. Domain/period summaries show moderate but persistent mixing, with CMI increasing from 6.1 in the 1999 to 2007 period to 8.4 in 2020. We identify 25,444 token-level adaptations, dominated by morphological (63.8%) and orthographic (35.9%) patterns, while lexical borrowings remain rare (0.3%). Diachronically, code-switching intensifies, and morphologically adapted borrowings grow from a base; French overwhelmingly supplies adapted items, with modest growth for German and negligible English. This study aligns with the objectives of the ENEOLI COST Action through its corpus-based analysis of borrowing as a form of lexical innovation in multilingual data.

Keywords: Luxembourgish, lexical borrowing, language identification, morphological adaptation, multilingual media, diachrony

1. Introduction

Lexical borrowing, the adoption of lexical items from a donor language, often with orthographic or morphological adaptation, is pervasive in multilingual media and directly affects the robustness of downstream NLP systems. In the Luxembourgish (LB) context, where sustained contact with German (DE) and French (FR) structures everyday communication in administration, education, and the press, borrowing is common yet insufficiently studied at scale. While code-switching co-occurs in this ecology, we foreground borrowing as our primary object of analysis and treat code-switching descriptively to contextualize borrowing rates. This choice follows the "Simple View" of borrowing and code-switching, which emphasizes that borrowings are typically lexically listed and grammatically integrated, whereas code-switches are more spontaneous and syntactically independent (Treffers-Daller, 2023). Prior work in multilingual European communities, including studies of Portuguese-speaking minorities in Luxembourg, further indicates that insertional code-switching patterns vary with language dominance, underscoring the need to distinguish code-switching from entrenched borrowings (Stell and Couto, 2012). More broadly, European multilingualism often blurs this boundary as speakers integrate foreign items in ways that challenge categorical distinctions (Chan, 2025).

We take a borrowing-first view of multilingual LB news and contribute the following: i) we annotate a 27-year LB news corpus (1999–2025) with hybrid token-level Language identification (LID) (LB/DE/FR/EN) and explicit borrowing/code-switching labels;, ii) we implement a morphological pattern matching system that distinguishes borrowed words from genuine code-switches using contextual analysis (foreign run length, LB neighborhood ratios);, iii) we present comprehensive code-switching metrics (Code-Mixing Index (CMI), entropy, M-index) with temporal evolution analysis; and, and iv) we develop a hybrid language identification approach combining sentence-level LID with token-level refinement and lexicon-based fallback for multilingual news text processing.

We organize our investigation around the following research questions:

- RQ1: How frequent is lexical borrowing across news domains and time periods, and how does borrowing integration compare to general code-switching patterns in multilingual contexts?
- RQ2: What are the patterns of language mixing at the document level in LB news texts, and to what extent does LB remain the dominant matrix language?
- RQ3: Can morphological adaptation patterns

[*] Both authors contributed equally to this work.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3171–3183
11–16 May 2026.

effectively distinguish established borrowing from spontaneous code-switching in closely related Germanic languages (LB/DE)?

- RQ4: How do borrowing integration patterns evolve diachronically (1999–2025), and what competitive dynamics emerge between borrowed forms and native LB alternatives?

Prior research on borrowing and code-switching spans descriptive linguistics, sociolinguistics, and computational modeling. Foundational accounts distinguish entrenched borrowings from in-situ alternation, viewing the two as a continuum governed by lexical integration and community norms (Treffers-Daller, 2023; Masojć, 2023; Deuchar, 2020). Recent NLP studies operationalize these phenomena through token-level language identification, normalization, and cross-lingual robustness benchmarks, often blurring linguistic boundaries for pragmatic classification (Winata et al., 2023; Prabhugaonkar et al., 2017). We build on both strands, with emphasis on resources and methods applicable to low-resource settings where code-switching is prevalent and linguistic resources are scarce. Beyond the Luxembourgish case, this work relates to the broader goals of the ENEOLI COST Action (CA22126), which promotes methodological exchange on the study of lexical innovation across languages and settings. More specifically, it connects with WG2 through its corpus-based treatment of borrowing as an observable and measurable form of lexical innovation in multilingual media.

2. Related Work

Borrowing vs. code-switching. Contact linguistics distinguishes *lexical borrowing*, items integrated into the recipient language's lexicon and grammar, from *code-switching*, i.e., spontaneous alternation between languages within discourse. Foundational work (e.g., Poplack, 2017; Poplack et al., 1988; Poplack and Sankoff, 1984) shows that entrenched borrowings exhibit morphological/phonological integration and community-wide diffusion, whereas code-switches retain structural independence and are often speaker-specific. The "Simple View" further operationalizes this distinction via *listedness* (membership in the mental lexicon) (Treffers-Daller, 2023). Concretely, lexically listed and community-entrenched items, attested in dictionaries, frequent, and morphologically/phonologically integrated into the recipient language, are classified as borrowings, whereas unlisted, ad hoc insertions that retain donor-language structure are treated as code-switches.

From generic mixing to borrowing identification. NLP has shifted from global "mixing" indices to explicit borrowing detection. Álvarez-Mellado and Lignos (2022) introduce Spanish newswire corpora annotated for unassimilated borrowings ("anglicisms") and benchmark CRF, BiLSTM-CRF, and Transformer taggers. This line seeded the ADoBo shared task at IberLEF and released public mBERT taggers (de la Rosa, 2021). Beyond monolingual scenarios, low-resource loanword discovery uses phonetic/semantic similarity to propose candidates with minimal supervision (Miller, 2021; Mi et al., 2020). In parallel, work on social-media corpora distinguishes borrowings from code-switches and integrates borrowing signals into sequence labeling pipelines (Kent and Claeser, 2018). Together, these efforts establish borrowing detection as a distinct token-level task rather than a by-product of LID.

Code-switching research popularized document-level indices (e.g., CMI), but subsequent analyses highlight their sensitivity to utterance length and annotation noise and their weak interpretability for edited text (Srivastava and Singh, 2021; Thara and Poornachandran, 2018; Chandu et al., 2018). Borrowing-centric studies instead report *itemized* statistics: the proportion of borrowed tokens and types, donor-language composition, and evidence of assimilation via morphological integration (Poplack et al., 1988). Following this tradition, we report i) Borrowed Token Rate and Borrowed Type Rate, ii) donor-language entropy over borrowed items (Rosillo-Rodes et al., 2025), and iii) an *assimilation ratio*, the share of borrowed tokens showing lexical/morphological integration.

