From lexical keywords to contextual embeddings: A neuro-symbolic study of French idiomatic semantics

Anonymous ACL submission

Abstract

Idiomatic expressions remain challenging for NLP because their meanings are often figurative and non-compositional. This study presents a semantic ontology of 764 French idioms built around body-part and animal keywords, such as *tête* ("head"), *cœur* ("heart"), *main* ("hand"), *vache* ("cow"), and *chien* ("dog"). The resource organizes idioms into 15 semantic categories. We compare a symbolic rule-based baseline, a TF-IDF + Logistic Regression model, and a fine-tuned CamemBERT classifier. Results show that symbolic lexical-semantic rules remain highly competitive and that symbolic and contextual models capture complementary aspects of figurative meaning.

Keywords: figurative language processing, idioms, semantic classification, ontology construction, CamemBERT, graph representation, French language processing

1 Introduction

Idiomatic expressions, a type of phraseological units & idiomatic multiword expression (MWE) (Polguère, 2002; Constant, 2012; Constant et al., 2017; Chen, 2021; Savary et al., 2023), remain challenging for natural language processing because their meanings are often figurative, non-compositional, and culturally grounded (Chen, 2022b; Dankers et al., 2022; fig, 2024; Miletić and im Walde, 2024; Madabushi et al., 2022). Expressions such as *poser un lapin à quelqu'un* (to stand someone up) or *avoir un bœuf sur la langue* (to be unable to speak) cannot be interpreted through literal lexical composition alone, making idiom understanding particularly difficult for computational models (Dankers et al., 2022). In this paper, we present a semantic ontology of 764 French idiomatic expressions involving body-part (Chen, 2022a) and animal (Chen, 2020) metaphors. The resource organizes idioms into 15 semantic categories related to emotion, cognition, behavior, social interaction, and embodied experience. We compare three semantic classification paradigms under a low-resource figurative language setting: (1) a symbolic rule-based baseline, (2) a TF-IDF + Logistic Regression classifier, and (3) a fine-tuned CamemBERT model. Our experiments show that simple symbolic lexical-semantic rules remain highly competitive and, under balanced semantic evaluation, outperform contextual transformer models. Error analysis further demonstrates that sparse lexical representations and contextual neural embeddings capture complementary aspects of figurative meaning. These findings suggest that figurative semantics preserves embodied lexical-semantic regularities that remain difficult for contextual transformers to learn under low-resource conditions (Wan et al., 2024; Körner et al., 2023).

2 Related work

2.1 Idiom processing and figurative NLP

Early computational approaches mainly relied on lexical matching, statistical co-occurrence patterns, and phraseological extraction methods, whereas more recent work has explored contextual transformer architectures for idiom detection, metaphor interpretation (Cowie, 1998; Granger and Meunier, 2008; Mel'čuk, 2023; Polguère, 2002, 2014; Chen, 2021), and figurative semantic classification (Madabushi et al., 2022; Miletić and im Walde, 2024). Despite these advances, figurative semantic modeling remains difficult because idiomatic expressions frequently involve overlapping conceptual domains such as emotion, cognition, and social interaction. In addition, most existing studies focus primarily on contextual representation learning while paying limited attention to explicit semantic ontology construction and embod-

ied conceptual organization (Körner et al., 2023; Wan et al., 2024).

2.2 Transformer models for figurative semantics

Transformer-based language models such as BERT and RoBERTa have achieved strong performance across many NLP tasks, including figurative language processing (Miletić and im Walde, 2024; Madabushi et al., 2022). Recent studies have applied contextual transformers to metaphor detection, sarcasm identification, and idiom interpretation (Dankers et al., 2022; Potamias et al., 2020). However, relatively little work has explored transformer-based semantic classification of idiomatic expressions under small annotated figurative dataset settings. Moreover, figurative meanings often involve implicit conceptual mappings and culturally grounded semantic associations that remain difficult to recover from contextual representations alone (Körner et al., 2023; Wan et al., 2024). Recent work additionally suggests that embodied semantic information and explicit conceptual structure may remain insufficiently represented in purely contextual transformer architectures.

