University of Otago

School of Computing

COSC385 Project Report

Talking in French Like Academia

Machine Learning Powered Verlan Identification

Author:
Yitian Li (4556502)

Supervisor(s):
Dr Lech Szymanski
Dr Veronica Liesaputra

October 14, 2025

Abstract

something.

1 Introduction

1.1 Context and Motivation

Since the early 19th century, the French people have started to talk using verlan. Just like Pig Latin[1] exists in English culture, verlan is an unusual and creative form of *argot* (slang) that is formed by flipping the syllables around in a word.[2][1, 2] Time flies, verlan has become more and more popular, and it is now widely used amongst teens and young people in francophone societies[3][3]. Examples of verlan can be as follows:

- bite = bi + te → te + bi → tebie (penis)
- shit = shi + t → t + shi → teuchi[3]
- bonjour = bon + jour → jour + bon → jourbon (greetings)

In real-life conversations, such can be used as in the example sentences below:

- *Le graff géant représente une tebie pixel art.*
  (The giant graffiti depicts a pixel art penis.)
- *Il a du bon teuchi du bled.*
  (He's got some good shit from the countryside.)
- *Un p'tit*[4]*jourbon et tout le monde sourit.*
  (A quick hello and everyone smiles.)

Indeed, verlan can be formed with different original languages, not only French, but also English and other languages. However, it always follows the same rule of flipping syllables, although, for better pronunciation reasons, certain minor amendments such as dropping unnecessary letters and applying accents (e.g., é, è) can be used from time to time[1]. Besides, due to the

[1] en.wikipedia.org/wiki/Pig_Latin

[2] In fact, the word *verlan* is a verlan from the word *l'inver* (the inversion).

[3] Such as France, Belgium, Switzerland, Luxembourg, and Canada.

[4] Standard spelling: petit.

universal trait of slang being used more often phonetically instead of written, verlan users tend to spell them differently when writing them down. As technology develops, this has been occurring more frequently than ever in daily texting[4].

Thinking internationally, when people are communicating with translators, it is possible that slang in their mother language can be brought to the conversation, which could be tricky for translators to translate[5]. Using translators such as DeepL[5] and Google Translate[6] to translate sentences that contain verlan from French to English can be a specific example to prove this. Furthermore, although both of the translators above are using Machine Learning (ML) for translation, their results of translating verlans are not ideal[6, 7]. For example, when attempting to translate the sentence above, *Le graff géant représente une tebie pixel art.*, both Google Translate1 and DeepL2 cannot translate the word *tebie* correctly. Specifically, for DeepL, there is no desired translation as *penis* in its alternative word list for *tebie*3.

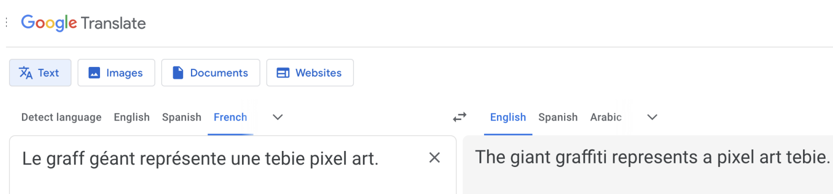


Figure 1: Google Translate cannot translate the verlan *tebie* correctly.

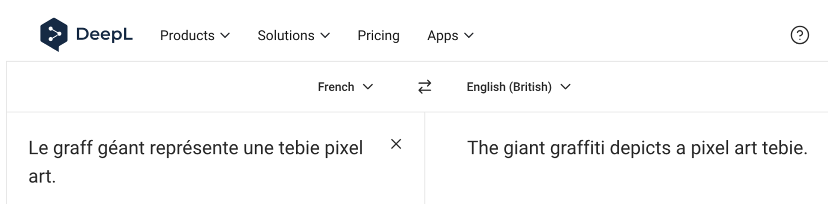


Figure 2: DeepL cannot translate the verlan *tebie* correctly.

[5] www.deepl.com
[6] translate.google.com

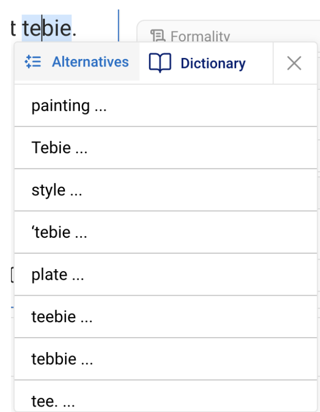


Figure 3: No desired translation for verlan *tebie* in DeepL's alternative word list.

Thus, a question shall naturally arise: Can we improve translators' performance in translating slang by improving the ML model? The answer is undoubtedly 'yes' in an era where artificial intelligence research is expanding rapidly. Researchers have been making progress in identifying slang using ML[13] and, moreover, in translating noisy text, of which slang is a part[8].