Because borrowing identification often requires donor-language labels and disambiguation of orthographic homographs, we treat token-level LID as a *supporting* component rather than the primary objective, aligning with prior distinctions between code-switching focused LID and granular borrowing detection (Treffers-Daller, 2023; Kent and Claeser, 2018). Standard code-switching benchmarks (LinCE, GLUECoS) remain useful for pretraining signals and protocol alignment (LID, Part-of-Speech (POS), Named Entity Recognition (NER)), even if their targets are broader than borrowing per se (Khanuja et al., 2020; Aguilar et al., 2020). Extensions to joint LID+POS (Soto and Hirschberg, 2018) and subword/intra-word LID (Sabty et al., 2021; Mager et al., 2019) further improve robustness in morphologically complex settings (Burchell et al., 2024).

For LB, resources have recently expanded. LuxBank provides the first UD treebank for syntax-aware modeling (Plum et al., 2024). The Lëtzebuerger Online Dictionnaire (LOD) supplies lexicon attestations relevant to entrenchment analyses (Zenter fir d'Lëtzebuerger Sprooch, 2025). Neural text normalization (TN) addresses LB's or-

thographic variation and improves downstream robustness for LID and borrowing detection (Lutgen et al., 2025). In generation tasks (e.g., LB headlines), numeric/date/currency TN quality matters for factual consistency; recent work emphasizes explicit factuality checks (e.g., FRANK/AlignScore) and decoding strategies that reduce hallucination (Malon and Zhu, 2024; Zha et al., 2023; Plum et al., 2025).

Gap and our contribution. Current code-switching benchmarks and the ADoBo framework do not close the borrowing-first gap: resources are sparse, and no borrowing-labeled corpus exists for LB (Álvarez-Mellado and Lignos, 2022; de la Rosa, 2021; Mellado and Lignos, 2022; Mi et al., 2020). We try to fill this gap with a token-level, borrowing-centric LB corpus and modeling baselines, using LOD to enable reproducible diachronic analysis and cross-linguistic comparisons beyond well-resourced pairs such as Spanish–English (List and Forkel, 2022).

3. Materials & Methodology

We study multilingual mixing in LB news by: i) identifying language at sentence and token level, ii) operationalizing lexical borrowing versus code-switching, and iii) quantifying diachronic and domain-specific trends. We treat *borrowing* as lexically integrated items (orthographic/morphological accommodation to LB or entrenched lexicalized forms) and *code-switching* as spontaneous alternation without LB-internal integration.

3.1. Dataset

This study uses a large-scale longitudinal corpus of RTL.lu news articles (1999–2025). The corpus was *not* constructed to elicit multilingual behavior; rather, as professionally edited LB reporting produced in Luxembourg's trilingual media environment, it naturally contains instances of lexical borrowing and occasional code-switching. We therefore treat it as an observational resource for contact-linguistic phenomena in contemporary written LB. The RTL news corpus comprises 259,305 articles distributed across six distinct socio-historical periods: Early Digital Era (1999-2007), Financial Crisis period (2008-2011), Post-Crisis Stability (2012-2019), COVID-19 pandemic coinciding with LB orthography reform (2020), COVID continuation (2021), and the ChatGPT/AI Era (2022-2025). This temporal stratification enables diachronic analysis of contact linguistic phenomena across major socio-economic and technological transitions, with notable publication peaks during the 2009 Financial Crisis (16,713 articles) and the 2022 COVID-era surge (29,762 articles). Figure 1 shows the uneven article volume over time and highlights key milestones such as the 2020 orthography reform (brown dotted line).

3.2. Language Identification

We adopt a hierarchical two-stage pipeline: (1) sentence-level LID as a context gate, and (2) token-level borrowing detection inside LB sentences.

1) Sentence-level LID (context gate). We use OpenLID (FastText-based) to predict the sentence language (Burchell et al., 2023). To curb LB over-assignment on short sentences, we apply a length-adaptive posterior threshold: a base of 0.50 is increased up to 0.80 for short inputs; low-confidence cases are routed to Other. Only sentences labeled LB by this gate proceed to the token-level detector.

2) Token-level borrowing detection (within LB sentences). Rather than generic token LID, we run a borrowing detector on LB sentences. Tokens are lemmatized and morphologically normalized with *spellux* (Purschke, 2020), then matched against a curated lexicon of 7 796 loanword entries with donor tags (FR/DE/EN) and assimilation patterns. Each token receives one of NATIVE, FR_LOAN, DE_LOAN, or EN_LOAN. The lexicon is versioned and frozen for this study.

To distinguish between entrenched lexical borrowings and code-switching events, we apply a compiled set of morphological and lexical patterns derived from empirical observations in the corpus. For each candidate token, the system computes contextual features including the run-length of foreign-language tokens and the local LB token ratio within a ±3-token window. Tokens matching a known borrowing pattern and occurring in predominantly LB contexts (short foreign spans and high local LB density) are classified as borrowings. Longer foreign spans are marked as code-switching, while inconclusive cases remain ambiguous.

Concretely, we treat items such as *De Sträit ass duerch e Malentendu entstan.* (**EN**: *The fight occurred due to a misunderstanding.*), as containing an entrenched borrowing, where *Malentendu* behaves like a LB noun in an otherwise LB sentence. By contrast, a sentence such as *D'Buch, ça n'a rien à voir mat dem Film.* (**EN**: *The book has nothing to do with the movie.*) is analyzed as a French code-switching span because it forms a longer contiguous FR clause with donor language syntax.

