A Subword Embedding Approach for Variation Detection in Luxembourgish User Comments

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke
University of Luxembourg, Esch-sur-Alzette, Luxembourg
{anne-marie.lutgen, alistair.plum, christoph.purschke}@uni.lu

Abstract

This paper presents an embedding-based approach to detecting variation without relying on prior normalisation or predefined variant lists. The method trains subword embeddings on raw text and groups related forms through combined cosine and n-gram similarity. This allows spelling and morphological diversity to be examined and analysed as linguistic structure rather than treated as noise. Using a large corpus of Luxembourgish user comments, the approach uncovers extensive lexical and orthographic variation that aligns with patterns described in dialectal and sociolinguistic research. The induced families capture systematic correspondences and highlight areas of regional and stylistic differentiation. The procedure does not strictly require manual annotation, but does produce transparent clusters that support both quantitative and qualitative analysis. The results demonstrate that distributional modelling can reveal meaningful patterns of variation even in "noisy" or low-resource settings, offering a reproducible methodological framework for studying language variety in multilingual and small-language contexts.

1 Introduction

Variation in language is often treated as noise in NLP pipelines (Eisenstein, 2013; Al Sharou et al., 2021). Spelling differences, orthographic inconsistencies, and regional forms are typically normalised or removed to simplify token space, which can erase sociolinguistic signal (Baron and Rayson, 2008). Work in sociolinguistics and large-corpus dialectology shows that such variation is systematic and informative for geography and social structure (Grieve et al., 2019). Subword modelling has long been used to handle non-standard forms in practice and improves classification in noisy settings (Munro and Manning, 2010).

For under-researched languages and varieties, identifying and extracting language variation remains challenging. Pre-processing tools such as VARD insert modern equivalents for historical spellings to aid search and tagging, relying on lexicons and edit-distance style methods, and target normalisation rather than discovery (Baron and Rayson, 2008). Embedding-based studies indicate that distributional models encode many types of spelling variation and near-orthographic similarity, though evaluations are typically based on curated sets of variant pairs and focus on representation quality (Nguyen and Grieve, 2020). Closely related work in context-sensitive spelling correction uses word and character n-gram embeddings to map misspellings to canonical forms, again optimising correction and not mining new variants (Fivez et al., 2017). Research on dialectal change detection has brought models for geographic differences, but does not allow for directly extracting unconstrained orthographic families from raw text (Jiang et al., 2020; Pham et al., 2024). Broader multilingual analyses of surface-form overlap highlight that form-level variation carries structure that models can exploit (Kallini et al., 2025).

There are methods to represent variation and to correct it, and there are resources that label dialectal differences. What is less supported are methods that detect and mine previously unlisted variant families directly from raw text without seed lexicons, and that do so beyond strictly dialectal contrasts. We propose the methodology laid out in this paper to address this gap. With this methodology, we are able to discover candidate variation families from distributional evidence and provide transparent scores for downstream qualitative and quantitative analysis.

The main contributions of the research carried out and presented in this paper are:

(1) A reproducible methodology[1] for inducing lexical and orthographic variation from raw

[1] https://github.com/plumaj/vadamt

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 113–122
March 29, 2026. ©2026 Association for Computational Linguistics

text using subword embeddings, similarity-based grouping, and controlled pruning without relying on predefined variant lists or normalisation rules.

(2) A large-scale empirical study of variation in Luxembourgish user comments, showing that the automatically induced families capture systematic patterns and provide a structured basis for qualitative linguistic analysis.

2 Background

Luxembourgish is a small language situated in a dense multilingual environment, with extensive contact to both German and French. Its grammatical structure derives from Moselle Franconian (Gilles, 2019), while sustained contact with French has shaped its lexicon, borrowing patterns, and code-switching practices. Written Luxembourgish displays considerable orthographic and lexical diversity, especially in informal online settings, which makes it a challenge for NLP.

Initial work on computational methods for Luxembourgish was limited but established an initial foundation. Adda-Decker et al. (2008) introduced the first tools and corpora for automatic processing. Subsequent studies examined characteristic orthographic phenomena (Snoeren et al., 2010) and provided early annotated resources for mixed-language processing (Lavergne et al., 2014).

Recently, research activity has increased noticeably. Work has expanded to sentiment analysis (Sirajzade et al., 2020; Gierschek, 2022), orthographic correction (Purschke, 2020), syntactic annotation (Plum et al., 2024), topic classification (Philippy et al., 2024), comment moderation (Ranasinghe et al., 2023), and automatic normalisation (Lutgen et al., 2025). A broader set of classification tasks, including named entity recognition, was provided by Lothritz et al. (2022), and the generative benchmark `LuxGen` was introduced by Plum et al. (2025). These works illustrate the rapid growth of Luxembourgish NLP but also reveal gaps in coverage, consistency, and domain diversity.