2.3 Ontologies and semantic representation of idioms

Ontology construction and semantic knowledge representation have become increasingly important in computational phraseology and figurative language research (Fiorelli et al., 2020; Abgaz, 2020; Lindemann et al., 2023). Previous work has explored semantic networks, lexical ontologies, and knowledge graphs for representing idiomatic meaning and conceptual relations between figurative expressions (Fiorelli et al., 2020; Lindemann et al., 2023). However, many existing resources rely on single-label taxonomies or manually curated semantic inventories without explicit computational evaluation. In addition, graph-theoretic analysis of idiomatic semantic organization remains relatively underexplored in NLP-oriented figurative language studies.

2.4 Embodied cognition and conceptual metaphor

The present study is also grounded in theories of embodied cognition and conceptual metaphor. Previous linguistic research has shown that figurative meaning is frequently structured through recurrent bodily experience and conceptual mappings linking physical perception to abstract semantic domains (Desai, 2022; Wan et al., 2024; Li et al., 2024). Body-part lexemes such as *tête* ("head"), *cœur* ("heart"), *main* ("hand"), and *œil* ("eye") often function as conceptual anchors for cognition, emotion, perception, and action in French idiomatic phraseology. Animal metaphors involving *chien* ("dog"), *chat* ("cat"), *cheval* ("horse"), and *loup* ("wolf") similarly encode socially and culturally salient evaluative meanings. However, relatively few NLP studies have explicitly integrated embodied semantic organization into computational idiom classification frameworks (Wan et al., 2024; Li et al., 2024).

2.5 Neuro-symbolic NLP

Recent work in neuro-symbolic NLP has emphasized the importance of combining symbolic semantic knowledge with contextual neural representations (DeLong et al., 2024; Liu et al., 2024). Hybrid approaches integrating ontologies, knowledge graphs, symbolic constraints, and neural language models have shown promising results in semantic parsing, commonsense reasoning, and knowledge-aware language understanding (DeLong et al., 2024; Liu et al., 2024). Figurative language processing constitutes a particularly relevant domain for neuro-symbolic modeling because idiomatic meaning simultaneously involves lexical embodiment, conceptual metaphor, and contextual interpretation. The present study contributes to this direction by comparing symbolic, sparse statistical, and contextual neural approaches for idiomatic semantic classification.

3 Dataset and ontology construction

Figure 1 provides an overview of the complete ontology construction and figurative semantic analysis pipeline used in this study, including dictionary-based idiom collection, semantic normalization, graph-based analysis, classification experiments, and interactive ontology visualization.

Figure 1: Pipeline for ontology construction, semantic analysis, and figurative idiom classification.

3.1 Dataset construction

The idiomatic inventory was compiled from monolingual and bilingual dictionary resources, focusing on French idiomatic expressions involving body-part and animal-related metaphors. We first collected an initial lexicon of approximately 750 idiomatic expressions. After normalization, manual validation, and ontology construction, the final ground-truth dataset contained 764 annotated idiomatic expressions. Each entry corresponds to one French idiomatic expression associated with lexical, semantic, and explanatory annotations. The resource was stored in CSV format and contains seven fields: *expression*, *keyword*, *literal_keyword_category*, *primary_label*, *secondary_labels*, *justification*, and *source_label*. The inventory mainly focuses on expressions involving body parts and animals, such as *faire la barbe à quelqu'un* ("to shave someone"), *avoir la gueule de bois* ("to have a hangover"), *garder bouche cousue* ("to keep one's mouth shut"), *prendre le taureau par les cornes* ("to take the bull by the horns"), and *pleurer comme un veau* ("to cry like a calf"). In addition, we used a large-scale French Wikipedia corpus extracted from OPUS v1.0, containing approximately 15.8 million words. This corpus was used for lexical exploration, corpus attestation, and figurative usage analysis.