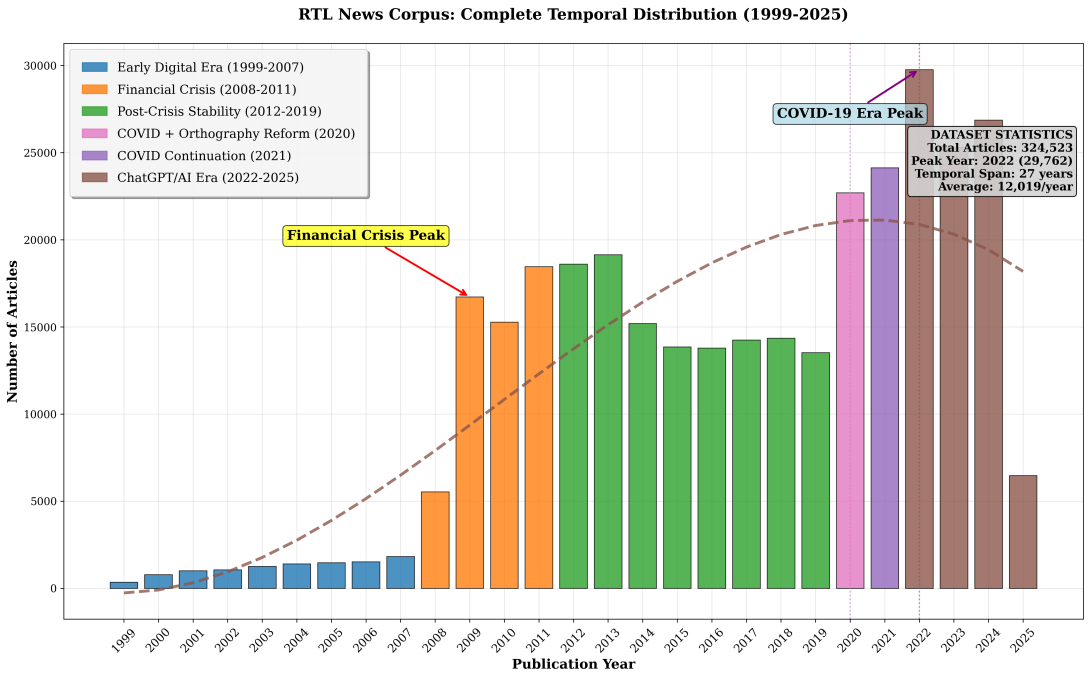

Figure 1: Temporal distribution of RTL news articles (1999–2025) with selected linguistic and technological milestones. The brown dotted line marks the 2020 orthography reform and onset of the COVID 19 period.

3.3. Loanword Identification

Seed Data Currently, there is no official list of loanwords in LB, even though it is widely acknowledged that the language contains a substantial number of borrowings from other languages. To identify such loanwords, we relied on the LOD as our primary resource. The LOD provides LB entries together with their possible multiple meanings and translations into German, French, English, and Portuguese. Since loanword borrowing into LB predominantly occurs from DE, FR, and EN, we restricted our analysis to these three languages and disregarded Portuguese. We further limit the scope to nouns, verbs, and adjectives, as these parts of speech are the most frequent categories for lexical borrowing. Dictionary entries explicitly marked as proper nouns (e.g., chemical elements, brand names, currencies, places, measurement units, etc.) are excluded as well.

Morphological and Orthographic Adaptation Patterns Loanwords in LB are either taken over directly from the donor language or frequently undergo systematic morphological or orthographic adaptations when integrated from source languages (e.g., FR *ajuster* → LB *ajustéieren* via the *-er* → *-éieren* adaptation). For each LB entry, we examine its translations into DE, FR, and EN to determine whether the relationship corresponds to one of these adaptation patterns.

Parallel Borrowing In some cases, a candidate match occurs simultaneously with more than one language. Such situations arise mainly for two reasons: (1) they often involve words that were independently derived across languages from a common root (e.g., *talentéiert*, *talentiert* and *talented* are borrowed from Latin *talentum* and Ancient Greek *tálenton*), or from another source language (e.g., *Alibi*, *Anorak*, or *Tsunami*). These cases do not represent genuine instances of borrowing into LB; (2) they reflect secondary borrowing chains. For example, the LB word *Successioun* matches with both FR and EN *succession*, but the EN form is itself a borrowing from FR.

Without accounting for such cases, these words would misleadingly appear as parallel borrowings rather than revealing the true donor-recipient relationship. To address this, we compiled external lists of known loanwords across the three languages (Wiktionary, 2025a,c,e,d), covering the most common borrowing directions: EN → DE, FR → DE, EN → FR, and FR → EN. This step is crucial not only for identifying which words are loanwords, but also for determining from which language they were originally borrowed.

Shared Inheritance Another issue that must be addressed concerns the linguistic proximity between LB and DE, which stems from their shared West Germanic origin. As a result, many LB words are identical or nearly identical to their DE equivalents. This makes it difficult to determine whether

a given term is genuinely a loanword or whether it developed independently in both languages from a common Old High German root, without any borrowing taking place. To resolve this, words previously identified as DE loanwords are reclassified as non-loanwords if the corresponding DE term can be traced back to Old High German (Wiktionary, 2025b), since it is likely that the LB form also evolved directly from the same ancestral source rather than being borrowed from modern DE.

Human Annotation After the automatic loanword detection, an initial identification and frequency count of the detected words within the corpus is carried out. A human annotator then reviewed these results to identify major shortcomings. The most significant issues included: (1) relevant and frequently occurring loanwords that were not captured by the automatic pattern-matching procedure (e.g., *Pompjee* from FR *pompier*, *Grupp* from FR *groupe*); (2) frequent compound words absent from the seed dictionary data (e.g., *Policepatrull* from FR *police* and *patrouille*); and (3) common orthographic variants or misspellings that, despite being nonstandard, occur frequently (e.g., *entretemps* instead of *entre-temps*, *Akteur* instead of *Acteur*). The annotator also removed words incorrectly tagged as loanwords by the automatic pipeline.

Final Corpus of Loanwords Using this strategy, we identify approximately 3,632 loanwords of DE origin, about 3,201 from FR, and around 535 from EN. For each tagged word, its officially documented spelling variants are also extracted from the source dictionary.

3.4. Statistical analysis

Our analysis combines descriptive statistics with selective inferential methods and advanced indexing. We implement compiled pattern indexing for constant-time morphological lookup and bidirectional lexical indexing for efficient retrieval across 43 M+ tokens. In total, we process 259,305 RTL articles comprising 43.7 M tokens. Token-level language distribution (OpenLID): LB 34.58 M (92.5%), FR 1.41 M (3.8%), DE 1.08 M (2.9%), and EN 0.31 M (0.8%). For RQ1 and RQ4, we compute document- and section-level code-mixing metrics, CMI, Shannon entropy, and M-Index, then aggregate them by news section and predefined year buckets (1999–2007, 2008–2011, 2012–2019, 2020, 2021, 2022–2025), exporting per-group means and token/article counts. For RQ2, we perform document-level code-mixing analysis using OpenLID token-level language labels (LB, DE, FR, EN). We calculate CMI, entropy, and M-index for each document and analyze multilingual scope patterns by comparing monolingual vs. multilingual documents. The analysis includes language combination frequencies, donor language prevalence, and mixing intensity distributions across the corpus using descriptive statistics. For RQ3, we tabulate borrowing frequencies by donor language (FR/DE/EN) and list the most frequent borrowed forms and associated morphological patterns. We compute several established code-switching metrics, including the CMI, Shannon entropy, and the M-Index. These are aggregated by domain (e.g., politics, sports, culture) and by temporal buckets: 1999–2007, 2008–2011, 2012–2019, 2020, 2021, and 2022–2025. Additionally, monthly aggregates are computed for diachronic tracking of code-switching rates, borrowing shares, and donor language distributions. Indexing systems enable real-time pattern matching and scalable lexical analysis, while outputs are summary tables and statistical comparisons intended for downstream visualization and hypothesis testing.