But what about verlan? There is no known ongoing or completed research on identifying *such* slang or their translations[7], nor does a proper dataset exist. The only work similar to this is an assignment published at the University of Toronto[8], asking students to train a Neural Machine Translation (NMT) model to transform standard English into Pig Latin. It is not only the other way around; instead of identifying Pig Latin and transforming it back to standard English, it is also more of an example for students to practice using NMT than a discussion on its identification and translation. Shouldn't we do something?

This report aims to change that.

1.2 Objective

The purpose of the project is to create two verlan datasets: one functioning as a dictionary, containing the verlan words and their normalised standard French equivalents; the other a dataset of sentences that contain verlan,

[7]Until September 2025.

[8]https://uoft-csc413.github.io/2022/assets/assignments/PA03.pdf

paired with the same sentences containing normalised words, with labels indicating whether a sentence contains verlan. After that, the project embeds and classifies verlan using Large Language Models (LLMs) and analyses the results.

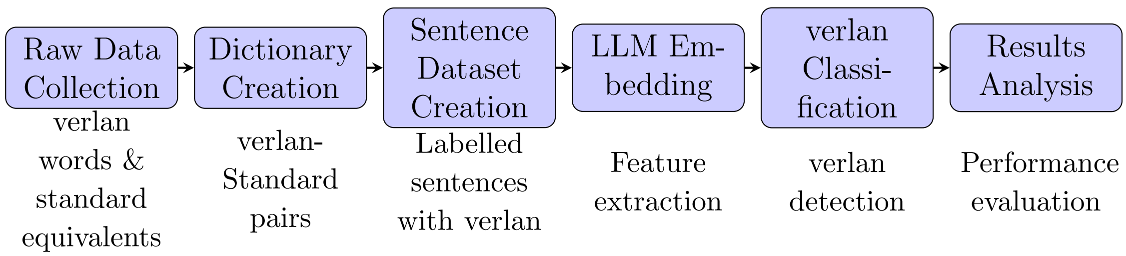


Figure 4: A visulisation of the objectives.

With the purpose above, the report contributes to the linguistics and the AI researchers two verlan datasets, for dictionary making or LLMs training. The report also evaluates how good we can achieve the identification of verlan with ML, to benefit machine translation in the future.

The code and the unannotated, un peer-reviewed dataset developed as part of the project are released under openlicences and aligns with open science best practices, with the usage of a version controlled software development platform (GitHub)[9]. The annotated, peer-reviewed dataset will be published shortly after this report, aiming by the end of 2025.

2 Background

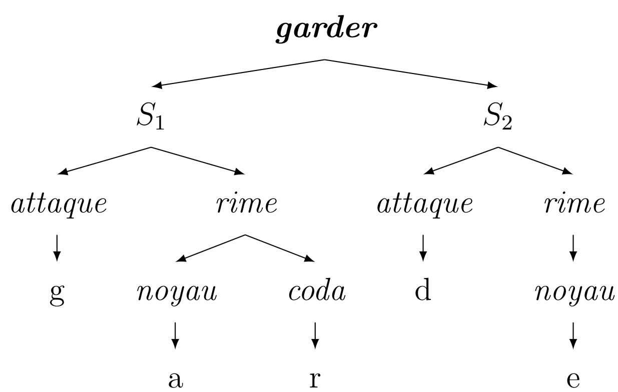
2.1 A Living Verlan

Vivienne Véla, a former scholar from Université Paris 8, poetically captured one of Verlan's most important traits: it pursues confusion instead of clarity[21]. One reason is that it is widely used among lower-class people, drug users, gangs, or those in jail. Thus, making the context unidentifiable is important — certain phenomena such as reverlanisation (flipping the Verlan again if it becomes too popular) and truncation are therefore applied.

[9] github.com/greateden/verlan-Identification-Normalisation

However, although Verlan is used for concealing meaning, it still follows certain rules. The most general rule is syllabic reversal, as mentioned in the introduction chapter of this report.

Specifically, to delve into the linguistic rules, Véla pointed out that the analytic model proposed by Kaye and Lowenstamm provides the best description[22]. The syllable can be disassembled into *attaque* (onset), *rime* (rhyme), *noyau* (nucleus), and *coda*. For example, here is a representation of the word *garder*, IPA[10] [garde].

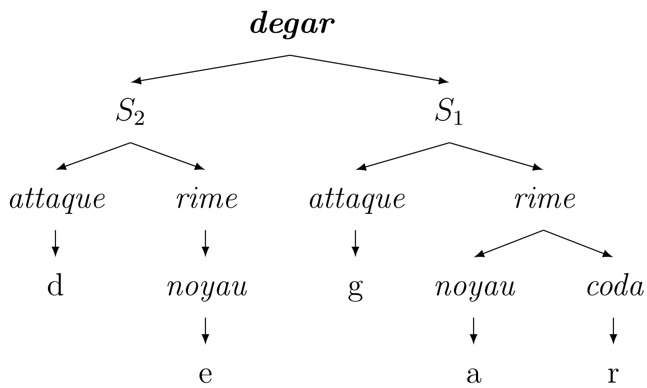


It has two syllables, S_1 and S_2. To create the Verlan form, we follow the permutation equation below:

$$(S_1 S_2) \rightarrow (S_2 S_1) \tag{1}$$

After the permutation, we obtain the Verlan form of *garder* as *degar*, represented below.

