Ontology-oriented lexico-semantic modeling and neural classification of Chinese *chéngyǔ*: A culture-aware NLP approach

CHEN Lian 陈恋
LLL Laboratory, University of Orléans, France
CRLAO, CNRS-INALCO, France
lian.chen@univ-orleans.fr

Abstract

This paper proposes a semi-automatic lexico-semantic modeling framework for Chinese *chéngyǔ* containing body-part and animal lexemes. The framework combines manual semantic annotation, lightweight RDF/OWL formalization and semantic classification in order to investigate whether lexical mediators such as 心 *xīn* "heart/mind", 口 *kǒu* "mouth" or 马 *mǎ* "horse" are sufficient to predict idiomatic semantic interpretation. Based on 440 annotated *chéngyǔ* normalized into 18 semantic categories, we compare three classification approaches: a rule-based keyword baseline, character n-gram TF-IDF with logistic regression, and BERT-base-chinese. The results show that lexical mediators cannot be directly equated with semantic categories and that TF-IDF achieves the best overall performance, suggesting that lightweight character-level representations remain robust for very short idioms in low-resource settings. The study contributes an interpretable RDF/OWL-compatible resource for culture-aware modeling of Chinese idioms.

1 Introduction

Chinese 成语 *chéngyǔ*, a type of phraseological units & idiomatic multiword expression (MWE) (Polguère, 2002; Constant, 2012; Constant et al., 2017; Chen, 2021; Savary et al., 2023), pose particular challenges for natural language processing (NLP) because their meaning cannot usually be inferred from the literal meaning of their components. These fixed idioms, typically composed of four Chinese characters, often encode historical narratives, cultural values and metaphorical conceptualizations (Cowie, 1998; Granger and Meunier, 2008; Mel'čuk, 2023; Polguère, 2002, 2014; Mejri, 2018; Chen, 2021). For example, the *chéngyǔ* 画蛇添足 *huà shé tiān zú* "to draw legs on a snake" does not denote a concrete action involving a snake, but refers to adding something unnecessary. In many *chéngyǔ*, body-part and animal lexemes play an important metaphorical role. Lexemes such as 心 *xīn* "heart/mind", 口 *kǒu* "mouth", 马 *mǎ* "horse" or 狗 *gǒu* "dog" frequently function as conceptual mediators linking concrete lexical material to more abstract semantic domains, including emotion, cognition, morality, speech, behaviour and social relations. However, these lexical components cannot be directly equated with semantic categories. The same keyword may appear in several idioms associated with different interpretations. This observation raises an important question for NLP and computational phraseology: to what extent can semantic categories of idioms be inferred from lexical cues alone? Unlike explicit factual relations typically addressed in relation extraction and knowledge graph construction (Dessí et al., 2025; Zhao et al., 2024; Stanovsky et al., 2018), idioms involve semantic relations that are not directly encoded by surface syntax or lexical composition. In expressions such as "lend a hand", the relevant interpretation corresponds to "help" rather than to a literal action involving a hand. Chinese *chéngyǔ* exhibit similar forms of semantic mediation, where lexical components function as metaphorical cues without determining the final idiomatic meaning. Most existing approaches either rely on manually constructed lexical resources or focus on surface-level distributional representations, without explicitly distinguishing between lexical mediators and idiomatic semantic interpretation. Structured semantic representations and graph-based knowledge modeling have nevertheless become increasingly important in NLP and Semantic Web research (Hogan et al., 2021, 2022; Choi and Jung, 2025). In this work, we propose a semi-automatic lexico-semantic framework for Chinese *chéngyǔ* that combines manual semantic annotation, RDF/OWL-compatible graph formalization, and semantic classification. Based on 440 annotated expressions normalized into 18 cat-

Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026), pages 150–160

egories, we compare a keyword-based baseline, character n-gram TF-IDF with logistic regression, and BERT-base-chinese. The results show that lexical mediators are not reliable predictors of semantic categories, while character-level TF-IDF remains robust for short idioms in low-resource settings.