4. Results & Discussion

Multilingual mixing in LB news is measurable and domain-sensitive, but sensitive to small-sample spikes. We therefore report summaries for overall mixing (RQ1/RQ4), document-level patterns (RQ2), and borrowing composition/assimilation (RQ3), noting where estimates are driven by sparse events.

RQ1: The analysis of 259,305 RTL news articles (1999–2025) reveals clear domain-level contrasts in multilingual mixing. Political subdomains exhibit the strongest code-mixing effects: CSV Süden reaches the highest which 38.5, PK Juncker shows the highest entropy (0.882), and DP Süden 2009 achieves the most balanced distribution (M=0.167). However, these extreme values are the result of small-sample events ($n \leq 3$), where a single multilingual speech or press release can inflate the metrics.

In contrast, large-scale sections such as National (52K articles, CMI=8.8) and International (45K articles, CMI=8.0) show stable yet persistent mixing across decades. Temporarily, the intensity of code mixing has increased from CMI=6.1 in 1999–2007 to a peak of 8.4 in 2020 (+38%), with corresponding rises in entropy (0.269 → 0.336) and M-index (0.022 → 0.035), indicating greater lexical diversity and more balanced multilingualism in recent years.

Overall, political coverage dominates multilingual expression, while national and international sections maintain moderate but consistent integration of lexical material from FR, DE, and EN. Table 1 summarizes CMI, entropy, and M-index values for representative domains and periods.

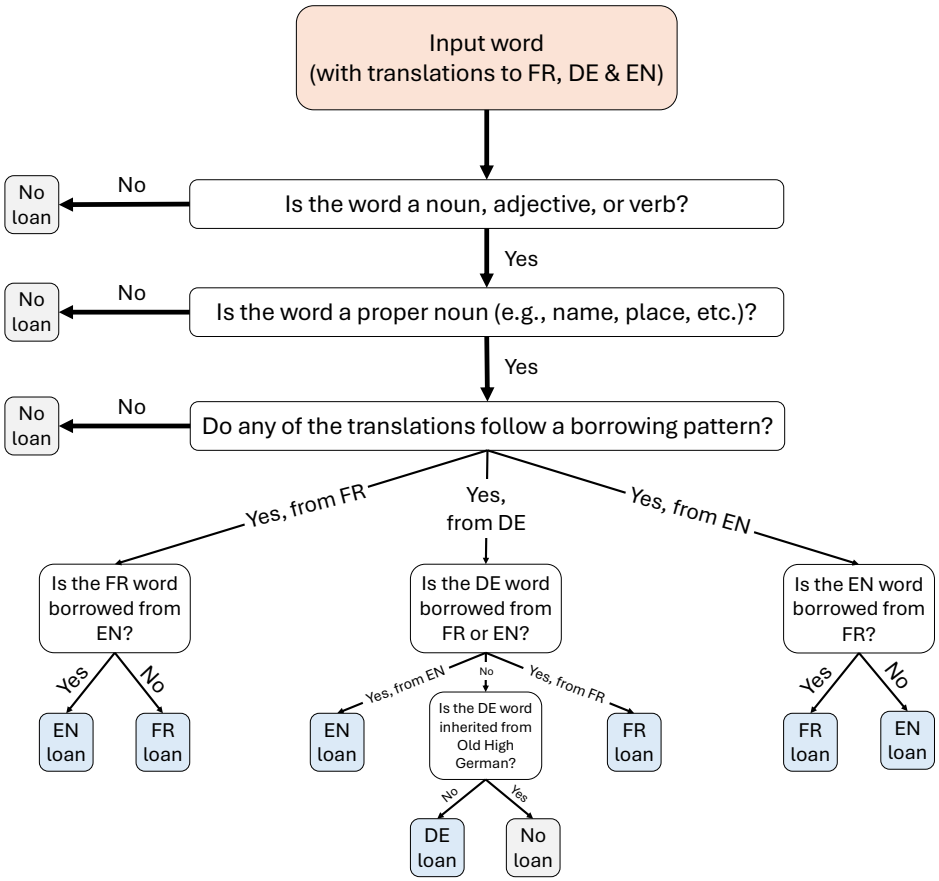

Figure 2: Pipeline for the initial automatic loanword detection. After completing this stage, the resulting loanword corpus was manually reviewed to identify major shortcomings, such as incorrectly tagged words, missing loanwords, or incomplete compound forms, and adjusted accordingly.

Domain / Period	CMI	Entropy	M-Index	#Articles
CSV Süden (Politics)	**38.46**	0.859	0.1542	1
PK Juncker(Press Conf.)	34.62	**0.882**	0.077	1
DP Süden(Politics)	16.67	0.451	**0.167**	2
National	8.82	0.351	0.035	**52,376**
International	8.00	0.322	0.035	45,486
1999–2007	6.10	0.269	0.022	9,795
2020 (peak)	**8.40**	0.336	0.035	17,831

Table 1: Computed code-mixing intensity and diversity (CMI, entropy, M-Index) for RTL news articles by domain and temporal slice.

RQ2: LB remains the matrix language across the corpus (100% of documents). Multilingual practices are nevertheless pervasive: 77.1% of documents mix LB with at least one of DE, FR, or EN, and 65.4% use three or four languages within a single article. Despite this breadth, the intensity of mixing is low on average ($\mathrm{CMI} = 5.25$), indicating LB dominance with localized insertions from donor languages rather than balanced bilingual text.

CMI rises monotonically with the number of co-present languages, but remains low in absolute terms even in highly multilingual documents. LB-only articles are, by definition, unmixed ($\mathrm{CMI} = 0$).

Scope / Combination	Count	%	Med. CMI	IQR
Document-level scope				
LB only	74,946	22.9	**0.00**	0.00–0.00
LB + 1	38,237	11.7	3.90	2.56–5.71
LB + 2	109,367	**33.5**	6.67	4.84–8.82
LB + 3	104,189	**31.9**	**7.00**	5.41–8.90
Most frequent donor combinations				
FR+LB	21,830	8.7	–	–
DE+LB	13,496	5.4	–	–
EN+LB	2,911	1.2	–	–
DE+FR+LB	93,138	**37.0**	–	–
EN+FR+LB	10,268	4.1	–	–
DE+EN+FR+LB	104,189	**41.4**	–	–

Table 2: Document-level language mixing and donor composition. LB dominates even in multilingual articles. FR and DE co-occur most frequently, with EN as a secondary donor.