3.2 Semantic normalization

To improve annotation consistency and reduce semantic sparsity, we applied a multi-stage normalization pipeline. Textual fields were normalized, repeated spaces were collapsed, empty or invalid values were filtered, and duplicate rows were removed. After preprocessing, the dataset still contained 764 valid entries. The initial annotation space contained 525 fine-grained semantic labels. These labels were manually reduced to 15 semantic macro-categories. For example, *colère* ("anger"), *frustration*, and *tristesse* ("sadness") were grouped under *émotion* ("emotion"), while *silence*, *insulte* ("insult"), and *franchise* ("frankness") were mapped to *langage* ("language"). Similarly, labels such as *travail* ("work"), *effort*, and *préparation* ("preparation") were grouped under *comportement* ("behavior"). This reduction simplified the annotation space while preserving semantic interpretability.

3.3 Multi-Label representation

Because idiomatic expressions often encode multiple semantic dimensions, we adopted a multi-label representation rather than a single-label annotation scheme. Both principal and secondary labels were processed through the semantic normalization pipeline. Secondary labels were segmented using multilingual delimiters and mapped to the predefined ontology categories. Labels outside the ontology were assigned to the residual category *autre* ("other"). The final ontology includes 15 semantic macro-categories covering domains such as emotion, cognition, behavior, social interaction, physical state, appearance, morality, capacity, value, and food-related meanings.

3.4 Dataset statistics

After semantic normalization and multi-label construction, the final dataset contained 764 French idiomatic expressions annotated with 15 semantic macro-categories.

Item	Value
Idiomatic expressions	764
Semantic categories	15
Single-label expressions	224
Multi-label expressions	540
Multi-label ratio	70.68%

Table 1: Dataset Statistics

The label cardinality distribution shows that 224 expressions contain one semantic label, 438 contain two labels, and 102 contain three labels.

Labels per expression	Count
1 label	224
2 labels	438
3 labels	102

Table 2: Multi-Label Cardinality Distribution.

Overall, 70.68% of the expressions are associated with more than one semantic category, confirming the need for a multi-label representation

of idiomatic meaning. The most frequent semantic categories are shown below.

Category	Frequency
other	722
emotion	116
cognition	75
behavior	74
social	57
morality	55

Table 3: Frequency Distribution of Semantic Categories.

The high frequency of *other* reflects the difficulty of assigning rigid semantic boundaries to figurative expressions. At the same time, categories such as *emotion*, *cognition*, *behavior*, and *social* show that French idioms frequently encode affective, mental, behavioral, and interactional meanings.

4 Corpus Validation and Semantic Graph Analysis

4.1 Corpus Attestation

After constructing the multi-label ontology, we conducted a corpus attestation experiment to examine whether the annotated idioms occurred in naturally occurring text. The validation corpus consisted of 50 French textual chunks, and attestation was performed using regular-expression matching over normalized surface forms. The ontology contains 764 annotated entries corresponding to 754 unique idiomatic expressions, since several expressions are associated with multiple semantic annotations. Among these 754 unique idiomatic expressions, 46 were detected in the corpus, corresponding to a coverage rate of 6.1%.

Item	Value
Idioms in ontology	754
Attested idioms	46
Corpus coverage	6.1%
Corpus files	50

Table 4: Corpus attestation statistics.

Although the attestation rate is low, it should be interpreted as a conservative estimate because the method relies on exact surface-form matching. French idioms frequently undergo lexical, morphological, and syntactic variation, which strict matching cannot capture. Several idioms appeared multiple times, including *corps et âme* ("body and soul") and *en chair et en os* ("in the flesh").

4.2 Lexical Distribution

We analyzed the distribution of lexical keywords across the complete ontology and the attested subset. The ontology is dominated by body-part keywords, including *tête* ("head", 46), *cœur* ("heart", 44), *pied* ("foot", 26), *nez* ("nose", 26), *main* ("hand", 24), *œil* ("eye", 22), and *yeux* ("eyes", 22). Animal-related keywords were less frequent, including *chien* ("dog", 17), *chat* ("cat", 13), *cheval* ("horse", 12), *coq* ("rooster", 8), and *loup* ("wolf", 8).

The attested subset preserved the main body-part profile, with *main* ("hand"), *œil* ("eye"), *tête* ("head"), *pied* ("foot"), and *cœur* ("heart") among the most frequent attested keywords. At the same time, many low-frequency keywords disappeared after corpus filtering, reflecting a long-tail distribution that creates challenges for statistical semantic modeling.