[10] International Phonetic Alphabet, https://en.wikipedia.org/wiki/International_Phonetic_Alphabet



Notably, the permutation occurs only at the syllable level (i.e., between S_1 and S_2); it does not affect the internal structure of each syllable tree, although in some cases, certain letters (such as *e*) might be dropped after permutation. That said, the example above is not an exhaustive explanation of forming a Verlan. To avoid confusing the readers, this report suggests that this example perfectly illustrates its regular rule. For further details, readers are advised to consult Véla's paper.

With such a sub-word permutation, researchers can not only discuss it within the linguistic realm, but it is also intriguing for computer scientists to explore how machines, such as LLMs, perceive this kind of difference. Just as Véla describes Verlan — ambiguous, sometimes violent, sometimes amazing, and always vivid.

2.2 Detecting Slang

To the best of our knowledge, there is no existing computational research[11] on the *detection* of Verlan — this particular form of French slang. However, there are a few scholars who have included Verlan in their research[9, 10, 11, 12]. Yet, these studies commonly included Verlan as a type of slang in their datasets or corpora. Moreover, they did not specifically focus on how to detect this particular type of slang, but rather approached it in a broader sense — they created slang datasets that contain Verlan, and some of them employed computational approaches to detect such slang.

Fortunately, there are several papers related to computational slang detection, and their approaches could contribute to Verlan detection to a large

[11] As of September 2025.

extent[13, 14, 15, 18]. These studies are not limited to French but also cover other Indo-European languages[12].

Therefore, regarding the history of Verlan detection, this report first generalises the task as slang detection, and then discusses possible methods that could be implemented for Verlan identification, in order to provide readers with a general and useful background.

2.2.1 1910s-2016: A Super-Condensed History of Slang Detection

The background of traditional slang detection often leverages fuzzy-matching methods. Two main methods were introduced and widely cited: Soundex, a phonetic indexing system for names introduced by Russell in 1918, and the edit-distance-based spelling-correction method introduced by Levenshtein in 1966[23, 24]. Afterwards, scholars introduced more algorithms, such as Philips's Metaphone and Double Metaphone, which improved on Russell's Soundex; Kukich's methods for detecting and correcting spelling; Sproat's normalisation of Non-Standard Words (NSW); and Aw et al.'s phrase-based Machine Translation (MT) approach for standardising SMS messages[25, 26, 27, 28, 29]. While these are not directly slang-detection research, over time their methodology became increasingly related to slang — some slang can be treated as misspelling or NSW, and people frequently use slang in text messages.

2.2.2 2016-2019: Dictionary Search

The easiest way we can think of dealing with slang is to use a dictionary — just like how we look up a word that we do not know. The pros and cons are highly similar to consulting a dictionary. It is fast (if using a digital one) and accurate. On the other hand, because it is purely fixed data, it only works with existing words and thus cannot identify newly invented ones.

Examples of existing slang dictionaries include SlangNet, SlangSD, and SLANGZY[17, 18, 19]. As for French slang dictionaries, we have, for example, *Dictionnaire du chilleur*[20]. Specifically for Verlan, the report identifies several online dictionaries, including *Dictionnaire Interactif du Verlan*[13],

[12] For example, English, German, and Russian. For more information, please refer to: https://en.wikipedia.org/wiki/Indo-European_languages.

[13] https://ecoleng.com/verlan-comprendre-argot-francais-parler/dictionnaire-interactif-du-verlan

Wiktionary[14], and *Dictionnaire Verlan*[15].

With these existing dictionaries, implementing a tool to identify Verlan should be straightforward. However, two major issues limit the possibility of directly using these dictionaries for Verlan identification: they lack comprehensive coverage, and some are fan-made, which neither captures the full extent of this slang nor guarantees accuracy. Licensing for certain dictionaries could also be a concern.

Although dictionaries have the drawbacks mentioned above, they remain essential resources for implementing LLM-based approaches, as discussed later. Consequently, new dictionaries continue to be produced.

2.2.3 Meanwhile, for Fuzzy Search

The 2010s belonged to social media and research on user-generated text. Representative work includes Beaufort et al.'s hybrid finite-state framework for SMS normalisation, Han and Baldwin's lexical normalisation for Twitter, and the W-NUT shared tasks on Twitter message normalisation[30, 31, 32].

While these works are not directly about slang recognition, they provided immensely useful background for the research specifically on slang that followed.