Model development reflects a similar trajectory. Strategies range from cross-lingual transfer from German, as in `LuxGPT` (Bernardy, 2022), to data augmentation with synthetic Luxembourgish text in `LuxemBERT` (Lothritz et al., 2022), and balanced multilingual pretraining for `LuxT5` (Plum et al., 2025). Other models include ENRICH4ALL (Anastasiou, 2022) for administrative-domain chatbots and the LUX-ASR speech recognition models (Gilles et al., 2023a,b). Together, these efforts demonstrate progress, yet available datasets remain fragmented and vary widely in size, annotation schemes, and linguistic phenomena.

One area that has received little explicit attention is lexical and orthographic variation. Lutgen et al. (2025) develop a qualitative performance test to evaluate normalisation models for specific orthographic variants. In linguistics, the Variation Atlas by Gilles (2021) represents the most comprehensive overview of phonological, lexical, grammatical and regional variants in Luxembourgish. This atlas is constructed by using an app to collect users' speech inputs for specific phenomena and socio-demographic data, which is then transcribed, analysed, and published (Entringer et al., 2021).

3 Methodology

The methodology adopted in this study combines semi-supervised modelling with targeted qualitative analysis to identify lexical and orthographic variation directly from raw text, without relying on predefined dictionaries or normalisation rules. Throughout, spelling diversity is treated as a source of linguistic information rather than noise, allowing the unsupervised detection of previously unrecorded orthographic and mixed variants while ensuring transparency and reproducibility. This design supports large-scale induction alongside qualitative interpretation, and aligns with recent work arguing that normalisation can obscure meaningful patterns in non-standard and partly standardised varieties (Grieve et al., 2019; Kallini et al., 2025).

First, subword embeddings[2] are trained on the raw corpus to obtain distributional representations that preserve orthographic detail. Second, these embeddings are used to induce groups of related forms through a combination of cosine and n-gram similarity, followed by controlled pruning and aggregation across relevant *dimensions* such as users, time periods, or domains. Third, the automatically identified groups are examined manually to assess their linguistic coherence and to trace patterns that are not fully captured by numerical criteria.

3.1 Distributional Embeddings

Before outlining the methodology in more detail, we briefly characterise what word- and subword-

[2] In this work, we use the term subword embeddings to refer to embeddings constructed from fixed character n-grams, rather than learned segmentation-based subword vocabularies.

level embeddings encode. Distributional embeddings represent lexical items based on their patterns of co-occurrence in context, such that similarity in embedding space reflects shared semantic content, syntactic behaviour, and usage environments (Turney and Pantel, 2010; Levy and Goldberg, 2014). Subword models extend this principle by incorporating character-level information, which allows orthographically related forms to be represented closely even when token frequencies are low or surface forms differ (Bojanowski et al., 2017). As a consequence, embedding similarity reflects semantic relatedness, morphosyntactic similarity, and orthographic overlap. This makes clustering in embedding space a suitable operation for identifying candidate groups of lexical variants, particularly in settings where variation manifests through both form and contextual usage (Munro and Manning, 2010; Nguyen and Grieve, 2020).

3.2 Stage 1: Training Subword Embeddings

Embeddings are trained with FastText (Bojanowski et al., 2017) using the configuration values `vector_size`, `window`, `min_count`, `epochs`, `min_n`, `max_n`, `sg`. These values are estimated in accordance with the size of the corpus, as well as with some testing of the variant families (as detected in the following stage).

The input is a JSON file containing a required `text_field`. Optional fields specify the comparison dimension, such as `user_id` or `date`. The corpus is streamed to manage memory, and basic token statistics are collected. Cleaning behaviour is minimal: Mentions beginning with `@` are removed before tokenisation, lowercasing is controlled by the `lowercase` flag.

3.3 Stage 2: Identifying Variant Families

After training, a candidate lexicon V is created from all tokens that meet the `min_count` threshold. For each seed $w \in V$, the method retrieves the top neighbours based on the values `open_TOPN` or `strict_TOPN`. Cosine similarity is computed as

$$\cos(\mathbf{w}, \mathbf{v}) = \frac{\mathbf{w} \cdot \mathbf{v}}{\|\mathbf{w}\| \, \|\mathbf{v}\|}.$$

Pairs are filtered according to the associated similarity threshold (`open_TH` or `strict_TH`). Then character n-gram Jaccard overlap is computed:

$$J(w, v) = \frac{|G(w) \cap G(v)|}{|G(w) \cup G(v)|},$$

where $G(\cdot)$ contains all n-grams in the range `min_n` to `max_n`. Cosine and Jaccard values jointly determine whether two tokens belong to the same group.