4.3 Semantic Distribution

The semantic distribution was highly imbalanced. The residual category *other* dominated the ontology with 722 occurrences, followed by *emotion* (116), *cognition* (75), *behavior* (74), and *social* (57). This imbalance motivates the use of balanced evaluation metrics in the classification experiments.

The attested subset showed a filtering effect: categories such as *cognition*, *social*, *emotion*, and *behavior* remained present, whereas food-related meanings and *desire* disappeared under exact matching. This suggests that corpus attestation is sensitive not only to lexical frequency but also to semantic domain productivity.

4.4 Graph Centrality Analysis

To examine the structural organization of the ontology, we represented lexical-semantic relations as a directed graph using NetworkX. The graph connects idiomatic expressions, lexical keywords, and semantic categories. We computed degree, in-degree, out-degree, and betweenness centrality for both the complete ontology graph and the corpus-attested subgraph.

Keyword	English gloss	Degree (All)	Degree (Attested)	Betweenness (Attested)
tête	head	60	10	0.002351
cœur	heart	55	7	0.000882
œil	eye	33	12	0.003379
main	hand	32	10	0.001322
pied	foot	37	5	0.000000

Table 5: Graph centrality measures for embodied keywords.

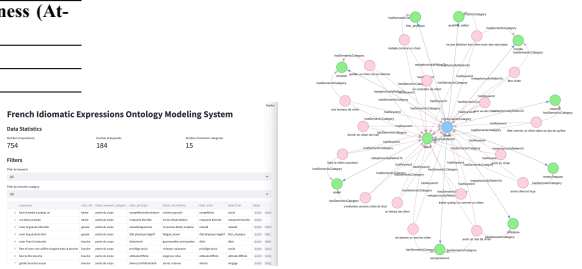

Figure 2: Streamlit platform for ontology exploration and graph visualization (left), and semantic graph centered on the lexical keyword *chien* (“dog”) (right).

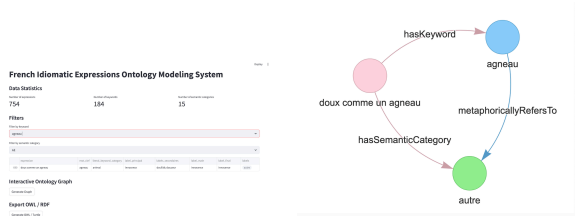

Figure 3: Minimal ontology graph of idiomatic semantic relations.

The highest-centrality nodes in the complete graph were mainly body-part keywords, including *tête* (“head”), *cœur* (“heart”), *nez* (“nose”), *pied* (“foot”), *œil* (“eye”), *yeux* (“eyes”), and *main* (“hand”). Animal-related keywords such as *chien* (“dog”), *chat* (“cat”), *cheval* (“horse”), *coq* (“rooster”), and *loup* (“wolf”) formed secondary lexical regions but showed lower centrality values.

The attested graph preserved the same general hub structure, especially around *œil* (“eye”), *tête* (“head”), *main* (“hand”), and *cœur* (“heart”). In particular, *œil* showed the highest betweenness centrality in the attested subgraph, suggesting that perception-related idioms occupy an important structural position in naturally occurring usage.

4.5 Interactive Ontology Exploration Platform

To facilitate ontology exploration and qualitative analysis, we developed an interactive Streamlit-based platform for navigating the idiomatic semantic resource. The interface supports filtering by lexical keyword and semantic category, annotation inspection, graph visualization, and RDF/OWL export.

The resource contains 764 annotated entries corresponding to 754 unique idiomatic expressions, since several expressions are associated with multiple semantic annotations. Graph visualizations represent relations between idiomatic expressions, lexical keywords, and semantic categories. Figure 2 shows the Streamlit interface, while Figure 3 illustrates a graph centered on the lexical keyword *chien* (“dog”).

4.6 Interpretation

The comparison between the complete graph and the attested subgraph shows that corpus filtering reduces lexical diversity while strengthening the relative importance of a small number of embodied keywords. For instance, the betweenness centrality of *œil* (“eye”) increased from 0.000221 in the complete graph to 0.003379 in the attested graph. Similar increases were observed for *tête* (“head”) and *main* (“hand”). These results suggest that idiomatic semantic organization is structured around recurring lexical domains related to perception, cognition, action, and bodily experience. From an NLP perspective, these hubs may serve as useful anchor nodes for semantic clustering, graph embeddings, ontology alignment, and graph-based figurative language models.