2 Related work

2.1 Ontology-oriented modeling of idioms

The idea of representing phraseological units & MWEs in ontological form is not new (Chen, 2025). Ran et al. (2010) proposed a manually constructed ontology of Chinese *chéngyǔ* using a skeleton-based method. This line of work shows that *chéngyǔ* can be modeled as structured semantic objects rather than merely as lexicalized expressions. However, such approaches remain largely handcrafted: they rely on expert-designed structures, predefined semantic relations and manually organized conceptual categories. This limits scalability (Choi and Jung, 2025) and makes it difficult to evaluate whether the proposed categories generalize beyond the constructed resource. More recent work has explored the ontological modeling of phraseological units & MWEs in RDF/OWL-based lexical resources. In monolingual contexts, Chen (2025) investigates the formal representation of French phraseological units & MWEs from the perspective of digital lexicography and ontophraseology. In bilingual contexts, Chen and Gasparini (2025) examine Chinese–French phraseological modeling, while Chen et al. (2025) extend this perspective to multilingual and trilingual resources involving Chinese, French and Vietnamese. These studies demonstrate the relevance of RDF/OWL structures for representing phraseological knowledge across languages, but they focus primarily on resource modeling rather than on controlled semantic classification or model evaluation. In the broader field of knowledge representation, RDF and OWL provide standard mechanisms for modeling entities, properties and semantic relations in interoperable graph-based formats (Hogan et al., 2021, 2022; McCrae et al., 2017). Such formats are useful for lexical resources because they allow idioms, lexical components and semantic categories to be explicitly linked and queried. However, applying RDF/OWL modeling to idioms raises a specific challenge: the semantic relation between an idiom and its interpretation is not directly compositional. A lexical component such as 心 *xīn* "heart/mind" may function as a mediator, but it does not determine the final idiomatic meaning by itself. The present work therefore adopts an operational and evaluative perspective. Rather than claiming to construct a complete ontology of Chinese *chéngyǔ*, we propose a semi-automatic lexico-semantic modeling framework based on manually validated semantic annotation, graph-based RDF/OWL formalization and controlled semantic classification. This allows us to test whether lexical mediators can predict idiomatic semantic categories, rather than assuming that keyword-to-category mappings are sufficient.

2.2 Semantic classification of short idioms

The semantic classification of *chéngyǔ* also raises methodological issues related to short-text classification. *chéngyǔ* are usually composed of only four characters, which means that models have very little textual context available. This makes the task different from sentence-level or document-level classification.

Traditional statistical representations such as TF-IDF remain useful in this setting because character n-grams can capture recurring morphographic patterns in very short expressions. Related methods such as PMI, C-value and rank fusion have also been widely used to identify salient lexical units and distributional regularities in specialized corpora. Although these methods do not provide deep semantic interpretation by themselves, they offer strong and interpretable baselines for short and highly lexicalized units. By contrast, neural models such as BERT and RoBERTa have achieved strong results in many semantic classification and relation extraction tasks (Devlin et al., 2019; Liu et al., 2019; Sun et al., 2021; Zhao et al., 2024). However, their effectiveness depends heavily on the size, balance and contextual richness of the annotated data. Their interpretability also remains limited (Rogers et al., 2020). These limitations are directly relevant to *chéngyǔ* classification, where the dataset is small, the labels are imbalanced and the input expressions are extremely short. For this reason, the present study compares symbolic, statistical and neural approaches under the same controlled evaluation setting: a rule-based keyword baseline, a character n-gram TF-IDF model with logistic regression, and BERT-base-chinese. This comparison allows us to evaluate whether idiomatic semantic categories can

be predicted from lexical mediators alone, from surface-level character patterns, or from contextualized neural representations.

2.3 Positioning of the present work

Compared with previous work, our contribution is not the construction of a complete ontology of Chinese *chéngyǔ*, nor a general framework for automatic relation extraction (Stanovsky et al., 2018). Instead, we propose a semi-automatic lexico-semantic modeling framework based on manually validated semantic annotation and lightweight RDF/OWL formalization. The study combines three components: (1) a controlled semantic annotation framework for Chinese *chéngyǔ* containing body-part and animal lexemes; (2) a comparative evaluation of symbolic, statistical and neural classification approaches; and (3) an RDF/OWL-compatible graph representation linking idioms, lexical mediators and semantic categories. From the perspective of culture-aware NLP, this work provides a interpretable case study of how culturally grounded phraseological knowledge can be represented, annotated and evaluated beyond purely compositional or distributional approaches.

3 Data and annotation

This section describes the resources and annotation framework used in the experiments. The semantic categories employed in this study are not automatically derived from lexical keywords. Instead, they are assigned through interpretative annotation of the global idiomatic meaning of each *chéngyǔ*. This distinction is central to the study, since the same lexical mediator may correspond to different semantic interpretations depending on the idiomatic context.