We implement two modes to help identify variant families. The *open* mode forms a local star around each seed. The *strict* mode builds an undirected graph and extracts connected components. Graph growth is limited by `DEGREE_CAP`. Groups that do not reach `SNN_MIN` members are removed. For the analysis presented in subsequent sections of this paper, we used *strict* mode.

Scoring and Pruning For each group F, we compute its size and the mean values of cosine similarity and Jaccard overlap. A cohesion score is the harmonic mean of these two averages. Groups are removed if they fail to reach the minimum size or if relative frequencies exceed the bound set by `MAX_FREQ_RATIO`. All pairwise scores are retained for inspection.

Dimension-Based Aggregation If a `dimension` field is provided, the method counts in how many distinct dimensions each variant appears and records the frequency of the most common dimension. For each variant we store its coverage, its top dimension, and the share that this dimension represents of its total frequency. These values feed into filters such as `MIN_USERS` and `MAX_FREQ_RATIO`. The summary CSV lists these quantities for all groups.

Parameter	Key	Default
Lowercasing	`lowercase`	true
Comparison dimension	`dimension`	user_id
Vector size	`vector_size`	100
Context window	`window`	5
Min frequency	`min_count`	10
Epochs	`epochs`	10
Skip-gram model	`sg`	1
Character n-gram range	`min_n–max_n`	3–7
Neighbours/seed (open)	`open_TOPN`	30
Similarity thr. (open)	`open_TH`	0.75
Neighbours/seed (strict)	`strict_TOPN`	100
Similarity thr. (strict)	`strict_TH`	0.73
Min family size	`SNN_MIN`	2
Degree cap	`DEGREE_CAP`	200
Min token length	`MIN_LEN`	3
Min users per variant	`MIN_USERS`	3
Max frequency ratio	`MAX_FREQ_RATIO`	25

Table 1: Main configuration parameters.

Configuration and Output Table 1 presents an overview of the parameters used and their defaults used for the purposes of this study. The method iterates through the vocabulary, computes cosine and Jaccard scores where needed, and constructs the final groups. The output consists of a JSONL file containing all groups with their members and a CSV summary with the main statistics. As this is an experimental study, the configuration of the values is based mainly on trial and error, by checking the variant families manually after each run. In contrast to normalisation tools such as VARD (Baron and Rayson, 2008), the procedure retains surface forms and measures their similarity instead of mapping them to canonical variants.

3.4 Stage 3: Qualitative Analysis

The model outputs variant families with their cosine and Jaccard values and the frequency of each member of the group. The first step in the qualitative analysis is to manually go over the families and approach the analysis with a bottom-up method. Based on the chosen dimension (i.e. user, time, etc.), the families represent for instance user variants in a similar context or variants over time in a similar context. Then, adopting a bottom-up approach, classifying the families is a straightforward way to analyse the families based on the type of variation (orthographic, morphological, lexical, stylistic, regional, etc.). The families could also be semantically or functionally related, or the relation is not identifiable, which could be a category in itself. With the help of categorisation, the identification of patterns in the data is more feasible.

4 Luxembourgish User Comments

We demonstrate the use of our methodology in the context of Luxembourgish user comments. The comments are part of the online media platform RTL[3], the main news broadcaster in Luxembourg and the only news broadcaster completely in Luxembourgish. The comments span from 2008 to 2024 and total roughly 1,42 million comments. As the use of Luxembourgish has been expanding in the written domain in the past 25 years and formal grammar teaching in school is still not properly regulated, we observe a high amount of variation in written texts. This is especially visible in informal domains, like online user comments. An in-depth analysis of variation in this domain in Luxembourgish has not been conducted yet. However, with our methodology, we can analyse a wide number of comments and classify the occurring variation. After applying stage 1 and 2, using the users as the comparative dimension, we start with the qualitative analysis. Using a bottom-up approach, we classify the families into 7 distinct categories, which are depicted in Table 2. As the families have a different number of words and also different phenomena appearing in one family, we decided for a multi-label approach. One family can have up to 3 categories. We illustrate the frequency of each category in Table 2. In the following, we describe each category and highlight findings related to language variation in Luxembourgish.