2.2.4 2020-2025: Fuzzy Search + Slang Corpus = BOOM

In the 2020s, everyone tended to check what could be done with Machine Learning (ML) for this task, using Natural Language Processing (NLP). Researchers started to apply NLP to slang detection. Wilson's paper used two million entries from *Urban Dictionary*, with terms, definitions, examples, and tags[33]. They pre-processed the dataset with techniques like lowercasing and removal of punctuation, followed by training a fastText[16] skip-gram for 10 epochs with a 300-dimensional vector space. Using a fastText classifier, they analysed properties such as sentiment and sarcasm. For evaluation, they used accuracy, precision, recall, and F1 score.

Notably, the report has found two theses highly related to this project, *Slang or not?* and *Toward Informal Language Processing*[14, 15]. Both cre-

[14] https://en.wiktionary.org/wiki/Category%3AVerlan

[15] https://zlang.fandom.com/fr/wiki/Dictionnaire_Verlan

[16] A library for learning of word embeddings and text classification created by Facebook's AI Research (FAIR) lab.

ated their own datasets that were manually annotated and validated. The former compared the performance of traditional ML (SVM[17]-linear with TF-IDF[18] + n-grams), Convolutional Neural Network (CNN)[19] / Bidirectional Long-Short Term Memory (BiLSTM)[20] with Bidirectional Encoder Representations from Transformers (BERT)[21] embeddings, Transformer models (e.g., BERT-large-uncased), and Large Language Models (LLMs) (GPT-4o and GPT-4o-mini), finding that a fine-tuned Transformer performed best. The latter compared traditional baselines, Language Models (LMs), and LLMs.

2.2.5 Detecting Verlan?

The results from the last section provide this report with a clear guideline regarding Verlan identification. They have absorbed and adapted the historical development of slang detection into a modern, up-to-date framework. Building upon these insights, this report argues that BERT and contemporary LLMs represent the most effective tools for the Verlan detection task.

3 Datasets

3.1 The Separated Structures

As of the time of writing, there are no published Verlan datasets. Thus, this report has created two datasets: one is a lookup table mapping words in Verlan to their standard French forms, named *GazetteerEntries* (hereafter *the dictionary*); the other contains example sentences for the words appearing in the table, both in Verlan and in standard French, with three entries per form, named *Sentences*. The general reasons for having two datasets are:

1. To separate rules and learning signals. The dictionary works as a lookup and a baseline for rule-matching: it provides word mappings, word variants, etc., whilst the sentences dataset is for detection and

[17]Support Vector Machine, https://en.wikipedia.org/wiki/Support_vector_machine

[18]https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[19]https://en.wikipedia.org/wiki/Convolutional_neural_network

[20]https://en.wikipedia.org/wiki/Long_short-term_memory

[21]https://en.wikipedia.org/wiki/BERT_(language_model)

evaluation and illustrates *how* verlan appears in context. If mixed together, the model will not be able to distinguish tokens as dictionary knowledge or usage.

2. To improve reusability. The dictionary can be used independently on any corpus for rule-based verification, while the sentences dataset can be updated separately to add more community examples without modifying the dictionary, supporting modularisation of the pipeline.

3. For a cleaner evaluation. The dictionary can serve as a baseline while the sentences dataset can be split for training and testing, making results easier to interpret.

Generally speaking, the separation of the datasets can potentially make the model and the experiments clearer, explainable, and easy to extend. They could also contribute to LLM training and corpus creation in the future.

3.2 Visualisation of the Datasets

Figure 5 presents the attributes in the datasets and highlights how they relate to each other.

3.3 The Creations

As mentioned in Section 2.2.2, this report first checked and scraped sources that were available and had researcher-friendly copyright policies. Among those mentioned, *Dictionnaire Verlan* and Wiktionary contributed the most in terms of quantity. However, as they are not curated or officially published, their quality is not guaranteed. Moreover, many entries do not provide example sentences, which makes the creation of the sentences corpus harder.

3.3.1 Sampling

Although there is no clear estimate of the overall quantity of Verlan, after searching, scraping, and combining, this report compiled a total of 1,086 Verlan items, though some are merely spelling variants of the same word. For example, *foncédé* and *foncedé*[22] are counted separately as two entries;

[22]Verlan of *défoncé*, often translated as *high (on drugs)* in English.