5 Experimental Setup

This section presents the experimental framework used to evaluate symbolic, statistical, and neural approaches for idiomatic semantic classification. The experiments compare three semantic modeling paradigms under the same low-resource figurative language setting: (1) a symbolic rule-based baseline grounded in lexical-semantic associations, (2) a TF-IDF + Logistic Regression classifier, and (3) a contextual transformer classifier based on CamemBERT. All models were evaluated using

the same supervised train/test split to ensure fair and reproducible comparison.

5.1 Experimental Protocol

After constructing the multi-label idiomatic ontology, we transformed the resource into a supervised single-label classification dataset. Since many standard classifiers assume one target label per instance, we introduced a deterministic label-selection strategy. For each idiomatic expression, the first semantic category different from *other* was selected whenever available. If no more specific semantic category existed, *other* was retained as the principal label.

The dataset was partitioned into training and test subsets using a stratified split that preserved semantic category distributions across both partitions. The final configuration contained 611 training expressions and 153 test expressions. The same train/test split was reused across all symbolic, statistical, and neural experiments to ensure methodological comparability. Semantic labels were encoded into numerical identifiers. Bidirectional *label2id* and *id2label* dictionaries were constructed to preserve interpretability between ontology categories and model outputs. The final label inventory contained 15 semantic classes.

5.2 Rule-Based Baseline

As a symbolic baseline, we constructed a rule-based classifier associating each lexical keyword with its most frequently observed semantic category. The rule system was built exclusively from the training split to avoid data leakage. The resulting rule dictionary contained 158 keyword-to-label mappings extracted from the training subset. During prediction, the classifier retrieved the keyword associated with each idiomatic expression and assigned its dominant semantic category. When an unknown keyword appeared in the test set, the system defaulted to the majority class, *other*.

5.3 TF-IDF + Logistic Regression

To evaluate sparse lexical-semantic representations, we trained a classifier based on TF-IDF features and Logistic Regression. Idiomatic expressions were represented using a hybrid TF-IDF feature space combining word-level and character-level n-grams. Word-level representations used n-grams from 1 to 3, while character-level representations used character windows from 3 to 5. The resulting sparse feature space contained 9,183 dimensions. A Logistic Regression classifier was trained using the *liblinear* solver with a maximum of 3,000 iterations. Since the semantic categories remained imbalanced, the classifier used class_weight="balanced" during optimization.

5.4 CamemBERT

To model contextual figurative semantics, we fine-tuned *camembert-base*, a RoBERTa-style transformer pretrained on large-scale French corpora. Idiomatic expressions were tokenized using the CamemBERT tokenizer with a maximum sequence length of 32. The classifier was initialized with *AutoModelForSequenceClassification* using 15 output classes corresponding to the ontology categories. The model was fine-tuned for 8 epochs with a learning rate of 2e-5, a batch size of 8 for both training and evaluation, and a weight decay of 0.01. Evaluation and checkpoint saving were performed at the end of each epoch. The best checkpoint was selected according to macro F1, which is appropriate for the imbalanced semantic classification setting. We report accuracy, macro F1, and weighted F1 for final evaluation. The same train/test split used for the symbolic and statistical baselines was reused for transformer fine-tuning to ensure methodological comparability across all experiments.

6 Results and Comparative Evaluation

This section compares the three semantic classification paradigms evaluated in this study: a symbolic rule-based baseline, a TF-IDF + Logistic Regression classifier, and a fine-tuned CamemBERT model. All systems were evaluated on the same held-out test set of 153 French idiomatic expressions to ensure fair comparison under identical low-resource conditions.

6.1 Rule-Based Baseline

The rule-based baseline achieved the strongest balanced performance among the three systems, with an accuracy of 0.2614, a Macro F1 score of 0.1494, and a Weighted F1 score of 0.2154. Despite its simplicity, this baseline captures stable lexical-semantic regularities associated with recurring body-part keywords such as *tête* ("head"), *cœur* ("heart"), *main* ("hand"), *œil* ("eye"), and *langue* ("tongue/language"). However, the rule-based model remains limited because it cannot distinguish different figurative meanings associated

with the same keyword. Categories such as *social*, *capacity*, *morality*, and *time/space* remained difficult to predict using lexical anchors alone.