Category	**Frequency**
Orthographic	394
Morphological	222
Lexical	115
Collocation	21
Tokenisation	14
Regional	8
Other	242

Table 2: Categories and frequency

Orthographic The orthographic category describes spellings that are not part of the official orthography (Zenter fir d’Lëtzebuerger Sprooch, 2019). This includes the use of different graphemes to express the same word, which are often based on the word’s phonological properties. Since Luxembourgish has a high phoneme-grapheme correspondence in addition to an ideology that you can write how you speak, the language presents a wide range of orthographic variation. One example is *laang* (lb. *long*) where one family includes the orthographic variant *lang*. The single vowel violates the quantity rule in the orthography (Zenter fir d’Lëtzebuerger Sprooch, 2019), as a long vowel is written doubly when more than one consonant follows. However, since the orthography is not well known, both variants appear frequently. Additionally, this category also includes families that encompass different lexemes that are spelled incorrectly. One example for this case is *krng, srng, dng, êng, öng* (lb. *none, his, yours, one, one*). The correct spelling is *keng, seng, deng, eng, eng*. In this instance, we can see two different spellings for *eng* which has two different sources. The first one *êng* represents more of a typical misspelling in comparison to *öng* which represents a phonological variant expressed with the choice of <ö>. The phonological variant of *eng* is pronounced with a rounded vowel which is then written as <ö> by some authors. This is still

[3] https://rtl.lu

#	Family	English	Standard
(1)	Zäit, Zeit, Zait, Zéit, Zaït	time	Zäit
(2)	mir, mer, mier, mär, maer, miir, mäer	we	mir
(3)	mat, matt, maat	with	mat
(4)	mecht, mécht, mescht, mëcht, mëscht	to do	mécht
(5)	wäit, weit, wait, wéit	far	wäit
(6)	sech, sëch, séch	himself	sech
(7)	Numm, Num	name	Numm
(8)	laang, lang	long	laang
(9)	Fehler, Feeler	mistake	Feeler

Table 3: Families in the category orthographic variation (One lexeme, different variants).

classified as orthographic variation since it violates the official orthography (Zenter fir d'Lëtzebuerger Sprooch, 2019).

Morphological The morphological category classifies all morphological variation. This encompasses conjugated verb forms, and inflections of nouns and adjectives in case, number, and gender. This also includes compounding nouns, clippings and conversions. One example is the family *fillen, fillt, fille* (lb. *to feel*) which includes the conjugated form *fillt* and the deletion of the final <-n> before specific characters, known as the n-rule, in *fille*.

Lexical This category includes different lexemes that are semantically related. This includes synonyms and antonyms like the family *méi, manner* (lb. *more, less*) and lexical variants like *dass, datt* (lb. *that*).

Collocation The collocation category describes families with lexical items that form conventionalised combinations in daily use. For this category, the distinct words in a family form a collocation together. For instance, *Gott, säi, Dank*[4] (lb. *god, be, thanks*) forms one family of distinct items that frequently appear in the same context. Together, they form the collocation *Gott säi Dank* meaning *thank god*.

Tokenisation The tokenisation category describes families where the same word appears twice, but in one instance without the definite article *d'* (lb. *the*) attached to it and once with it attached to the word. One example for this is the family *Zukunft, d'Zukunft* (lb. *the future*).

Regional The regional category encompasses categories where regional varieties are visible in the graphemic representation. This category overlaps with the orthographic category, as it violates the official orthography (Zenter fir d'Lëtzebuerger Sprooch, 2019). However, in this case the regional influence is visible and can be verified in the official Luxembourgish Variation Atlas (Gilles, 2021).

Other The other category includes families that have no distinct features or are not identifiable as meaningful words. This includes pragmatic expressions that are often used in online discourse like *bla, blabla, ehm, hoho, tzzz* or non-identifiable words like *fisk, ragnax, har, sed.*

4.1 Orthographic Variation: Insights

In this section, we present new insights from the orthographic category revealed by the clustering method. We examine several representative examples in more detail and discuss their impact.

One Lexeme, Different Variants We start with examples, where one lexeme has several orthographic varieties in one single family. As the families are constructed based on similar neighbouring tokens in the embedding space, this grouping is straightforward. Interestingly, often not only variants of the same lexeme are part of the family but also the correctly spelled variant. This indicates that at least for these families, the lexemes of the orthographic Luxembourgish and the non-orthographic variants of Luxembourgish are aligned in the embedding space (Cao et al., 2020). Additionally, this shows that the model is able to capture orthographic variants. In these cases, we either see multiple variants for one lexeme in one family, or only two variants, where one item is usually spelled correctly as indicated in Table 3. We observe that these categories include the most common orthographic variants in Luxembourgish.

[4] For readability, nouns in families are capitalized.

#	Family	English	Standard
(10)	déi, wéi, méi, ewéi	that, as, more, as	déi, wéi, méi, ewéi
(11)	dèi, wèi, mèi	that, as, more	déi, wéi, méi
(12)	dat, wat	this, which	dat, wat
(13)	dad, wad	this, which	dat, wat
(14)	weéi, deéi, eweéi, meéi	as, that, as, more	wéi, déi, ewéi, méi
(15)	méh, déh, wéh, ewéh	more, that, as, as	méi, déi, wéi, ewéi

Table 4: Families in the category orthographic variation (Different lexemes, one variant).