3.1 Lexical corpus of *chéngyǔ*

The first resource consists of a closed lexical corpus of 2,210 Chinese *chéngyǔ* manually compiled from specialized monolingual and bilingual dictionaries. The corpus focuses on two major semantic domains: body-part lexemes and animal lexemes. This controlled thematic subset provides the basis for the semantic and graph-based analyses conducted in the study. For example, expressions such as 心高气傲 *xīn gāo qì ào* contain body-related morphemes such as 心 *xīn* “heart/mind”, while 虎视眈眈 *hǔ shì dāndān* belongs to the animal domain through the lexeme 虎 *hǔ* “tiger”. This lexical corpus serves as the starting point for semantic annotation and RDF/OWL formalization.

3.2 Annotated *chéngyǔ* dataset

From the lexical corpus, a manually annotated dataset of 440 *chéngyǔ* was constructed for semantic classification and evaluation. Each entry includes the idioms, a lexical keyword (*mot_clef*), its literal category (e.g., body part or animal), a primary semantic label, optional secondary labels, and annotation metadata. The annotation process explicitly distinguishes lexical mediators from semantic categories. Semantic labels are assigned on the basis of idiomatic interpretation rather than literal lexical meaning. For example, 赤口毒舌 *chìkǒu dúshé* “venomous mouth and poisonous tongue” contains body-related lexemes but is interpreted through both speech and moral evaluation. Table 1 presents several examples of annotated *chéngyǔ*.

Chengyu	Keyword	Literal category	Secondary labels
闭口不言 (*bì kǒu bù yán*)	口 (mouth)	身体部位 (Body part)	言语; 表达 (Speech; expression)
牛马生活 (*niú mǎ shēnghuó*)	牛; 马 (ox; horse)	动物 (Animal)	劳累; 压迫; 辛苦 (Exhaustion; oppression; hardship)

Table 1: Examples of annotated *chéngyǔ*.

These examples illustrate the central hypothesis of the study: lexical mediators function as semantic cues, but they cannot be directly equated with idiomatic semantic categories.

3.3 Contemporary media corpus

In addition to the annotated dataset, a contemporary media corpus covering the period 2015–2025 was compiled in order to verify the attestation and circulation of the selected *chéngyǔ* in recent Chinese discourse. The corpus contains 1,286 news articles collected from three major Chinese media sources: *People's Daily, Xinhua News and The Paper*. The articles were collected through targeted web scraping and stored in structured CSV format. The corpus is not used to define semantic categories; instead, it serves as a contextual resource for validating contemporary usage and observing the diversity of semantic environments in which lexical mediators occur.

3.4 Annotation guidelines and reliability

The annotation process followed explicit guidelines distinguishing lexical mediators from idiomatic semantic categories. Annotators were instructed not to assign semantic labels solely on the basis of literal lexical components such as 心 *xīn* "heart/mind", 口 *kǒu* "mouth" or 马 *mǎ* "horse", but to consider the global idiomatic meaning of each expression. When multiple semantic dimensions were activated, a primary label and optional secondary labels were assigned. The dataset was manually validated by the main author following iterative normalization and consistency checking procedures. Although no full-scale inter-annotator agreement was computed in the current version, the annotation process was designed to maintain semantic consistency across the resource. Future work will extend the annotation protocol to multiple annotators and evaluate agreement using standard measures such as Cohen's kappa or Krippendorff's alpha.

4 Methodology

4.1 Experimental pipeline

The proposed framework consists of six main stages, illustrated in Figure 1: (1) lexical resource construction, (2) manual semantic annotation and category normalization, (3) lexico-semantic graph construction, (4) RDF/OWL formalization, (5) symbolic, statistical and neural semantic classification, and (6) evaluation and qualitative error analysis. This pipeline allows us to investigate whether idiomatic semantic categories can be inferred from lexical mediators, character-level surface patterns, or contextualized neural representations.

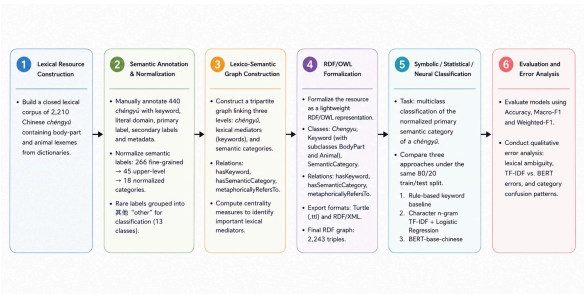

Figure 1: Overview of the experimental pipeline for culture-aware lexico-semantic modeling and neural semantic classification of Chinese *chéngyǔ*.