6.2 TF-IDF + Logistic Regression

The TF-IDF + Logistic Regression classifier achieved an accuracy of 0.2092, a Macro F1 score of 0.1375, and a Weighted F1 score of 0.2084. It captured some surface phraseological regularities, especially for categories involving recurrent lexical patterns. However, its performance remained slightly below the symbolic baseline. The classifier obtained partial recognition for categories such as *appearance*, *cognition*, and *behavior*, but struggled with minority categories such as *desire*, *social*, *time/space*, and *quantity/value*. This confirms that surface lexical similarity alone is insufficient for robust figurative semantic classification, since idioms with similar lexical material may express different semantic meanings.

6.3 CamemBERT

The fine-tuned CamemBERT model achieved an accuracy of 0.2614, a Macro F1 score of 0.0704, and a Weighted F1 score of 0.1637. Although its accuracy matched the rule-based baseline, its Macro F1 score was substantially lower, indicating poor generalization across minority semantic categories. The final model strongly favored the majority category *other*. Most minority categories, including food-related meanings, *appearance*, *capacity*, *language*, *social*, and *time/space*, received F1 scores close to zero. Only a few categories, such as *cognition*, *emotion*, and *problem*, obtained partial recognition. This weak performance can be explained by the small training set of 611 expressions, the strong semantic imbalance, and the non-compositional nature of idiomatic meaning.

6.4 Final Comparison

The final results are shown in Table 6.

Model	Accuracy	Macro F1	Weighted F1
Rule baseline	0.2614	0.1494	0.2154
TF-IDF + LR	0.2092	0.1375	0.2084
CamemBERT	0.2614	0.0704	0.1637

Table 6: Classification performance across models.

The comparison reveals that the symbolic rule-based baseline achieved the best Macro F1 score, despite being the simplest model. TF-IDF + Logistic Regression provided moderate lexical discrimination but remained limited by surface-form dependence. CamemBERT reached the same accuracy as the rule baseline but performed worse on balanced semantic evaluation, indicating poor generalization to minority categories. These results show that accuracy alone is misleading under strong class imbalance. Macro F1 is more informative for this task because it better reflects performance across all semantic categories. The main finding is that symbolic lexical-semantic knowledge remains highly competitive under low-resource figurative language conditions.

7 Error Analysis and Discussion

This section presents a qualitative and quantitative error analysis of the three semantic classification systems: the rule-based baseline, the TF-IDF + Logistic Regression classifier, and the CamemBERT model. Beyond aggregate scores, the goal is to examine why the models fail, which semantic categories are difficult to classify, and how symbolic, sparse statistical, and contextual neural representations capture different aspects of figurative meaning.

Expression	English gloss	Gold	TF-IDF	CamemBER
avoir des oreilles en feuilles de chou	"to have cauliflower ears"	appearance	appearance	other
tendre l'oreille	"to listen carefully"	cognition	cognition	other
se ronger les sangs	"to worry oneself sick"	emotion	social	emotion
avoir une mémoire d'éléphant	"to have an elephant's memory"	cognition	capacity	cognition

Table 7: Comparative error examples for TF-IDF and CamemBERT.

7.1 General Error Patterns

Several recurrent error patterns appeared across the three systems, including figurative ambiguity, semantic overlap between categories, lexical polysemy, and attraction toward the majority category *other*. Many idiomatic expressions contain lexical units associated with one semantic domain while expressing another figurative meaning. For example, idioms involving *tête* ("head") may express *cognition*, *emotion*, or *behavior* depending on the construction. Similarly, *main* ("hand") may encode *social* relations, *competence*, or *action*.

7.2 Majority-Class Bias

The most dominant error phenomenon concerns the influence of the residual category *other*. All three systems exhibited some degree of majority-class bias, although it was particularly severe for

the rule-based baseline and CamemBERT. Expressions such as *se crêper le chignon* ("to have a fight"), *sortir par les trous du nez* ("to be fed up with something"), and *un panier de crabes* ("a nest of vipers") were frequently predicted as *other* despite belonging to more specific categories such as *social* or *emotion*.