Especially in the families with only two variants, one variant is often the most common spelling variant.

Different Lexemes, One Variant The second insight for orthographic variation is the clustering of families of the same type of variation for different words across multiple families. For these families, we observe similar words, for instance *dèi, wèi, mèi* (Table 4 (11)) or *dad, wad* (Table 4 (13)) clustering together but with the same variant pattern in every word. A variant pattern is the same type of grapheme writing in different words, for example instead of the correct spelling of the diphthong <éi>, the graphemes <èi> are used consistently to represent the diphthong. Instances for these patterns are shown in Table 4. Additionally, not only variants cluster together but also the standard variants of these function words as shown in the instances (10) and (12) in Table 4. Since the instances are mostly function words, the clustering in itself is evident, however, the clustering of the identical variation pattern in different families indicates that these variants have social meaning which is represented in the embedding space. Therefore, the clustering shows that these words appear in similar topics or similar sentence structures that are linked via this variant. Another option would be a higher frequency for multiple authors writing about a similar topic. A more in-depth corpus analysis would give more insights into the social meaning of these variants.

4.2 Regional Variation: Insights

Overall, our method only made 8 instances of regional variation visible. Due to an advanced state of dialect levelling in Luxembourgish, regional dialects have evolved into a national variety with some remaining lexical, phonological, and grammatical features (Gilles, 1999). Due to the informal nature of the comments, some phonological regional features are visible in the grapheme representation of the written lexeme. As these written forms differ from the official orthography, they also classify as orthographic variants. There are also instances of regional variants that are not orthographic variants but are part of the official dictionary for Luxembourgish[5] like *mar* which is a regional variant of *muer* (lb. *tomorrow*). With the Variation Atlas (Gilles, 2021) we can verify specific variants that are part of the 811 variant maps included in the atlas.

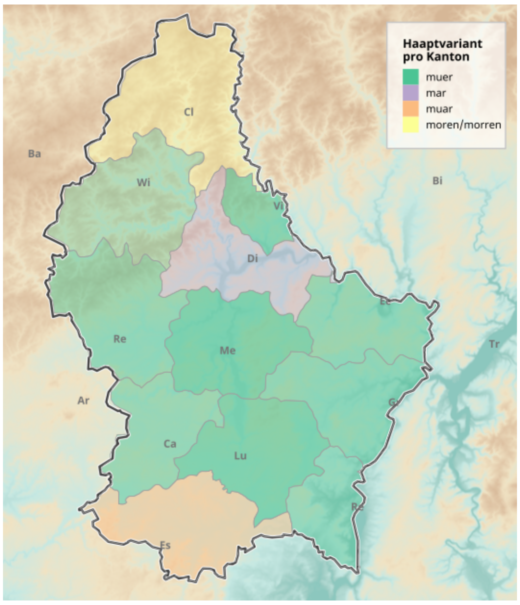

Figure 1: Map of *muer* variants (Gilles, 2021)

One of those variants is the family *muer, muar, moar* (lb. *tomorrow*). Figure 1 shows the map of the variant *muer* and the regional variants *mar, muar, moren*. Not shown in the map due to a low frequency of instances is the variant *moar*, which is also part of the family in our analysis. The variant *muar* is prevalent in the south of Luxembourg, whereas *muer* is the most common variant in Lux-

[5] https://lod.lu

embourg and considered the standard variant[6]. The variant *moar* is mostly localised in the east of the country, but only a few participants (19) of the variation atlas survey have used that variant. However, in the comments 338 instances are recorded in comparison to 604 instances of *muar*, and 6577 instances of *muer*. This shows that *moar* is still the least used variant in contrast to the most frequent variant *muer*. However, it is still a common variant used in the comments, which was not clearly recorded before as the findings of the Variation Atlas did not indicate this. Additionally, the Jaccard values show a low overlap between authors using these variants, which indicates a consistent use of a variant instead of switching between variants in different contexts.

4.3 Lexical Variation: Insights

The lexical variation category is one of the most heterogeneous categories in our analysis, as this category includes every family that encompasses different lexemes which are semantically related to each other. In addition to synonyms and antonyms, we also found common lexical variants in the data that are included in the Variation Atlas (Gilles, 2021).

Two interesting families are *séier, schnell, seier* (lb. *fast*) and *säit, seit, sait, zanter, zenter, zënter, séit* (lb. *since*). These families do not only show lexical variation, but also orthographic variation. In this section, we focus on the lexical side. Figure 2 illustrates the regional distribution of the use of *séier* and *schnell*. Overall, the frequency of use of both variants is nearly identical, at an almost equal split. The variant map illustrates some preference for the *séier* variant in the north of the country, and some for *schnell* in the south. However, the statistical analysis shows that only the north is significantly favouring the *séier* variant (Gilles, 2021). One factor that influences the use of this variant significantly is age. The older the participants of the variation atlas survey are, the more the variant *séier* is preferred.

