

UNIVERSITY OF OTAGO

SCHOOL OF COMPUTING

COSC385 PROJECT REPORT

Talking in French Like an Academia

Machine Learning Powered Verlan Identification

Author:

Yitian LI (4556502)

Supervisor(s):

Dr. Lech SZYMANSKI

Dr. Veronica

LIESAPUTRA

October 8, 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 Latin¹ 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 societies³[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'tit⁴ 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

¹en.wikipedia.org/wiki/Pig_Latin

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

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

⁴Standard 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 DeepL⁵ and Google Translate⁶ 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.

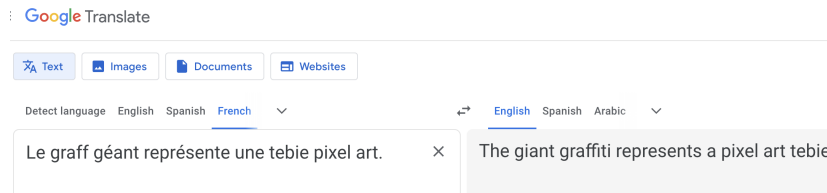


Figure 1: Google Translate cannot translate the Verlan *tebie* correctly.

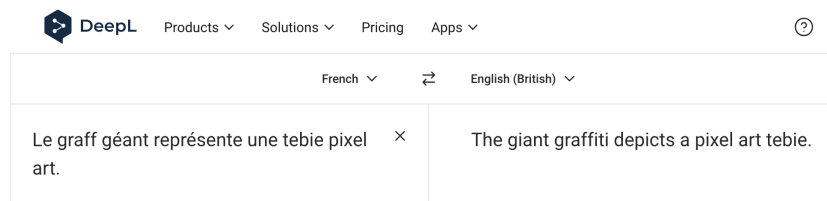


Figure 2: DeepL cannot translate the Verlan *tebie* correctly.

⁵www.deepl.com

⁶translate.google.com

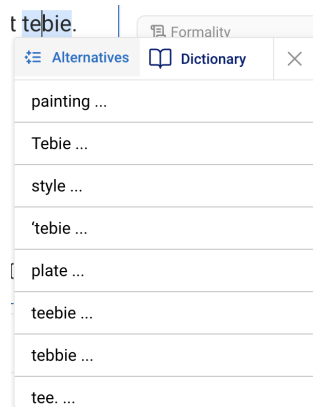


Figure 3: No desired translation for Verlan *tebie* in DeepL’s alternative word list.

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 translations⁷, nor does a proper dataset exist. The only work similar to this is an assignment published at the University of Toronto⁸, 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,

⁷Until 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.

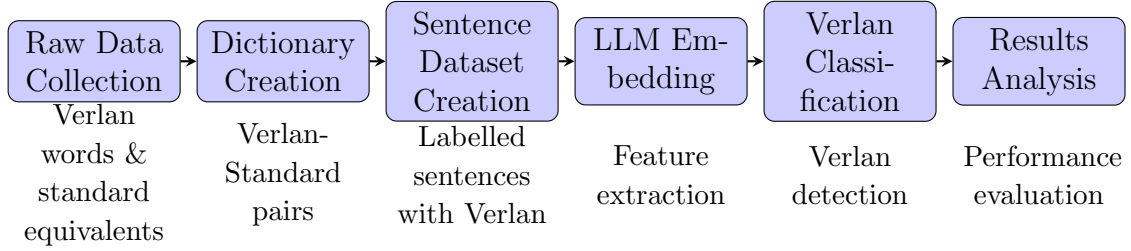


Figure 4: A visulisation of the objectives.

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 How Francophone People Use Verlan

This section is to help those readers who don't have a linguistic background in Verlan to clean their fogs. TODO ref the Paris VIII thesis. mention academie francaise.

⁹github.com/greateden/Verlan-Identification-Normalisation

2.2 How to Detect Verlan

To the best of our knowledge, there is no existing computational research¹⁰ on the *detection* of Verlan — this particular form of French slang. However, there are a few scholars who included Verlan in their research[9, 10, 11, 12]. Yet, these research commonly included Verlan as a type of slang in their dataset or corpus. Moreover, they did not extract and research on how to detect this specific type of slang, but in a broader way — they create slang datasets which contain Verlan, and some of them used computational approaches to detect those slang.

Luckily, there are quite a few papers related to computational slang detection, and their approaches could contribute to Verlan detection to a large extent[13, 14, 15, 16], which are not limited to French, but also cover other Indo-European languages¹¹.

Therefore, regarding the history of Verlan detection, this report firstly generalises the task as slang detection, constructed based on a slang detection paper’s background chapter[13]. After each key historical point, there will be discussions related to theoretical ways to implement those methods for Verlan detection, in order to provide the readers with a general and useful background.

2.2.1 Dictionary Search

The easiest way we can think of is to use a dictionary, or a corpus — just like how we look up a word that we don’t know of. The pros and cons are just like using a dictionary. It is fast (if using a computer) and accurate. On the other hand, the resource is limited, and it only works with those words which already existed, thus it cannot identify newly invented words. The existing works for slang glossaries are, for example, This is

¹⁰As of September 2025.

¹¹For example, English, German, and Russian. For more information, please refer to: https://en.wikipedia.org/wiki/Indo-European_languages.

| Item | Quantity |
|---------|----------|
| Widgets | 42 |
| Gadgets | 13 |

Table 1: An example table.

2.2.2 Traditional ML

3 L^AT_EX markup examples

3.1 Sections

Use `\section{}` and `\subsection{}` commands to organise your document. L^AT_EX handles all the formatting and numbering automatically. Use `\label{}` and `\ref{}` commands for cross-references.

3.2 Comments

Comments might be useful during the writing process, as reminders or questions to your supervisor (who should get a chance to comment on your report). Comments can be added to the margins of the document using the `\todo{}` command, as shown in the example on the right. You can also add inline comments:

This is an inline comment.

Here's
a com-
ment
in the
mar-
gin!

3.3 Tables and Figures

Use the `table{}` and `\tabular{}` commands for basic tables — see Table 1, for example. You can include a figure (JPEG, PNG or PDF) with the `\includegraphics{}` command as in the code for Figure ?? below.

3.4 Mathematics

L^AT_EX is great at typesetting mathematics. Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with $E[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_i^n X_i \quad (1)$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$. You can also reference labeled equations, such as Equation 1.

3.5 Lists

You can make lists with automatic numbering ...

1. Like this,
2. and like this.

...or bullet points ...

- Like this,
- and like this.

4 Conclusion

Concluding remarks. Send the pdf (not the `*.tex` file) to your supervisor for comments (as early as possible). Don't forget to change the `\usepackage[draft]{cosc4x0style}` setting to `\usepackage{cosc4x0style}` to produce the pdf in the format for the final submission.

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] Hajiyevea, 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. Available at: <https://aclanthology.org/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. Available at: <https://aclanthology.org/2021.ranlp-1.108/>.
- [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. Available at: <https://journals.openedition.org/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. Available at: <https://aclanthology.org/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. Available at: <https://link.springer.com/article/10.1007/s10579-018-9416-0>.

Appendix A Some extra things

If you have anything more to add such as:

- not essential details - things that might be too much for first time reading, or could be distracting from the main points...but are still important for reproducibility or deeper understanding
- work that was done in the project but doesn't go with the main work, or detracts/is not essential for the main narrative.

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 L^AT_EX 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.