Median CMI increases from 3.90 in LB+1 to 6.67 in LB+2 and 7.00 in LB+3 (Table 2). This shows that *breadth* (three–four languages) does not entail *high intensity* of mixing: LB continues to dominate, with insertions from donor languages rather than balanced bilingual text.

Among multilingual documents ($n = 251{,}793$), FR is the most frequent donor language (91.1%), followed by DE (86.1%) and EN (49.0%). At

the combination level, the fully multilingual set (LB+DE+FR+EN) accounts for 41.4% of multilingual documents, while the most common triad is DE+FR+LB (37.0%). Dyads are comparatively rare (15.2%), led by FR+LB (8.7%) and DE+LB (5.4%). Overall, FR and DE jointly drive most document-level multilinguality, with EN as a secondary co-present donor.

RQ3: Using our token-independent, rule-based detector on LB sentences, we identified 25,444 borrowing instances distributed across three adaptation types: morphological (16,221; 63.8%), orthographic (9,134; 35.9%), and lexical (89; 0.3%). Within the *ten most frequent concrete patterns* ($n = 25{,}353$), the distribution shows a mixed morphological-orthographic dominance: *on→oun* (8,750; 34.5%), *eur→er* (7,500; 29.6%), and *é→éiert* (5,284; 20.8%) together account for $\approx 85\%$ of the top-10 set. Orthographic shifts such as *on→oun* (8,750; 34.5%), *le→el* (296; 1.2%), *que→ck* (58; 0.2%), and *É→E* (30; 0.1%) represent a substantial component of the adaptation inventory. Morphological patterns including *er→éieren* (2,635; 10.4%), *t→tt* (559; 2.2%), and *isch→esch* (118; 0.5%) remain productive but secondary. Pure lexical borrowings (unadapted forms) are minimal at both class level (0.3%) and individual pattern frequency (*exact*: 89; 0.4%). Overall, the distribution in Figure 3 indicates a mixed orthographic–morphological profile: orthographic adaptation rules produce the highest individual counts, morphological patterns dominate collectively, and lexical borrowings form a minimal tail.

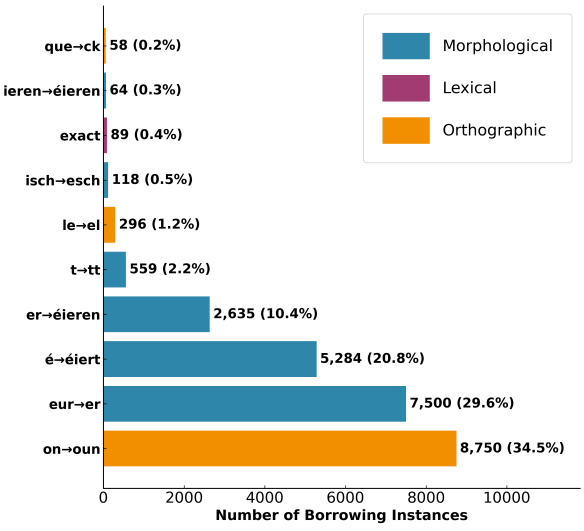

Figure 3: Distribution of observed borrowing pattern types across the RTL news corpus. Borrowing adaptation types across.

RQ4: Figure 4 illustrates the diachronic evolution of code-switching in Luxembourgish news media across 307 months of data (1999–2025). The left panel displays the monthly CS rate (black line) with five-year analytical periods highlighted in distinct colors and period means overlaid as dashed orange lines; dotted vertical lines mark two external milestones, namely the 2020 orthography reform and the onset of the AI era in 2022. The right panel summarizes the mean CS rate and standard deviation for each interval. Code-switching metrics, calculated in five-year analytical intervals, demonstrate a consistent upward trend: 5.76% (1999–2004), 6.49% (2005–2009), 7.28% (2010–2014), 7.62% (2015–2019), and 7.92% (2020–2025). This progression represents a 37.4% increase in CS frequency over the 26 year period in this corpus, suggesting a gradual intensification of multilingual mixing in written discourse. Standard deviations across periods (ranging from 0.5114 in the early years to 0.1752–0.3642 more recently) suggest increasingly stable mixing patterns in most periods.

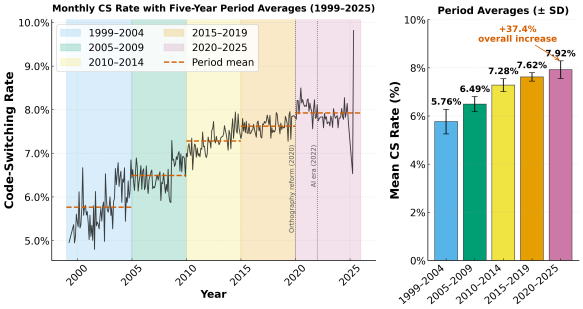

Figure 4: Diachronic evolution of code-switching in Luxembourgish news media (1999–2025). *Left:* monthly CS rate with five-year period means (dashed lines). *Right:* mean CS rate per period; error bars denote one standard deviation.

Morphological borrowing accounts for a small but measurable fraction of multilingual mixing. The borrowing share averages 0.73% across all periods (range: 0.00–6.25%), indicating that the vast majority of observed multilingual elements are insertional code-switches rather than morphologically adapted borrowings. This low borrowing share confirms that code-switching in LB news is predominantly characterized by direct lexical insertion rather than systematic morphological integration.

Donor language patterns in morphological borrowing show strong FR dominance with emerging DE presence. FR maintains overwhelming prevalence: 99.2% (1999–2004), 98.7% (2005–2009), 99.7% (2010–2014), 99.2% (2015–2019), and 97.3% (2020–2025). DE shows modest but consistent growth from 0.8% to 2.7% (+1.9 percentage points), while EN contributions remain negligible (0.0% across all periods). This distribution reflects

the systematic morphological adaptation of FR lexical material into LB, with DE playing a secondary but growing role.

The absolute volume of morphological borrowings has grown substantially, but from a small base. Between the first (1999–2004) and most recent (2020–2025) periods, total borrowing instances increased by a factor of 13.2×, from 483 to 6,398. This growth pattern parallels the broader code-switching trends and reflects the expanding application of morphological adaptation rules over time, particularly for FR lexical material.