4.2 Category normalization

After annotation, the dataset contained a highly fine-grained label inventory. The 440 annotated *chéngyǔ* included 266 distinct primary semantic labels. Although this granularity captures rich semantic distinctions, it is not directly suitable for supervised classification because many labels occur only once or only a few times, increasing sparsity and class imbalance. To obtain a more stable and interpretable label space, we applied a two-step normalization procedure. First, hierarchical labels were reduced to their upper-level category. Labels encoded in hierarchical form using the separator "/" were simplified by retaining only the higher-level semantic category. For example, labels such as 社会/关系行为 *shèhuì/guānxì xíngwéi* "society/relational behaviour" and 认知/决策变化 *rènzhī/juécè biànhuà* "cognition/decision change" were reduced to 社会 *shèhuì* "society" and 认知 *rènzhī* "cognition". This first step reduced the number of semantic categories from 266 to 45. Second, semantically related categories were manually merged into broader conceptual domains. For instance, labels related to social identity or social evaluation were grouped under 社会 *shèhuì* "society"; affective and psychological labels were grouped under 情感 *qínggǎn* "emotion/affect"; and labels related to action, intention or strategy were grouped under 行为 *xíngwéi* "behaviour/action". After normalization, the final inventory contains 18 semantic categories. The distribution of the most frequent categories is shown in Table 2. This distribution shows that the semantic interpretation of *chéngyǔ* is primarily organized around abstract domains such as social relations, cognition, behaviour and emotion, rather than around literal body-part or animal meanings.

Category	Occurrences
社会 "society"	89
行为 "behaviour/action"	72
认知 "cognition"	63
情感 "emotion/affect"	57
言语 "speech/language"	45

Table 2: Distribution of the most frequent semantic categories after normalization.

The resource also preserves multi-label information. Some expressions activate several semantic dimensions simultaneously. For example, 赤口毒舌 *chìkǒu dúshé* "venomous mouth and poisonous tongue" combines speech and moral evalu-

ation, while 集腋成裘 *jíyè chéngqiú* “to accumulate small contributions into a larger result” combines accumulation and action. After normalization, 417 expressions have one label, 20 have two labels, and 3 have three labels. For the classification experiments, however, we formulate the task as a multiclass problem based on the normalized primary label. Categories with fewer than five occurrences were grouped into a single class named 其他 *qítā* “other” in order to reduce sparsity and enable stratified train/test splitting. As a result, the full resource retains 18 semantic categories, while the classification experiments are conducted on 13 classes. The final dataset is split into 352 training examples and 88 test examples using a stratified 80/20 split.

4.3 Lexico-semantic graph construction

After category normalization, the annotated dataset was represented as a directed lexico-semantic graph linking three levels of description: *chéngyǔ* as phraseological units & MWEs, lexical mediators, and semantic categories.

Each annotated entry generates three types of nodes: (1) a *chéngyǔ* node; (2) a keyword node corresponding to a body-part or animal lexeme; and (3) a semantic category node. Three directed relations are defined:

- hasKeyword
- hasSemanticCategory
- metaphoricallyRefersTo

The resulting graph contains 566 nodes and 1,079 edges, including 440 *chéngyǔ* nodes, 108 keyword nodes and 18 semantic category nodes. Since many expressions share the same lexical mediators and semantic categories, the graph reveals a highly interconnected semantic structure. To identify the most important lexical mediators, we computed degree and betweenness centrality measures on keyword nodes. The most central keywords are 心 *xīn* “heart/mind”, 口 *kǒu* “mouth”, 马 *mǎ* “horse”, 牛 *niú* “ox”, 狗 *gǒu* “dog” and 鸡 *jī* “chicken”. The keyword 心 *xīn* “heart/mind” exhibits the highest degree centrality, followed by 口 *kǒu* “mouth” and 马 *mǎ* “horse”.

Keyword	Degree	Betweenness
心 xīn “heart/mind”	123	0.003191
口 kǒu “mouth”	68	0.001638
马 mǎ “horse”	55	0.001155
牛 niú “ox”	27	0.000417
狗 gǒu “dog”	20	0.000220
鸡 jī “chicken”	20	0.000264

Table 3: Centrality of the most frequent keyword nodes.

These results suggest that body-part and animal lexemes function as important lexical mediators connecting idioms to multiple semantic domains. In particular, 心 *xīn* “heart/mind” is associated with emotional, cognitive and moral interpretations across a wide range of *chéngyǔ*, while 口 *kǒu* “mouth” is strongly connected to speech, social interaction and verbal evaluation. Rather than constituting a complete ontology of Chinese idiomatic knowledge, the graph provides an interpretable lexico-semantic representation of culturally grounded phraseological relations.