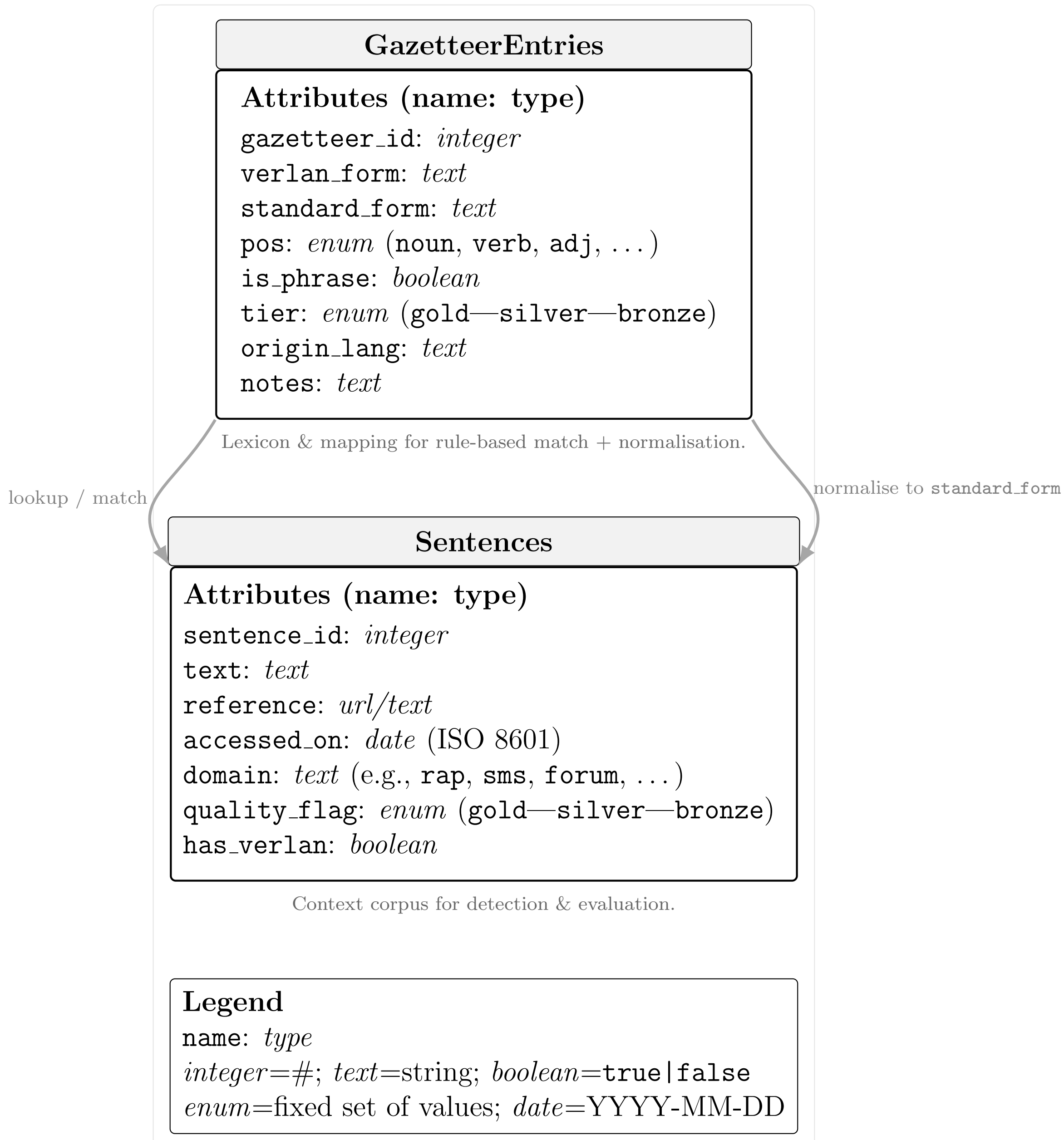


Figure 5: Overview of the GazetteerEntries lookup table and the Sentences corpus, including their key attributes.

so are *keus* and *keuss*[23]. Notably, there are around 150 entries for which the report did not find their standard form; thus they have been categorised as *bronze* regarding their quality. To the best of our knowledge, the dictionary we have created contains the largest number of Verlan entries among the public dictionaries we could find.

After creating the dictionary, the report searched for and scraped usage examples online to create the sentences corpus. The report also used Artificial Intelligence (AI) tools — specifically, ChatGPT Deep Research[24] — for sentence scraping. The results for sentences with a verifiable reference have been marked as *gold* quality; those without a verifiable reference have been marked as *silver* quality. For items for which the report could not find example sentences, we prompted ChatGPT-o3[25] to generate example sentences; these results have been marked as *bronze* quality. All results have been reviewed by the author of this report and are intended to undergo annotation in the future.

3.3.2 Quality Tiers

To provide readers with a clearer understanding of the tier/quality schema introduced in this report, we have created a table for clarity:

Table 1: Quality tiers of the Verlan datasets.

Tier	Definition	Source
Gold	Verified with public reference	Public reference (URL/citation)
Silver	Plausible sentence without verifiable source	Scraped / semi-auto
Bronze	LLM-generated and manually reviewed	ChatGPT-o3

3.4 Final Dataset

At the time of writing, the datasets are not yet officially finalised. Although the structure of the two datasets is as shown in Section 3.2, in the dictionary dataset this report did not invest much effort in annotating the original language of the Verlan items; the *note* column is also scarcely used. In

[23]Verlan of *sec*, translated as *dry* in English.
[24]https://openai.com/index/introducing-deep-research/
[25]https://en.wikipedia.org/wiki/OpenAI_o3

the sentences corpus, the *accessed_on* and *domain* columns are also scarcely annotated.