Year-over-year data confirm these trends, with code-switching rates climbing steadily: 5.19% in 1999, 7.04% by 2010, and 8.10% in 2020, stabilizing around 7.88% by 2024. DE's morphological borrowing share exhibits parallel growth from 0.6% in 1999 to 2.7% in 2024. These year-specific milestones reinforce the broader period-based observations and underscore the long-term nature of these linguistic developments.

Our 27-year analysis of 259,305 RTL articles shows that LB remains the matrix language while multilingual practices are pervasive but low-intensity. At the document level (RQ2), 77.1% of articles mix LB with at least one of FR/DE/EN, yet median CMI rises only modestly with breadth (LB+1→LB+3: 3.90 → 7.00), indicating localized insertions rather than balanced bilingual text. This profile—broad but shallow mixing with a stable matrix language—is consistent with typological accounts that separate entrenched borrowing from alternational code-switching and emphasize structural dominance of the recipient language in edited prose (Bullock and Toribio, 2009; Poplack, 2017). We summarize the supporting evidence per RQ1–RQ4:

- RQ1. Domain comparisons reveal small-sample spikes: the highest CMI/entropy/M-index values (CSV Süden, PK Juncker, DP Süden 2009) stem from $n \leq 3$ items and should be read as event-driven outliers, consistent with prior findings that entropy and mutual information measures can be inflated or unstable in small datasets (Schroeder, 2004). In contrast, large sections, National (52K, CMI = 8.82) and International (45K, CMI = 8.00), exhibit stable, moderate mixing across decades, aligning with research showing more reliable entropy metrics in large datasets (Hendrix and Sun, 2020).

- RQ2. Widespread but low-intensity multilingual mixing at the document level. Despite 77% of documents containing multiple languages, LB consistently functions as the matrix language. This pattern of broad but shallow mixing aligns with findings that dominant local languages tend to retain structural control even in heavily multilingual media environments (Lam and Matthews, 2020; Stell and Couto, 2012; Bullock and Toribio, 2009).

- RQ3. Token-level adaptation shows a mixed orthographic–morphological profile. From 25,444 borrowing instances, morphological changes are the collective majority (63.8%), orthographic accounts for 35.9%, and lexical accounts are rare (0.3%). The largest single rules are orthographic *on*→*oun* (34.5%) and *eur*→*er* (29.6%), with productive but smaller morphological rules such as *é*→*éiert* (20.8%) and *er*→*éieren* (10.4%). This supports a Zipf-like concentration, a few highly productive rules dominate the head, while lexical-only borrowing forms a minimal tail (Poplack, 2017; Riionheimo, 2002).

- RQ4. Diachronically, code-switching intensifies (≈ +37% from 1999–2004 to 2020–2025), with increasingly stable period-level variation. Within morphological borrowing specifically, the overall share is small (≈ 0.73%) but grows in absolute counts (13.2×, 1999–2004 → 2020–2025). FR overwhelmingly supplies morphologically adapted items (≈ 97–99% by period), while DE shows modest growth (to 2.7%), and EN remains negligible. These trends indicate persistent insertional code-switching as the main driver of multilingual mixing, with selective morphological assimilation concentrated in a small set of highly productive rules (Grimstad, 2017).

For borrowing-aware NLP, sentence-level LID plus token-level, rule/lexicon-guided resolution reduces ambiguity where integrated forms resemble LB natives (Bullock and Toribio, 2009; Poplack, 2017). For borrowing-aware NLP, sentence-level LID plus token-level, rule/lexicon-guided resolution reduces ambiguity where integrated forms resemble LB natives (Poplack, 2017; Bullock and Toribio, 2009), and remains critical given the limitations of sequence-based LID in code-switched environments (Burchell et al., 2024). Given the head-heavy rule distribution and class imbalance, compact, constraint-aware decoders remain attractive, while document-level indices (CMI, entropy, M-index) are best used as context. The Simple View of Borrowing further reinforces the utility of lexical/formulaicity-based classification and low-complexity lexicon integration (Treffers-Daller, 2023), while mutual information scores and morphosyntactic listedness offer compact proxies for borrowing likelihood in ambiguous units. Reporting borrowed token/type rates, donor entropy over borrowed items, and assimilation ratios provides more actionable diagnostics for normalization and

lexicon management (Álvarez-Mellado and Lignos, 2022; Poplack, 2017). Similar metrics have been shown effective in tracking lexical innovation and entropy shifts during rapid borrowing events like COVID-19 media spikes (Foster and Welsh, 2021), supporting their use in modern NLP diagnostics. Finally, the edited-news bias likely underestimates rare adaptations; complementary conversational/social data will help surface tail phenomena. This aligns with findings that code-switching norms vary significantly between written and spoken domains, and that conversational data are more likely to surface nonce or low-frequency borrowings (Dias, 2017; Hickey, 2009).

5. Conclusion & Future Work

Our 27-year study of LB news shows a stable matrix language with pervasive but low-intensity multilingual practice: 77% of articles mix LB with FR/DE/EN, yet median CMI remains modest even for LB+3 documents. Domain extremes are driven by small samples, whereas large sections (e.g., *National*, *International*) display moderate, persistent mixing. Token-level analysis reveals a head-heavy adaptation profile: borrowing instances are dominated collectively by morphological rules, alongside strong orthographic contributors (e.g., *on→oun*, *eur→er*); fully lexical, unadapted items are rare. Diachronically, overall code-switching intensifies, while morphologically adapted borrowings remain a small share and are overwhelmingly FR-sourced, with a modest rise in DE. Methodologically, sentence-level LID coupled with token-level, rule/lexicon-guided resolution offers practical robustness. Borrowing-first diagnostics (borrowed token/type rates, donor entropy over borrowed items, assimilation ratios) are more actionable than headline document-level indices alone for tasks such as lexicon enrichment, normalization of entrenched hybrids, and evaluation of Automatic Speech Recognition (ASR) or Machine Translation (MT) systems on loanwords.

Future work will extend LuxBorrow in four directions. First, we will create a manually annotated gold standard subset with inter-annotator agreement to better calibrate the rule-based pipeline and quantify remaining error. Second, we will add conversational and social media data to capture rarer adaptation patterns that are underrepresented in edited news. Third, we plan to combine our pattern-based detector with data-driven models, for example, sequence taggers that use the loanword registry as weak supervision, to improve recall and portability to other language pairs. Finally, we will evaluate the impact of an `is_loan` feature in downstream tasks such as Luxembourgish normalization, NER, ASR, and machine translation.