7.3 Cases Where TF-IDF Succeeds

Several expressions were correctly classified by TF-IDF but misclassified by CamemBERT. These cases mostly involve stable phraseological constructions with highly discriminative lexical patterns. Appearance-related expressions such as *avoir des oreilles en feuilles de chou* ("to have cauliflower ears") and *avoir une mine de papier mâché* ("to look pale and exhausted") were correctly associated with *appearance*, because their lexical combinations strongly encode descriptive figurative imagery. TF-IDF also performed well on cognition-related idioms such as *tendre l'oreille* ("to listen carefully") and *bête comme une oie* ("as stupid as a goose"), as well as behavior-oriented expressions such as *prendre son courage à deux mains* ("to pluck up one's courage") and *de pied ferme* ("firmly/resolutely"). These examples suggest that sparse lexical representations remain effective for stable idiomatic constructions with recurrent phraseological patterns.

7.4 Cases where CamemBERT succeeds

Conversely, several expressions were correctly classified by CamemBERT but misclassified by TF-IDF. These cases suggest that contextual neural representations capture complementary aspects of figurative semantics unavailable to sparse lexical models. CamemBERT performed particularly well on emotionally charged idioms such as *se ronger les sangs* ("to worry oneself sick"), *rester en travers de la gorge* ("to stick in one's throat"), and *se faire du mauvais sang* ("to worry"). The transformer also succeeded on cognition-related expressions such as *avoir un petit pois dans la tête* ("to be stupid") and *avoir une mémoire d'éléphant* ("to have an elephant's memory"). In several cases, CamemBERT captured the overall figurative interpretation more effectively than TF-IDF, particularly for expressions such as *faire de l'œil* ("to make eyes at someone") and *partir les pieds devant* ("to leave feet first / to die").

7.5 Discussion: Low-Resource Figurative Semantics

The error analysis reveals several linguistic and computational properties of idiomatic semantic processing. First, figurative semantics remain difficult to model under low-resource conditions: even pretrained transformers struggle to infer stable semantic categories when idiomatic constructions are sparse, heterogeneous, and culturally implicit. Second, lexical-semantic structures remain important. Keywords such as *tête* ("head"), *cœur* ("heart"), *main* ("hand"), *œil* ("eye"), and *langue* ("tongue/language") repeatedly participate in figurative mappings across semantic domains, partially explaining the competitiveness of the symbolic baseline. Third, contextual neural representations alone are insufficient for robust idiomatic semantic classification when training data are limited and semantic categories are imbalanced. These findings motivate hybrid approaches combining symbolic ontology knowledge, graph-based semantic representations, and contextual neural embeddings.

8 Conclusion

This paper presented a semantic ontology of 764 French idiomatic expressions involving body-part and animal metaphors. We analyzed the resource through corpus attestation, lexical distribution, and graph centrality, and compared three classification paradigms: a rule-based baseline, TF-IDF + Logistic Regression, and CamemBERT.

The results show that symbolic lexical-semantic rules remain highly competitive under low-resource figurative language conditions. Although CamemBERT captured some contextual semantic patterns, it did not outperform symbolic and sparse lexical approaches on balanced evaluation. These findings suggest that figurative semantics preserves embodied lexical-semantic regularities that remain difficult for contextual transformers to learn from limited annotated data. Future work will explore ontology-aware neural architectures, graph-based semantic learning, and hybrid neuro-symbolic models for figurative language understanding.

Limitations

This study has several limitations. First, although the dataset contains 764 manually annotated French idiomatic expressions, it remains

small for supervised neural classification, especially given the 15 semantic categories and the strong imbalance caused by the residual category *other*. Second, the annotations were manually curated, but no inter-annotator agreement was reported, which limits the assessment of annotation reliability. Third, corpus attestation was based on exact surface-form matching, which likely underestimates idiom occurrence because French idioms frequently undergo lexical, morphological, and syntactic variation. Fourth, the ontology focuses mainly on body-part and animal-related idioms and therefore does not cover the full diversity of French phraseology. Finally, the experiments evaluate single-label classification derived from an originally multi-label ontology, which simplifies the semantic multidimensionality of idiomatic expressions. Future work should expand the dataset, refine the ontology with hierarchical labels, and explore more flexible idiom detection methods and ontology-aware neural architectures.