The main reason is that these columns were not used in the implementations. In fact, this report only used *gazetteer_id*, *verlan_form*, and *standard_form* in the dictionary, and *sentence_id*, *text*, and *has_verlan* in the sentences corpus. To us, the remaining columns are primarily for publishing the datasets and for potential advanced experiments in the future.

4 Building the Pipelines — Model Architectures and Specifications

4.1 Zero-Shot Models

4.2 Training Models

4.2.1 General Pipeline

4.2.2 The Usage of the Datasets

4.2.3 Environment and Hyperparameters

4.3 Other Models

4.3.1 Justifying for LR

4.3.2 Gazetteer Gate

4.3.3 Calibration: Temp / Platt / Isotonic / Threshold Tuning

5 Experiments, Results, and Analyses

5.1 Evaluation Methodology

5.1.1 Testing Datasets

5.1.2 Testing Schema

5.2 Results

Draft note. Tables 3 and 2 sketch the raw confusion counts gathered from the six shortlisted systems. The learned verlan detectors (especially the end-to-end BERT variant) reach the best overall balance on the main held-out split, whereas both zero-shot models remain brittle on slang and invented verlan. Invented forms are the hardest slice for every model, with even the strongest system missing more than half of the positives.

Model	Existed verlan (TP/FN)		Invented verlan (TP/FN)		French slang (TN/FP)	
	TP	FN	TP	FN	TN	FP
Frozen+LR	20	9	8	17	18	7
E2E+LR	19	10	6	19	16	9
Frozen+BERT	24	5	12	13	20	5
E2E+BERT	19	10	4	21	19	6
Mistral-7B Zero Shot	22	7	17	8	11	14
ChatGPT 5 Thinking Zero Shot	23	6	23	2	20	5

Table 2: Draft confusion-count summary on the auxiliary targeted suites: curated historical verlan pairs (29 samples), self-created verlan forms (25 samples), and contemporary slang paraphrases (25 samples). Only verlan variants are shown for the verlan suites, so TN/FP are omitted there. ChatGPT operates on 158 prompt responses in total across both tables.

5.3 Analyses

5.3.1 Zero-Shot Mistral-7B Model

5.3.2 Frozen Neurons Embedding with Linear Regression Head as Classifier

5.3.3 Frozen Neurons Embedding with BERT Head as Classifier

5.3.4 Full Fine-Tune Neurons with Linear Regression Head as Classifier

5.3.5 Full Fine-Tune Neurons with BERT Head as Classifier

5.3.6 Zero-Shot ChatGPT 5 Codex (High)

5.4 Conclusion and Limitation

6 Discussion and Outlook

References

[1] Radjabov, Ruslan Rajabmurodovich. *Understanding "verlan" in the French Language*. Web of Scientist: International Scientific Research Journal, vol. 6, no. 3, 2025, pp. 368-372. Available at: https://webofjournals.com/index.php/3/article/view/3264.

[2] Bach, Xavier. *Tracing the origins of verlan in an early nineteenth century text*. Journal of French Language Studies, vol. 28, no. 1, 2018, pp. 1-18. Cambridge University Press. doi:10.1017/S0959269516000221.

[3] Olivier Sécardin. *Évolution du verlan, marqueur social et identitaire, comme reflet de la langue et de la société françaises*. Synergies Europe, no. 3, 2008, pp. 223-232. Available at: https://journal.lib.uoguelph.ca/index.php/synergies/article/download/1037/1859?inline=1.

[4] Rúa, Paula López. "Shortening Devices in Text Messaging." *Journal of Computer-Mediated Communication*, vol. 10, no. 4, July 2005. Wiley. doi:10.1111/j.1083-6101.2005.tb00268.x.

[5] Hajiyeva, Bulbul. "Translating Idioms and Slang: Problems, Strategies, and Cultural Implications." Acta Globalis Humanitatis et Linguarum, vol. 2, no. 2, 2025, pp. 284-293. doi:10.69760/aghel.025002123.

[6] DeepL. "DeepL Translator translates texts using artificial neural networks. These networks are trained on many millions of translated texts." *DeepL Blog*, 2020. Available at: https://www.deepl.com/en/blog/how-does-deepl-work.

[7] Wu, Yonghui, et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." arXiv preprint arXiv:1609.08144, 2016. Available at: https://arxiv.org/abs/1609.08144.

[8] Michel, Paul, and Graham Neubig. "MTNT: A Testbed for Machine Translation of Noisy Text." *Proceedings of EMNLP*, 2018. Available at: https://aclanthology.org/D18-1050/.