5 Classification experiments

This section evaluates whether the semantic category of a *chéngyǔ* can be predicted automatically from its lexical form. The objective is not to maximize classification performance, but to compare how lexical keywords, character-level patterns and contextualized neural representations capture idiomatic semantic information.

5.1 Task definition

The task is formulated as a multiclass classification problem: given a *chéngyǔ*, the model predicts its normalized primary semantic category. Although the resource preserves multi-label information, the experiments use the primary label for controlled comparison across models. After category normalization and grouping of rare labels into the class 其他 *qítā* (“other”), the classification setup contains 13 semantic classes. The dataset is split into 352 training examples and 88 test examples using a stratified 80/20 split. Because the dataset remains imbalanced, we report accuracy, macro-F1 and weighted-F1.

5.2 Models

We compare three classification approaches under the same train/test setting.

Rule-Based keyword Baseline The rule-based baseline assigns to each keyword the most frequent semantic category observed in the training

data. At test time, the model predicts the category associated with the lexical keyword of the input *chéngyǔ*. Unseen keywords are mapped to the majority class.

TF-IDF + Logistic regression The second model represents each *chéngyǔ* using character n-grams of length 1 to 4 combined with TF-IDF weighting. Classification is performed using multiclass logistic regression with balanced class weights.

BERT-base-chinese The third model fine-tunes BERT-base-chinese for sequence classification. The model uses 13 output labels, a maximum sequence length of 16, a batch size of 16, a learning rate of 2e-5 and 8 training epochs. Model selection is based on macro-F1.

5.3 Results

Model	Acc.	Macro-F1	Weighted-F1
Rule baseline	0.273	0.190	0.231
TF-IDF + Logistic Regression	0.443	0.351	0.406
BERT-base-chinese	0.330	0.173	0.286

Table 4: Classification results on the test set.

The rule-based baseline obtains the lowest overall performance, confirming that lexical mediators cannot be directly equated with semantic categories. A lexical component such as 心 *xīn* "heart/mind" may occur in idioms associated with emotion, cognition, morality or social relations, making deterministic keyword-to-category mappings insufficient. The TF-IDF model achieves the best performance across all metrics. These results suggest that character-level representations remain robust for very short idioms, where local morphographic patterns carry important semantic information. BERT-base-chinese performs better than the rule-based baseline in terms of accuracy and weighted-F1, but does not outperform TF-IDF. This result is likely related to the small size, class imbalance and context-free nature of the dataset, since the input expressions are extremely short and no contextual examples or glosses are provided to the model.

The experiments show that idiomatic semantic interpretation cannot be reduced to simple lexical keyword matching and that lightweight character-based representations may remain more robust than large contextual neural models for low-resource classification of very short idioms.

6 Error analysis

To better understand the behaviour of the different models, we conducted a qualitative error analysis on the test set. For each *chéngyǔ*, we compared the gold category with the predictions produced by the rule-based baseline, the TF-IDF model and BERT-base-chinese. The analysis focuses on lexical ambiguity, category overlap and model-specific limitations.

6.1 Lexical ambiguity and limits of the Rule-Based Baseline

The rule-based baseline is strongly dependent on the dominant semantic category associated with each lexical keyword in the training data. This leads to systematic errors when the same keyword appears in idioms with different semantic interpretations. For example, the keyword 心 *xīn* "heart/mind" is frequently associated with 情感 *qínggǎn* "emotion/affect" in the training set. As a result, the rule-based model predicts 情感 *qínggǎn* "emotion/affect" for expressions such as 痴心妄想 *chīxīn wàngxiǎng* "to indulge in unrealistic fantasies", 三心二意 *sānxīn èryì* "to be of two minds" and 包藏祸心 *bāocáng huòxīn* "to harbour evil intentions", although their gold labels correspond to 认知 *rènzhī* "cognition" or 道德 *dàodé* "morality". These errors confirm that lexical mediators cannot be treated as equivalent to semantic categories. The semantic interpretation of a *chéngyǔ* depends on the idiomatic meaning of the expression as a whole rather than on isolated lexical components.