Data and code availability. Code for preprocessing, LID, borrowing detection, and reproduction of the figures is available at `github.com/NinaKivanani/LuxBorrow-LREC`. Due to copyright and database rights of RTL Luxembourg, the full article texts are not redistributed, but access can be requested via `ai@rtl.lu`; we release scripts, schemas, pattern lists, and aggregate outputs.

Limitations

Our pipeline is largely automated and not anchored in a fully manually annotated gold standard corpus. This improves scale and reproducibility but introduces uncertainty. Sentence gating and token routing depend on LID confidence, so short or ambiguous sentences can misclassify borrowing vs. code switching. We plan to mitigate these limits via targeted human annotation with inter annotator agreement, active learning driven expansion of patterns and lexicon, cross genre validation (speech and social media), and mixed effects modelling with domain and time controls.

LuxBorrow reflects the editorial choices, style, and topic mix of a single major media outlet, RTL Luxembourg, over 27 years. As such, it is not a representative sample of all Luxembourgish language use. Models or tools trained on this corpus may inherit RTL specific biases in topic coverage, register, and language choices. We explicitly document domain distributions and mixing patterns and encourage users of LuxBorrow to take these biases into account when drawing sociolinguistic conclusions or building downstream NLP systems.

6. Ethical Considerations

Data source and legal basis The LuxBorrow corpus is built from online news articles published by RTL Luxembourg between 1996 and 2025. The data were obtained under a formal research collaboration between RTL Luxembourg and the University of Luxembourg, with controlled access to the internal news archive provided by an RTL collaborator. No user accounts were accessed, no technical protection measures were bypassed, and we did not perform large-scale scraping of the public website. All processing follows the applicable terms of use agreed with RTL Luxembourg and EU text and data mining provisions for non-commercial scientific research.

Copyright and database rights All articles remain under the copyright and database rights of RTL Luxembourg. Our preprocessing, annotation, and analysis operate on copies stored on secure institutional infrastructure, including the MeluXina

high performance computing system, a EuroHPC supercomputer hosted and operated by LuxProvide in Luxembourg, which we access via a research allocation from the University of Luxembourg, for non commercial scientific research purposes. In line with RTL's rights and applicable EU copyright and database rights legislation, we do not redistribute the full text of the corpus. Instead, we release only derived artefacts that do not substitute for the original content, such as (i) the borrowing annotation schema and pattern lists, (ii) scripts to reproduce our pipeline on any legally obtained Luxembourgish news corpus, (iii) aggregate statistics and plots reported in this paper, and (iv) small, non substitutable excerpted examples where required for illustration. Researchers who require access to the underlying RTL corpus may request it via our RTL collaborator, subject to RTL's approval and any necessary data sharing agreements.

Data protection and privacy News articles may mention identifiable individuals. These mentions concern public journalistic content that is already lawfully available online, but they remain sensitive textual data. We do not link the corpus to external personal records, profile individual persons, or attempt to infer sensitive attributes. All analyses are carried out at the level of tokens, sentences, and aggregated document sets, not at the level of individuals. The work therefore adds minimal incremental risk to the persons mentioned in the articles. Data are stored and processed on secure institutional infrastructure, including the MeluXina high performance computing system hosted and operated by LuxProvide in Luxembourg, in compliance with the GDPR and applicable national data protection requirements.

Normative use and impact As discussed in Section 5, LuxBorrow reflects the editorial choices and topic mix of a single media outlet and is not a representative sample of all Luxembourgish language use. Our borrowing first diagnostics are intended as a descriptive tool for contact linguistics and multilingual NLP, not as a prescriptive standard for "correct" Luxembourgish. We discourage the use of this resource for normative policing of lexical borrowing or for applications that marginalise code switching practices in everyday multilingual communication. Aggregate patterns of borrowing or code mixing may correlate with social or regional factors, but such correlations should be interpreted with caution to avoid reifying stereotypes or overgeneralising from a single media source.

7. Acknowledgements

We thank RTL Luxembourg and Tom Weber for providing access to the news archive and for supporting its use for research purposes. This work highly benefited from the collaborative network fostered by the **ENEOLI COST Action (CA22126)**, supported by COST (European Cooperation in Science and Technology), and also within the project LuxVoice (project reference 19205922) from the FNR.

8. Bibliographical References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. Lince: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813.

Elena Álvarez-Mellado and Constantine Lignos. 2022. Detecting unassimilated borrowings in spanish: An annotated corpus and approaches to modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3868–3888.

Barbara E Bullock and Almeida Jacqueline Toribio. 2009. *The Cambridge handbook of linguistic code-switching*. Cambridge university press.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, Robert Thompson, and Kenneth Heafield. 2024. Code-switched language identification is harder than you think. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 646–658.

Brian Hok-Shing Chan. 2025. Borrowing or code-switching? single-word english prepositions in hong kong cantonese. *Open Linguistics*, 11(1):20250045.

Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Günter Neumann, Manoj Chinnakotla, Eric Nyberg, and Alan W Black. 2018. Code-mixed question answering challenge: Crowd-sourcing data and techniques. In

Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, pages 29–38.

Javier de la Rosa. 2021. Adobo 2021: The futility of stilts for the classification of lexical borrowings in spanish. In *IberLEF@ SEPLN*, pages 947–955.

Margaret Deuchar. 2020. Code-switching in linguistics: A position paper. *Languages*, 5(2):22.

Silvia Fortuna Dias. 2017. *Code-Switching and Lexical Borrowing among Brazilian Portuguese and English Bilinguals in Chicagoland*. Ph.D. thesis, University of Illinois Chicago.

Stuart Mannix Foster and Alistair Welsh. 2021. A 'new normal'of code-switching: Covid-19, the indonesian media and language change. *Indonesian Journal of Applied Linguistics*, 11:1.

Maren Berg Grimstad. 2017. The code-switching/borrowing debate: evidence from english-origin verbs in american norwegian. *Lingue e linguaggio*, 16(1):3–34.

Peter Hendrix and Ching Chu Sun. 2020. The role of information theory for compound words in mandarin chinese and english. *Cognition*, 205:104389.

Tina Hickey. 2009. Code-switching and borrowing in irish 1. *Journal of Sociolinguistics*, 13(5):670–688.

Samantha Kent and Daniel Claeser. 2018. Incorporating code-switching and borrowing in dutch-english automatic language detection on twitter. In *Proceedings of the Future Technologies Conference*, pages 418–434. Springer.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585.