References

2024. Figlang 2024: Workshop on figurative language processing. https://sites.google.com/view/figlang2024.

Yalemisew Abgaz. 2020. Using ontolex-lemon for representing and interlinking lexicographic collections of bavarian dialects. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 61–69, Marseille, France. European Language Resources Association (ELRA).

Lian Chen. 2020. Les stéréotypes dans la zoo-anthroponymie française et chinoise. *Itinéraires. Littérature, textes, cultures*, (2020-2).

Lian Chen. 2021. *Analyse Comparative des Expressions Idiomatiques en Chinois et en Français (Relatives au Corps Humain et aux Animaux)*. Ph.D. thesis, CY Cergy Paris Université.

Lian Chen. 2022a. Analyse phraséoculturelle contrastive : représentation et motivation du coeur en français et en chinois. In *SHS Web of Conferences*, volume 138, page 11001.

Lian Chen. 2022b. Phraséoculturologie : une sous-discipline moderne indispensable de la phraséologie. In *SHS Web of Conferences*, volume 138, page 04011.

A. P. Cowie, editor. 1998. *Phraseology: Theory, Analysis, and Applications*. Clarendon Press, Oxford.

Verna Dankers, Christopher G. Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of ACL 2022*.

Lauren Nicole DeLong, Ramon Fernández Mir, and Jacques D. Fleuriot. 2024. Neurosymbolic ai for reasoning over knowledge graphs: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Rutvik H. Desai. 2022. Are metaphors embodied? the neural evidence. *Psychonomic Bulletin & Review*.

Manuel Fiorelli, Armando Stellato, Tiziano Lorenzetti, Andrea Turbati, Peter Schmitz, Enrico Francesconi, Najeh Hajlaoui, and Brahim Batouche. 2020. Editing ontolex-lemon in vocbench 3. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7194–7203, Marseille, France. European Language Resources Association.

Sylviane Granger and Fanny Meunier, editors. 2008. *Phraseology: An Interdisciplinary Perspective*. John Benjamins, Amsterdam and Philadelphia.

Anita Körner, Mauricio Castillo, Linda Drijvers, Martin H. Fischer, Fritz Günther, Marco Marelli, Olesia Platonova, Luca Rinaldi, Samuel Shaki, James P. Trujillo, Oksana Tsaregorodtseva, and Arthur M. Glenberg. 2023. Embodied processing at six linguistic granularity levels: A consensus paper. *Journal of Cognition*, 6(1).

Yu Xi Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. 2024. Embodiedbert: Cognitively informed metaphor detection incorporating sensorimotor information. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16868–16876. Association for Computational Linguistics.

David Lindemann, Sina Ahmadi, Anas Fahad Khan, Francesco Mambrini, Federica Iurescia, and Marco Passarotti. 2023. When ontolex meets wikibase: Remodeling use cases. In *Proceedings of the 4th Wikidata Workshop*.

Lihui Liu, Zihao Wang, and Hanghang Tong. 2024. Neural-symbolic reasoning over knowledge graphs: A survey from a query perspective. In *Proceedings of ACM Conference*, New York, NY, USA. ACM.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Igor A. Mel'čuk. 2023. *General Phraseology: Theory and Practice*, volume 36 of *Lingvisticæ Investigationes Supplementa*. John Benjamins, Amsterdam.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

Alain Polguère. 2002. *Notions de base en lexicologie*. Ophrys, Paris.

Alain Polguère. 2014. Principes de modélisation systémique des réseaux lexicaux. In *Proceedings of TALN 2014*, pages 79–90. ATALA.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*.

Mingyu Wan, Qi Su, Kathleen Ahrens, and Chu-Ren Huang. 2024. Perceptional and actional enrichment for metaphor detection with sensorimotor norms. *Natural Language Engineering*, 30(6):1181–1209.