Similarly, we can observe these tendencies with *säit*, *zënter* and *zanter*. Gilles (2021) illustrates in the Variation Atlas that the variant *säit* is overall favoured and regionally the most used variant. However, if we look at the factor age, the variant *zënter* and *zanter* are more popular with an older age group. It needs to be noted that *zënter* and *zanter* are phonological variants but belong the same lexical variant.

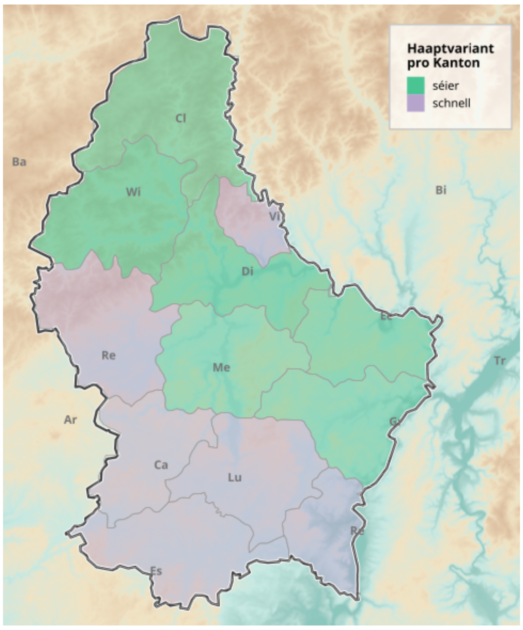

Figure 2: Map of *schnell* variants (Gilles, 2021)

Further, we also found families that are topically related. The family *chat, chatgpt, gpt, chatbot* illustrates the more recent advances in AI and how these entered into the discourse in the user comments. Another example is the family *Grexit, Frexit, Brexiteers, Lexit, Nexit, Luxexit* which encompasses word formations inspired by the expression *Brexit*. By combining country names with the word *exit* (e.g. *France* = *Frexit*, *Netherlands* = *Nexit*, *Luxembourg* = *Luxexit*) we get a family of variants that is topically related to a withdrawal from the European Union for different countries. Interestingly, different countries cluster together, indicating a similar discourse on the topic for different countries.

5 Discussion

Previous work has demonstrated that specific lexical and orthographic variants tend to cluster in distributional space. Studies such as Nguyen and Grieve's (2020) show that subword-based embeddings encode systematic spelling variation, and related work in dialectology and sociolinguistics has used clustering to analyse similarity between known varieties or regional forms. Crucially, much of this research starts from predefined units, such as known variants, dialect labels, or geographic groupings, and then examines how these cluster together. The analytical focus is typically on confirming that related forms or varieties occupy nearby positions in the embedding space.

[6] Verified via the official dictionary for Luxembourgish.

The present study reverses this perspective. Instead of beginning with predefined variants, it detects and mines variants directly from co-occurrence and similarity patterns in the data. Clustering is thus used not as a validation tool, but as an exploratory mechanism that makes candidate forms visible without prior assumptions.

This methodological change is particularly relevant for Luxembourgish. Written Luxembourgish exhibits high orthographic diversity, partial standardisation, frequent borrowing, and code-switching, which make manual enumeration of variants difficult and incomplete. These properties create favourable conditions for unsupervised discovery of variation. This also applies to the dataset at the centre of our analysis. User comments provide dense, repetitive, and informal language use across many contributors, which supports distributional modelling of variation while retaining socially meaningful diversity.

Beyond identifying variant families, the approach opens up several directions for further analysis. One avenue concerns the role of different dimensions in shaping cluster structure. Future work will examine which types of variation cluster along which dimensions, for example whether certain families are associated with particular groups of users or whether historical and more recent forms of Luxembourgish can be distinguished through their distributional profiles. Exploring these questions will help clarify how different sources of variation interact and which analytical perspectives are best supported by this methodology.

The resulting clusters reveal patterns that are well aligned with linguistic intuition, but are not explicitly encoded in existing resources. This suggests that substantial lexical and orthographic variation remains undocumented. Resources such as the one analysed here therefore warrant further investigation, both to enrich descriptive accounts of Luxembourgish and to inform the development of computational models that better reflect actual language use.