6.2 Comparative error patterns between TF-IDF and BERT

We identified 17 cases in which TF-IDF predicts the correct category while BERT fails. These examples show that character n-gram representations remain highly effective for short and fixed idioms. For example, 口诛笔伐 *kǒuzhū bǐfá* "to denounce verbally and in writing" is correctly classified by TF-IDF as 言语 *yányǔ* "speech/language", while BERT predicts 行为 *xíngwéi* "behaviour/action". Similarly, 众口一词 *zhòngkǒu yīcí* "everyone says the same thing" is correctly classified by TF-IDF as 言语 *yányǔ* "speech/language", whereas BERT predicts 社会 *shèhuì* "society". Conversely, BERT outperforms TF-IDF in 7 cases,

especially when local lexical cues are misleading. For example, 笑口常开 *xiàokǒu chángkāi* is correctly classified by BERT as 情感 *qínggǎn* "emotion/affect", while TF-IDF predicts 言语 *yányǔ* "speech/language" because of 口 *kǒu* "mouth".

6.3 Main confusion patterns

The analysis reveals several recurring confusion patterns between semantically related categories.

First, 情感 *qínggǎn* "emotion/affect" is frequently confused with 认知 *rènzhī* "cognition", especially in idioms containing 心 *xīn* "heart/mind", which may refer to emotion, intention, thought or judgement depending on the expression. Second, 社会 *shèhuì* "society" is often confused with 行为 *xíngwéi* "behaviour/action", because many *chéngyǔ* describe actions embedded in social situations or social evaluation. Third, 言语 *yányǔ* "speech/language" may overlap with 社会 *shèhuì* "society" in expressions involving public opinion, collective judgement or social interaction, such as 众口一词 *zhòngkǒu yīcí* "everyone says the same thing". Finally, 道德 *dàodé* "morality" frequently overlaps with 情感 *qínggǎn* "emotion/affect", particularly in expressions involving intention, sincerity or moral judgement.

These confusion patterns suggest that the semantic categories are not strictly disjoint. Many *chéngyǔ* simultaneously activate cognitive, emotional, social and moral dimensions, reinforcing the importance of preserving multi-label semantic information in the resource.

7 RDF/OWL representation and visualization

The annotated resource is formalized as a lightweight RDF/OWL lexico-semantic representation linking *chéngyǔ*, lexical mediators and semantic categories. The objective of this formalization is not to construct a complete ontology of Chinese idiomatic knowledge, but to provide an interoperable semantic structure compatible with Semantic Web and graph-based NLP environments.

Each *chéngyǔ* is represented as an instance of the class Chengyu. Lexical mediators are represented as instances of Keyword, with the subclasses BodyPart and Animal used when the keyword belongs to one of these semantic domains. Normalized semantic labels are represented as instances of SemanticCategory.

The resource defines three main object properties:

- hasKeyword, linking a *chéngyǔ* to its lexical mediator;
- hasSemanticCategory, linking a *chéngyǔ* to its interpreted semantic category;
- metaphoricallyRefersTo, linking a lexical mediator to an abstract semantic category.

This representation follows the tripartite structure used throughout the study: Chengyu → Keyword → SemanticCategory, while preserving the distinction between lexical mediators and idiomatic semantic interpretation. Figure 2 illustrates the simplified RDF/OWL schema underlying the resource.

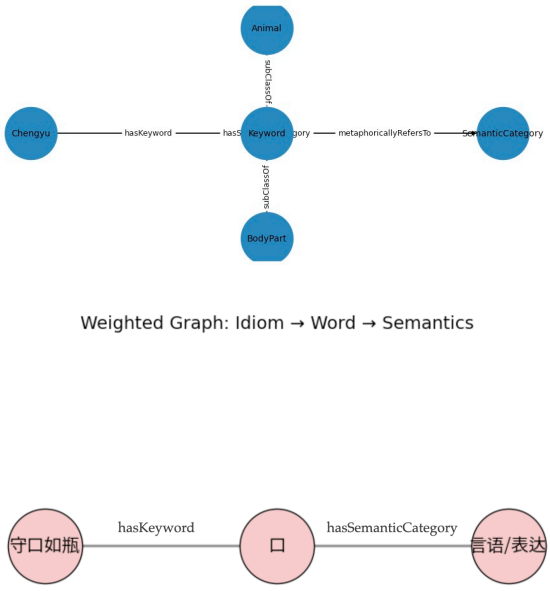

Figure 2: Simplified RDF/OWL schema linking chéngyǔ, lexical mediators and semantic categories. Bottom: Streamlit-based interface for exploring the annotated chéngyǔ resource.