[9] Zurbuchen, Lucas, and Rob Voigt. *A Computational Analysis and Exploration of Linguistic Borrowings in French Rap Lyrics*. In *Proceedings

of the 62nd Annual Meeting of the Association for Computational Linguistics — Student Research Workshop (ACL SRW 2024)*, 2024, pp. 200-208. DOI: 10.18653/v1/2024.acl-srw.27.

[10] Podhorná-Polická, Alena. *RapCor, Francophone Rap Songs Text Corpus*. In *Proceedings of the Fourteenth Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2020)*, 2020, pp. 95-102. Available at: https://nlp.fi.muni.cz/raslan/raslan20.pdf#page=95.

[11] Mekki, Jade; Lecorvé, Gwénolé; Battistelli, Delphine; Béchet, Nicolas. *TREMoLo-Tweets: A Multi-Label Corpus of French Tweets for Language Register Characterization*. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Held Online, INCOMA Ltd., Sep 1-3, 2021, pp. 950-958. DOI: 10.26615/978-954-452-072-4_108.

[12] Panckhurst, Rachel; Lopez, Cédric; Roche, Mathieu. *A French text-message corpus: 88milSMS. Synthesis and usage*. Corpus [En ligne], 20 — 2020 (mis en ligne le 28 janvier 2020). DOI: 10.4000/corpus.4852.

[13] Pei, Zhengqi, Zhewei Sun, and Yang Xu. *Slang Detection and Identification*. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL 2019)*, Hong Kong, China, 2019, pp. 881-889. Available at: https://aclanthology.org/K19-1082/.

[14] Sun, Zhewei, Qian Hu, et al. *Toward Informal Language Processing: Knowledge of Slang in Large Language Models*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*, 2024. DOI: 10.18653/v1/2024.naacl-long.94.

[15] Anonymous. *Slang or Not? Exploring NLP Techniques for Slang Detection Using the SlangTrack Dataset*. ACL ARR (OpenReview) submission, December 2024 (ACL ARR 2024 December). Available at: https://openreview.net/forum?id=bIS03DD8sU.

[16] Wu, Tianyang; Morstatter, Fred; Liu, Huan; et al. *SlangSD: Building, Expanding, and Using a Sentiment Dictionary of Slang Words for*

Short-Text Sentiment Classification. Language Resources and Evaluation (2018). DOI: 10.1007/s10579-018-9416-0.

[17] Dhuliawala, Shehzaad; Kanojia, Diptesh; Bhattacharyya, Pushpak. *SlangNet: A WordNet like Resource for Slang Words*. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia (2016). Available at: https://www.cse.iitb.ac.in/~pb/papers/lrec16-slangnet.pdf.

[18] Wu, Tianyang; Morstatter, Fred; Liu, Huan; et al. *SlangSD: Building, Expanding, and Using a Sentiment Dictionary of Slang Words for Short-Text Sentiment Classification*. Language Resources and Evaluation (2018). DOI: 10.1007/s10579-018-9416-0.

[19] Gupta, Vishal; Rani, Rekha; et al. *SLANGZY: A Slang Word Recognition System for Hindi-English Code-Mixed Social Media Text*. In: Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2019). Kolkata, India (2019). Available at: https://aclanthology.org/K19-1082.pdf.

[20] Parent, Philippe; Parent, André. *Dictionnaire du chilleur*. Éditions Somme toute (2024). ISBN: 9782925124351.

[21] Méla, Vivienne. *Le verlan ou le langage du miroir*. Langages, No. 101, Les javanais (Mars 1991), pp. 73–94. Published by Armand Colin. Available at: https://www.jstor.org/stable/23906698.

[22] Kaye, Jonathan D.; Lowenstamm, Jean. *De la syllabicité*. In: Dell, François; Hirst, Daniel; Vergnaud, Jean-Roger (eds.), *Forme sonore du langage*. Hermann, Paris (1984), pp. 123–159. Available at: https://archive.org/details/formesonoredulangage.

[23] Russell, Robert C.; Odell, Margaret K. *Soundex system of indexing names*. U.S. Patent 1,261,167, filed June 3, 1918, and issued April 2, 1918. Available at: https://patents.google.com/patent/US1261167A/en.

[24] Levenshtein, Vladimir I. *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady, vol. 10, no. 8, 1966, pp. 707-710. Available at: https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

[25] Philips, Lawrence. *Hanging on the Metaphone*. Computer Language (1990). Available at: https://aspell.net/metaphone/.

[26] Philips, Lawrence. *The Double Metaphone Search Algorithm*. C/C++ Users Journal (June 2000). Available at: https://xlinux.nist.gov/dads/HTML/doubleMetaphone.html.

[27] Kukich, Karen. *Techniques for automatically correcting words in text*. ACM Computing Surveys (1992). DOI: 10.1145/146370.146380.