Building on these considerations, we emphasise that the full codebase is released publicly to encourage reuse, scrutiny, and extension by other researchers. At the same time, our findings and limited further testing suggest that the proposed approach is not suited to highly standardised languages, where orthographic variation is limited and many relevant distinctions are already captured by existing lexical resources and normalisation pipelines. Its strengths instead lie in settings characterised by non-standardisation, activate variation, or incomplete codification, where spelling diversity encodes sociolinguistic and contextual information rather than noise. We therefore hope that this work enables and motivates applications to similar language varieties and research situations, including under-resourced or emerging standards. Beyond language-specific use cases, the method is also applicable to other domains and previously unexplored corpora, such as large web crawls or collections with limited metadata. In this sense, the contribution is less a universal solution for all languages than a transferable framework for studying variation in contexts where standard assumptions do not hold.

6 Conclusion

In this paper, we have made two main contributions. First, we describe a transparent and reproducible methodology for inducing lexical and orthographic variation directly from raw text, without relying on predefined variant lists or normalisation. Second, we present a large-scale empirical analysis of Luxembourgish user-generated text that documents variation as it is used in practice. Across seven analytically defined categories, the method identifies around 800 variant families, revealing systematic patterns of spelling and lexical diversity. The findings confirm earlier observations that related variants cluster in distributional space (Nguyen and Grieve, 2020), while extending this insight by showing how such clusters can be mined directly from data rather than used only for validation.

Looking ahead, we plan to apply this methodology to additional Luxembourgish corpora in order to compare domains and writing contexts. This includes exploring whether user-level variation patterns can be characterised more systematically and assessing how well different available corpora reflect everyday language use. At the same time, initial experiments suggest that the approach is less effective for highly standardised languages, where orthographic variation is limited and distributional signals are weaker. This highlights that the method is particularly suited to languages and domains with active variation, and that its applicability depends on the sociolinguistic properties of the data.

Limitations

The findings in this study need to be interpreted with care. The corpus consists of user comments from a single online platform, which represents only a small portion of the Luxembourgish-speaking population. Patterns observed in this dataset therefore do not necessarily generalise to the wider speech community, nor do they capture the full range of regional, social, or stylistic variation present in Luxembourgish. The method identifies orthographic and lexical families based on distributional and subword similarity, which makes it sensitive to corpus composition and frequency effects. Rare variants may be missed, and high-frequency items can dominate neighbourhood structures. While the induced families provide useful candidates for analysis, their linguistic validity still depends on qualitative assessment. The results should thus be seen as a structured starting point for investigating Luxembourgish variation rather than a comprehensive account of the language.

A further limitation concerns the interpretation of the induced clusters. While the method identifies groups of closely related forms, it cannot by itself determine whether these patterns reflect linguistic variation, author-specific preferences, or temporal effects. In practice, these sources of variation are often intertwined in user-generated text, and distributional similarity alone does not allow them to be disentangled with certainty. Although dimension-based aggregation provides partial insight into how variants are distributed across users or time periods, the clustering process itself is agnostic to the underlying cause of similarity. As a result, the identified families should be interpreted as candidates for linguistic variation that require contextual and qualitative analysis to establish their nature.

Ethical Considerations

This study uses publicly accessible user comments, but they remain sensitive textual data. All processing follows the terms of use of the platform from which the comments were collected. No attempt is made to identify individual users, and the analysis relies only on aggregated patterns such as variant frequencies and distribution across dimensions. Even though user identifiers are present in the raw data, they are treated only as categorical variables and are not (and could not be) interpreted as personal attributes. The dataset represents a self-selected set of online participants whose linguistic behaviour may differ from that of the wider population, and care should be taken not to attribute group-level characteristics to individual users. Finally, automatically induced variant families can reflect social or regional differences, but these patterns should be interpreted with caution to avoid reifying stereotypes or overgeneralising from limited data. The methodological framework is intended for linguistic analysis rather than profiling or prediction of individuals.

Acknowledgments

This research was supported by the Luxembourg National Research Fund (Project code: C22/SC/117225699). The experiments reported in this paper were conducted on the MeluXina high-performance computing infrastructure, made available through an allocation granted by the University of Luxembourg on the EuroHPC supercomputer hosted by LuxProvide. We would like to thank the TRAVOLTA project partners for their invaluable advice and guidance on this project.

References

Martine Adda-Decker, Thomas Pellegrini, Eric Bilinski, and Gilles Adda. 2008. Developments of "Lëtzebuergesch" Resources for Automatic Speech Processing and Linguistic Studies. In *Proceedings of LREC*.

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a Better Understanding of Noise in Natural Language Processing. In *Proceedings of RANLP*.

Dimitra Anastasiou. 2022. ENRICH4ALL: A First Luxembourgish BERT Model for a Multilingual Chatbot. In *Proceedings of SIGUL*. ELRA.

Alistair Baron and Paul Rayson. 2008. VARD2: A tool for dealing with spelling variation in historical corpora.