The final RDF graph contains 2,243 triples, including class declarations, instance types, labels and object-property assertions. The resource is exported in Turtle (.ttl) and RDF/XML formats, making it compatible with Semantic Web tools such as Protégé, RDFLib and SPARQL-based environments.

To facilitate qualitative exploration of the resource, we developed an interactive Streamlit-based interface integrating NetworkX and PyVis for graph visualization. The interface enables

users to inspect the annotated data, visualize summary statistics, filter entries by lexical mediator or semantic category, and explore semantic relations between idioms and abstract conceptual domains. Due to space limitations, screenshots of the Streamlit interface, graph visualizations, and RDF/Turtle examples are provided in the Appendix.

Rather than constituting a complete ontology of Chinese cultural concepts, the resulting graph should be understood as an interpretable lexico-semantic representation of culturally grounded phraseological knowledge.

8 Discussion

This study explored how culturally grounded Chinese *chéngyǔ* can be represented, classified and formalized through a combination of manual semantic annotation, lexico-semantic graph modeling and lightweight RDF/OWL representation. The results provide several methodological observations concerning lexical mediation, semantic ambiguity and the limits of surface-based semantic modeling for idioms.

8.1 Lexical mediators and semantic interpretation

One of the central findings of this work is that lexical mediators cannot be directly equated with semantic categories. The weak performance of the rule-based keyword baseline confirms that a lexical component such as 心 *xīn* "heart/mind" or 马 *mǎ* "horse" does not determine the final semantic interpretation of a *chéngyǔ*. Instead, lexical mediators function as partial metaphorical cues whose interpretation depends on the idiomatic context. For example, 心 *xīn* may refer to emotion, cognition, morality or intention depending on the expression in which it occurs, while 口 *kǒu* "mouth" may be associated with speech, persuasion, social interaction or verbal aggression. These results support the distinction introduced throughout the paper between: (1) the lexical mediator as a concrete morpho-semantic cue, and (2) the semantic category as an interpreted idiomatic meaning. The semantic interpretation of *chéngyǔ* is therefore not reducible to simple keyword matching. Rather, it involves culturally grounded conceptual interpretation in which lexical forms act as metaphorical entry points rather than deterministic semantic labels.

8.2 Character-Level patterns and neural limitations

A second important result concerns the comparison between character-based statistical models and contextual neural models. TF-IDF combined with Logistic Regression achieves the best overall quantitative performance, outperforming BERT-base-chinese in terms of accuracy, macro-F1 and weighted-F1. Although Transformer models generally achieve strong results in many NLP tasks, several characteristics of the present task explain this outcome. First, *chéngyǔ* are extremely short expressions, usually composed of only four Chinese characters. Character n-grams therefore capture stable morphographic patterns particularly well. Second, the dataset remains relatively small and imbalanced, with only 352 training examples distributed across 13 semantic categories. Finally, the experiments intentionally evaluate isolated surface forms without contextual examples, glosses or definitions. Under these conditions, lightweight character-level representations remain highly competitive. At the same time, the qualitative analysis shows that BERT occasionally captures global idiomatic meaning better than TF-IDF when local lexical cues are misleading. This suggests that contextual neural models may become more effective when enriched with definitions, usage examples or broader discourse contexts.

8.3 Multi-Dimensional semantics of *chéngyǔ*

The error analysis also highlights the semantic complexity of Chinese idioms. Many *chéngyǔ* simultaneously activate emotional, cognitive, social and moral dimensions, producing recurrent overlaps between categories such as:

情感 *qínggǎn* "emotion" and 认知 *rènzhī* "cognition",

社会 *shèhuì* "society" and 行为 *xíngwéi* "behaviour/action",

言语 *yányǔ* "speech/language" and 社会 *shèhuì* "society",

道德 *dàodé* "morality" and 情感 *qínggǎn* "emotion".

These overlaps suggest that idiomatic meaning cannot always be reduced to a single semantic dimension. The resource therefore preserves multi-label semantic information even though the classification experiments are conducted in a multi-class setting for controlled comparison purposes. These observations reinforce the idea that *chéngyǔ*

encode culturally grounded conceptual structures rather than purely compositional lexical meanings. Their interpretation frequently depends on shared metaphorical and social knowledge embedded in Chinese linguistic culture.

8.4 Implications for Culture-Aware NLP

The proposed resource provides an interpretable framework for representing culturally grounded phraseological knowledge in NLP systems. By explicitly separating idioms, lexical mediators and semantic categories, the RDF/OWL formalization models how culturally salient lexical elements contribute to idiomatic interpretation without reducing meaning to surface lexical forms alone. More broadly, the results suggest that culture-aware NLP benefits from combining manual semantic interpretation, lightweight semantic formalization and interpretable lexical representations.