[28] Sproat, Richard; Black, Alan W.; Chen, Stanley; Kumar, Shankar; Ostendorf, Mari; Richards, Christopher. *Normalization of non-standard words*. Computer Speech & Language, 15(3):287–333 (2001). DOI: 10.1006/csla.2001.0169.

[29] Aw, AiTi; Zhang, Min; Xiao, Juan; Su, Jian. *A Phrase-Based Statistical Model for SMS Text Normalization*. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions (2006), pages 33–40. Available at: https://aclanthology.org/P06-2005/.

[30] Beaufort, Richard; Roekhaut, Sophie; Cougnon, Louise-Amélie; Fairon, Cédrick. *A Hybrid Rule/Model-Based Finite-State Framework for Normalizing SMS Messages*. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010). Available at: https://aclanthology.org/P10-1079.pdf.

[31] Han, Bo; Baldwin, Timothy. *Lexical Normalisation of Short Text Messages: Makn Sens a #twitter*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), pages 368–378. Available at: https://aclanthology.org/P11-1038/

[32] Baldwin, Timothy; de Marneffe, Marie Catherine; Han, Bo; Kim, Young-Bum; Ritter, Alan; Xu, Wei. *Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition*. Proceedings of the Workshop on Noisy User-generated Text (W-NUT 2015). DOI: 10.18653/v1/W15-4319.

[33] Urban Dictionary Embeddings. *Urban Dictionary Embeddings for Slang NLP Applications*. LREC / ACL Anthology (2020). Available at: https://aclanthology.org/2020.lrec-1.586/.

[34] Sun, X.; et al. *Knowledge of Slang in Large Language Models*. Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-Long 2024). Available at: https://aclanthology.org/2024.naacl-long.94/.

Appendix A Some extra things

[System message] You are a linguist who identifies verlan (French reversed syllable slang). Ignore any prior memories or cached context and follow only the instructions in this conversation. Do not browse the internet or use external tools; base your reasoning purely on the text you receive here. Reply with a single digit: "1" if the sentence contains verlan; otherwise reply "0". Do not include extra words, punctuation, or explanations.

[User message] You will be given one or more French sentences. For each sentence, decide whether it contains verlan and answer with a single digit (0 or 1) per sentence, in the same order that the sentences appear.

Sentences to evaluate:

Model	Overall (main test)				Standard French subset			
	TN	FP	FN	TP	TN	FP	FN	TP
Frozen+LR	434	99	49	271	393	40	2	0
E2E+LR	449	84	46	274	363	55	1	1
Frozen+BERT	449	84	43	277	399	35	1	1
E2E+BERT	468	65	35	285	387	42	0	3
Mistral-7B Zero Shot	339	194	67	253	325	108	2	2

Table 3: Draft confusion-count summary on the primary held-out test split (853 sentences) and its standard-French subset.

Appendix B Aims and Objectives

Interim report only! – you do not need to include this appendix in the final report. However, in your interim the last appendix should include your original Aims and Objectives, and, if the things have changed, the revised Aims and Objectives. If you used the LaTeX template provided for your Aims and objectives document, just copy the `\paragraph{Aims}` and `\paragraph{Objectives}` sections and paste them here.

Original

Aims Here you are describing the term goal of the project. What do you want to achieve by the end? What is the ultimate goal of this work? For example, the primary aim of this document is to have students produce suitable aims and objectives for their COSC480/490 project. While the aims and objectives document is not an assessed deliverable, a clear definition of what is to be done, and a bit of planning of how it is to be accomplished is paramount to the project's success. It is important to establish the scope of the project.

Objectives Objectives list the milestones that you need to achieve in order to achieve the projects aim(s). It's a rough plan for what needs to happen in what order. It's best to list the objectives in bullet point form. For many projects the structure to these objectives might follow the following pattern (objective names are just examples – you can have different objective names):

- background reading; going through the literature; learning about the research field;
- setting up of some kind of system for the project; getting the environment for experiments working;
- conducting preliminary experiments; implementation of a basic/simple approach; producing base case results;
- trying method 1; recording the results;
- trying method 2; recording the results.

Revised

Aims Here you are describing the term goal of the project. What do you want to achieve by the end? What is the ultimate goal of this work? For example, the primary aim of this document is to have students produce suitable aims and objectives for their COSC480/490 project. While the aims and objectives document is not an assessed deliverable, a clear definition of what is to be done, and a bit of planning of how it is to be accomplished is paramount to the project's success. It is important to establish the scope of the project.

Objectives Objectives list the milestones that you need to achieve in order to achieve the projects aim(s). It's a rough plan for what needs to happen in what order. It's best to list the objectives in bullet point form. For many projects the structure to these objectives might follow the following pattern (objective names are just examples – you can have different objective names):

- background reading; going through the literature; learning about the research field;
- setting up of some kind of system for the project; getting the environment for experiments working;
- conducting preliminary experiments; implementation of a basic/simple approach; producing base case results;
- trying method 1; recording the results;
- trying method 2; recording the results.