Laura Bernardy. 2022. A Luxembourgish GPT-2 Approach Based on Transfer Learning. Master's thesis, University of Trier.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. In *Transactions of ACL*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proceedings of ICLR*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*.

Nathalie Entringer, Peter Gilles, Sara Martin, and Christoph Purschke. 2021. Schnëssen. surveying language dynamics in luxembourgish with a mobile research app. *Linguistics Vanguard*, 7.

Pieter Fivez, Simon Šuster, and Walter Daelemans. 2017. Unsupervised Context-Sensitive Spelling Correction of English and Dutch Clinical Free-Text with Word and Character N-Gram Embeddings. *Computational Linguistics in the Netherlands Journal*.

Daniela Gierschek. 2022. *Detection of Sentiment in Luxembourgish User Comments*. Ph.D. thesis, University of Luxembourg.

Peter Gilles. 1999. *Dialektausgleich im Lëtzebuergeschen: Zur phonetisch-phonologischen Fokussierung einer Nationalsprache*. Niemeyer, Tübingen, Germany.

Peter Gilles. 2019. 39. Komplexe Überdachung II: Luxemburg. Die Genese Einer Neuen Nationalsprache. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum - Ein internationales Handbuch der Sprachvariation. Volume 4 Deutsch*. De Gruyter Mouton, Berlin, Boston.

Peter Gilles. 2021. Variatiounsatlas vum lëtzebuergeschen. `https://infolux.uni.lu/variatiounsatlas`. Accessed: 15.10.2025.

Peter Gilles, Léopold Edem Ayité Hillah, and Nina Hosseini Kivanani. 2023a. ASRLUX: Automatic Speech Recognition for the Low-Resource Language Luxembourgish. In *Proceedings of the International Congress of Phonetic Sciences*.

Peter Gilles, Nina Hosseini Kivanani, and Léopold Edem Ayité Hillah. 2023b. LUX-ASR: Building an ASR system for the Luxembourgish language. In *Proceedings of IEEE Spoken Language Technology Workshop*.

Jack Grieve, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in AI*, 2.

Hang Jiang, Haoshen Hong, Yuxing Chen, and Vivek Kulkarni. 2020. DialectGram: Automatic Detection of Dialectal Changes with Multi-geographic Resolution Analysis. In *Proceedings of the Society for Computation in Linguistics*.

Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. 2025. False Friends Are Not Foes: Investigating Vocabulary Overlap in Multilingual Language Models. *arXiv preprint*. ArXiv:2509.18750 [cs].

Thomas Lavergne, Gilles Adda, Martine Adda-Decker, and Lori Lamel. 2014. Automatic language identity tagging on word and sentence-level in multilingual text sources: a case-study on Luxembourgish. In *Proceedings of LREC*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of NeurIPS*.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish. In *Proceedings of LREC*.

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural text normalization for Luxembourgish using real-life variation data. In *Proceedings of VarDial (COLING)*.

Robert Munro and Christopher D. Manning. 2010. Subword Variation in Text Message Classification. In *Proceedings of NAACL-HLT*.

Dong Nguyen and Jack Grieve. 2020. Do Word Embeddings Capture Spelling Variation? In *Proceedings of ICCL*.

Nhi Pham, Lachlan Pham, and Adam Meyers. 2024. Towards Better Inclusivity: A Diverse Tweet Corpus of English Varieties. In *Proceedings of LAW (ACL)*.

Fred Philippy, Shohreh Haddadan, and Siwen Guo. 2024. Forget NLI, Use a Dictionary: Zero-Shot Topic Classification for Low-Resource Languages with Application to Luxembourgish. In *Proceedings of SIGUL (LREC-COLING)*.

Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. LuxBank: The first Universal Dependency treebank for Luxembourgish. In *Proceedings of TLT*.

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for Luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of VarDial (COLING)*.

Christoph Purschke. 2020. Attitudes Toward Multilingualism in Luxembourg. A Comparative Analysis of Online News Comments and Crowdsourced Questionnaire Data. *Frontiers in AI*, 3.

Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. Publish or Hold? Automatic Comment Moderation in Luxembourgish News Articles. In *Proceedings of RANLP*.

Joshgun Sirajzade, Daniela Gierschek, and Christoph Schommer. 2020. An Annotation Framework for Luxembourgish Sentiment Analysis. In *Proceedings of SLTU-CCURL (LREC)*.

Natalie D. Snoeren, Martine Adda-Decker, and Gilles Adda. 2010. The Study of Writing Variants in an Under-resourced Language: Some Evidence from Mobile N-Deletion in Luxembourgish. In *Proceedings of LREC*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37.

Zenter fir d'Lëtzebuerger Sprooch, editor. 2019. *D'Lëtzebuerger Orthografie*. Zenter fir d'Lëtzebuerger Sprooch, Stroossen.