Chit Fung Lam and Stephen Matthews. 2020. Intersentential code-switching and language dominance in cantonese–english bilingual children. *Journal of Monolingual and Bilingual Speech*, 2(1):73–105.

Johann-Mattis List and Robert Forkel. 2022. Automated identification of borrowings in multilingual wordlists. *Open Research Europe*, 1:79.

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural text normalization for luxembourgish using real-life variation data. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 115–127.

Manuel Mager, Özlem Çetinoğlu, and Katharina von der Wense. 2019. Subword-level language identification for intra-word code-switching. 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 2005–2011.

Christopher Malon and Xiaodan Zhu. 2024. Self-consistent decoding for more factual open responses. *arXiv preprint arXiv:2403.00696*.

Irena Masojć. 2023. Borrowings or code-switching? analysis of ways of morphosyntactic integration of lithuanian and russian nouns in the novel" robczik" by bartosz połoński. *Slavistica Vilnensis*, 68(1):87–101.

Elena Álvarez Mellado and Constantine Lignos. 2022. Borrowing or codeswitching? annotating for finer-grained distinctions in language mixing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3195–3201.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2020. Loanword identification in low-resource languages with minimal supervision. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(3):1–22.

John Edward Miller. 2021. Neural borrowing detection with monolingual lexical models. In *Proceedings of the student research workshop associated with RANLP 2021*.

Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. Luxbank: The first universal dependency treebank for luxembourgish. *TLT 2024*, page 30.

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 93–104.

Shana Poplack. 2017. *Borrowing: Loanwords in the Speech Community and in the Grammar*. Oxford University Press.

Shana Poplack and David Sankoff. 1984. Borrowing: the synchrony of integration.

Shana Poplack, David Sankoff, and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104.

Neha Prabhugaonkar, Sai Kiran Peketi, Kavita Ganeshan, and Unnikrishnan Sureshkumar. 2017. Differentiating code-borrowing from code-mixing. In *Proceedings of the 4th ACM IKDD Conferences on Data Sciences*, pages 1–2.

Christoph Purschke. 2020. spellux – automatic text normalization for luxembourgish.

Helka Riionheimo. 2002. How to borrow a bound morpheme? evaluating the status of structural interference in a contact between closely-related languages. *Finnish Journal of Linguistics*, (15):187–218.

Pablo Rosillo-Rodes, Maxi San Miguel, and David Sánchez. 2025. Entropy and type-token ratio in gigaword corpora. *Physical Review Research*, 7(3):033054.

Caroline Sabty, Islam Mesabah, Özlem Çetinoğlu, and Slim Abdennadher. 2021. Language identification of intra-word code-switching for arabic–english. *Array*, 12:100104.

Marcin J Schroeder. 2004. An alternative to entropy in the measurement of information. *Entropy*, 6(5):388–412.

Victor Soto and Julia Hirschberg. 2018. Joint part-of-speech and language id tagging for code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–10.

Vivek Srivastava and Mayank Singh. 2021. Challenges and limitations with the metrics measuring the complexity of code-mixed text. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 6–14.

Gerald Stell and Maria Del Carmen Parafita Couto. 2012. Code-switching practices in luxembourg's lusophone minority: a pilot study on how an immigrant community linguistically behaves differently from the majority. *Zeitschrift fur Sprachwissenschaft*, 31(1):153–185.

S Thara and Prabaharan Poornachandran. 2018. Code-mixing: A brief survey. In *2018 International conference on advances in computing, communications and informatics (ICACCI)*, pages 2382–2388. IEEE.

Jeanine Treffers-Daller. 2023. The simple view of borrowing and code-switching. *International Journal of Bilingualism*, 29(2):347–370.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

9. Language Resource References

Wiktionary. 2025a. *Category:English terms borrowed from French – Wiktionary*. [Accessed on 20-10-2025]. Available at: https://en.wiktionary.org/wiki/Category:English_terms_borrowed_from_French.

Wiktionary. 2025b. *Category:German terms derived from Old High German – Wiktionary*. [Accessed on 20-10-2025]. Available at: https://en.wiktionary.org/wiki/Category:German_terms_derived_from_Old_High_German.

Wiktionary. 2025c. *Catégorie:Anglicismes en français – Wiktionary*. [Accessed on 20-10-2025]. Available at: https://fr.wiktionary.org/wiki/Cat%C3%A9gorie:Anglicismes_en_fran%C3%A7ais.

Wiktionary. 2025d. *Verzeichnis:Deutsch/Anglizismen – Wiktionary*. [Accessed on 25-09-2025]. Available at: https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Anglizismen.

Wiktionary. 2025e. *Verzeichnis:Deutsch/Gallizismen – Wiktionary*. [Accessed on 25-09-2025]. Available at: https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Gallizismen.

Zenter fir d'Lëtzebuerger Sprooch. 2025. *Lëtzebuerger Online Dictionnaire (LOD)*. Official reference dictionary for Luxembourgish.

10. Appendices

10.1. Loanword Identification

Table 3 summarizes the selected transformation patterns used to identify loanwords.

Pattern	LU Word	Source Word
nner→néieren	sanctionéieren ordonéieren deconéieren	sanctionner ordonner deconner
exact	Vignette sektoriell Absence	vignette sektoriell absence
er→éieren	filetéieren immuniséieren marginaliséieren	fileter immuniser marginaliser
le→el	Decapotabel accessibel favorabel	décapotable accessible favorable
on→oun	Motivatioun Abstentioun Exploitatioun	motivation abstention exploitation
-e	Karwoch Schlei Manikür	Karwoche Schleie Maniküre
é→éit	Stabilitéit Majestéit Disponibilitéit	stabilité Majesté disponibilité
é→éiert	bornéiert arméiert accidentéiert	borné armé accidenté
on→oun + c→k	Fiktioun Kollektioun Attraktioun	fiction collection attraction
ir→éieren	reusséieren investéieren demoléieren	reussir investir demolir
ät→éit	Skurrilitéit Aggressivitéit Fakultéit	Skurrilität Aggressivität Fakultät
que→ck	Barock Attack Pick	baroque attaque pique
eur→er	Rapporter Coiffer Entreprener	rapporteur coiffeur entrepreneur
É→E	Eclat Epicière Ecart	éclat épicière écart
i→éiert	introvertéiert etabléiert	introverti etabli
nné→néiert	proportionéiert passionéiert	proportionné passionné

Table 3: Examples of orthographic and morphological transformations in LB loanwords.