Reading & Research Proposal

Title: Toward a Grammar for French Idioms

Student: Jackson Eshbaugh

Proposed Advisor: Dr. Maria Hernandez

Semester: Spring 2026

1 Thesis Context

What makes idioms unique? This is at the center of my thesis research. While we have some understanding of this interesting literary device, I am interested in expanding our understanding to a much more mathematical level, precisely defining idiomaticity (insofar as it is possible). I plan to approach the question from a computational angle by training a neural network to classify French multi-word expressions (MWEs) as idiomatic or non-idiomatic. Neural networks are a type of machine learning (ML) algorithm. The network I propose will accept a French MWE and decide if it is idiomatic or not and perhaps attribute some value of "idiomaticity" to the expression.

This question represents a clarification of our understanding of idioms for two reasons. First, since neural networks are mathematical devices, there must be some mathematical grammar (set of precise rules) that the network uses to delineate between idiomatic and non-idiomatic. In the thesis and in future work, I plan to uncover these rules, using tools such as linear probes, which can contextualize what a model is "thinking."

The neural network will be optimized (trained) with a loss function that utilizes back-translation techniques. This method has been demonstrated to work well for English (Yayavaram et al. 2024); however functionality in any language is not a given. Hence a central goal of my thesis is to apply this paradigm to the French language. In order to train a network, data is required. I plan to acquire this data by scraping `expressio.fr` for idiomatic expressions and mining other well-known corpora for instances of these idioms. Below, I provide a preliminary outline of thesis chapters:

1. Introduction
2. Idioms in French
3. An Introduction to Machine Learning
4. FRIdiom: a Dataset
5. A Neural Network-Based Model to Classify MWEs
6. Idioms *in Silico*: What Makes Idioms...Idioms? (Mathematically) [This would involve probing and other interpretability methods to understand *how* the model makes its

decision]

7. Idioms *in Silico* 2: A Continuum of Idioms [Can we regard idiomaticity as a gradient? Do machines do this?]
8. Conclusion

The question of idioms in computational linguistics is still open—and the majority of scholarship is focused on English. Hence work in computational linguistics focused on French constitutes important work. Before I can address this question computationally, it is wise to first consider the question linguistically. This study corresponds directly to the thesis plan outlined below, and its successful completion will provide the linguistic foundation needed for the computational and mathematical analyses that follow.

2 Objectives of the Reading & Research

- To survey existing linguistic literature on idioms and other multi-word expressions (MWEs), with a focus on French sources where available.
- To characterize idioms in terms of their semantic opacity and syntactic rigidity, contrasting them with other MWEs such as collocations and light-verb constructions.
- To examine and assess existing diagnostic criteria proposed in the literature for distinguishing idioms from other MWEs, and to evaluate their applicability to French.
- To apply these characterizations to a bounded subset of French idioms as a focused case study.
- To synthesize findings into a written report that will serve as the foundation for Chapter 2 of the thesis ("Idioms in French").

3 Methodology / Approach

This study will combine literature review and linguistic analysis. First, I will survey scholarship on idioms and MWEs, paying particular attention to French idioms. This survey will focus on definitions, semantic opacity, and syntactic constraints. From this review, I will extract and summarize diagnostic criteria that have been proposed for distinguishing idioms from other MWEs.

Next, I will select a bounded subset of French idioms from resources such as `expressio.fr` and representative corpora. Using these data, I will apply the identified diagnostic criteria in order to characterize the idioms' semantic and syntactic behavior. Where criteria prove insufficient or inconsistent, I will note limitations and potential refinements for further exploration in the thesis.

The outcome will be a structured report synthesizing the findings, which will form the basis of Chapter 2 of the thesis.

4 Deliverables

- A written report characterizing French idioms in terms of semantic opacity and syntactic rigidity, with comparison to other MWEs.
- A structured summary of diagnostic criteria from the linguistic literature, annotated with observations about their applicability to French.
- A set of rules or "pattern classes" which can be used to identify idioms.

5 Phases

The independent study will proceed in three phases:

- **Phase 1: Literature Review**—Many sources centering on idioms will be reviewed. Books, chapters, journal articles, and conference papers are in scope for this review, and they can speak about idioms in general, or idioms in French. However, a specific emphasis will be placed on French idioms. Please see Section 6 for a preliminary list of possible sources for this phase.
- **Phase 2: Creation of a Descriptive Grammar or Rule-Based Framework for Identifying Idiomatic Pattern Classes**—Inductively, I will define a set of rules that seem to cover a majority of idioms. This will be done with work reported in the literature reviewed in Phase 1 and a set of gathered idioms. These rules can also be regarded as "pattern classes." The set of idioms will be used both to develop and validate this theoretical framework.
- **Phase 3: Reporting of Findings**—A linguistics journal-style paper will be prepared in French as the main deliverable from this Reading & Research. In it, I will include sections similar to the following: Introduction, Related Works, Methodology, Results & Discussion, and Conclusion. This paper will lay out my findings from the semester, and will be adapted to form Chapter 2 ("Idioms in French") of my thesis.

6 Preliminary References

Booth, Trudie Maria (2005). *French verbs and idioms*. Lanham: University Press of America. ISBN: 978-0-7618-3194-5.

Espinal, M. Teresa and Jaume Mateu (May 2019). *Idioms and Phraseology*. DOI: 10.1093/acrefore/9780199384655.013.51. URL: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-51.

Everaert, Martin et al. (2014). *Idioms: Structural and psychological perspectives*. Psychology Press.

Linden, Erik-Jan van der (1992). "IDIOMS, NON-LITERAL LANGUAGE AND KNOWLEDGE REPRESENTATION". In: *Computational Intelligence* 8.3, pp. 433–453. DOI: https://doi.org/10.1111/j.1467-8640.1992.tb00374.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.1992.tb00374.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.1992.tb00374.x.

Lupson, J. P. and M. L. Pélissier (1998). *Guide to French idioms =: Guide des locutions françaises*. Lincolnwood, Ill: Passport Books. ISBN: 978-0-8442-1502-0.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow (Sept. 1994). "Idioms". In: *Language* 70.3, p. 491. ISSN: 00978507. DOI: 10.2307/416483. JSTOR: 416483. (Visited on 08/22/2025).

Owens, Jonathan and Robin Dodsworth (2017). "Semantic mapping: What happens to idioms in discourse". In: *Linguistics* 55.3, pp. 641–682.

Van Goethem, Kristel and Dany Amiot (Jan. 2019). "Compounds and Multi-Word Expressions in French". In: *Complex Lexical Units: Compounds and Multi-Word Expressions*. Ed. by Barbara(ed.) Schlücker. De Gruyter, pp. 127–152. ISBN: 978-3-11-063244-6. DOI: 10.1515/9783110632446-005.

Yayavaram, Arnav et al. (May 2024). "BERT-based Idiom Identification using Language Translation and Word Cohesion". In: *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*. Ed. by Archna Bhatia et al. Torino, Italia: ELRA and ICCL, pp. 220–230. URL: https://aclanthology.org/2024.mwe-1.26/.