9 Conclusion

This study presented a lexico-semantic resource for Chinese *chéngyǔ* containing body-part and animal lexemes, together with a lightweight RDF/OWL formalization and a comparison of symbolic, statistical and neural classifiers. The results show that lexical mediators should be distinguished from idiomatic semantic interpretation, and that character-level TF-IDF outperforms BERT-base-chinese in this low-resource, short-expression setting. Future work may extend the resource with richer contextual information, multilingual alignments and larger-scale graph-based semantic representations.

Limitations

This study has several limitations. First, the annotated dataset remains relatively small, with class imbalance across some semantic categories, which limits the evaluation of neural models. Second, the experiments are conducted on isolated *chéngyǔ* surface forms without contextual examples or definitions, a setting that is disadvantageous for contextual models such as BERT-base-chinese. Third, the resource focuses specifically on *chéngyǔ* containing body-part and animal lexemes, which limits its semantic coverage. Fourth, the semantic normalization process involves manual decisions, and no full-scale inter-annotator agreement was computed in the current version. Finally, the RDF/OWL formalization remains intentionally lightweight and does not yet encode richer conceptual or historical semantic relations.

References

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. 2025. Modeling and structuring of a bilingual french–chinese phraseological dictionary: Neural automatic approach for ontology and lexicography. In *Proceedings of eLex 2025: Electronic Lexicography in the 21st Century – Intelligent Lexicography*, pages 830–851. Lexical Computing CZ s.r.o.

Lian Chen and Nicolas Gasparini. 2025. Modélisation et structuration d'un dictionnaire bilingue français-chinois d'expressions idiomatiques à l'aide du modèle ontolex: Une approche lexicographique et ontologique. In Giovanni Dotoli, Salah Mejri, and Philippe Golda, editors, *Les Cahiers du dictionnaire: Dictionnaire et Polylexicalité / Dictionary and Polylexicality*, volume 2, pages 335–357. Classiques Garnier, Paris.

Lian Chen, Nicolas Gasparini, H.-L. Dao, and D.-T. Do-Hurinville. 2025. Toward a trilingual ontology of phraseological units: Lexicographic and computational modeling in chinese, french, and vietnamese. In *Proceedings of AsiaLex 2025: The 18th International Conference of the Asian Association for Lexicography*, pages 13–21.

S. Choi and Y. Jung. 2025. Knowledge graph construction: Extraction, learning, and evaluation. *Applied Sciences*, 15(7):3727.

Mathieu Constant. 2012. Mettre les expressions multimots au coeur de l'analyse automatique de textes: Sur l'exploitation de ressources symboliques externes. Habilitation à diriger des recherches, Université Paris-Est.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing—a survey. *Computational Linguistics*, 43(4):837–892.

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

Danilo Dessí, Francesco Osborne, Davide Buscaldi, Diego Reforgiato Recupero, and Enrico Motta. 2025. Cs-kg 2.0: A large-scale knowledge graph of computer science. *Scientific Data*, 12:964.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

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

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, and 1 others. 2022. *Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool.

Aidan Hogan, Eva Blomqvist, Michael Cochez, and 1 others. 2021. Knowledge graphs. *ACM Computing Surveys*, 54(4):1–37.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, and 1 others. 2017. The ontolex-lemon model: Developments and applications. In *Proceedings of eLex 2017*, pages 587–597.

Salah Mejri. 2018. La phraséologie: cotexte, contexte et contenus culturels. *Modern Languages and Literature*, 42(4):12–38.

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

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.

Jun Ran, Yu Sun, X. Chang, X.-J. Zhang, and J. Li. 2010. Research of construction of the idiom story ontology based on owl. *Computer Technology and Development*, 20(5):63–66.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and 1 others. 2023. Parseme meets universal dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of NAACL-HLT 2018*, pages 885–895.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxian Lu, Weipeng Liu, Zijun Wu, Wei Gong, Jing Liang, Zhengyan Shang, Peiyuan Sun, Wei Liu, Xi Ouyang, Dong Yu, and 3 others. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Xiang Zhao, Yong Deng, Minghao Yang, Liang Wang, Rui Zhang, Hao Cheng, Wai Lam, Yelong Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *Preprint*, arXiv:2306.02051.

A Interactive Visualization and RDF/OWL Resources

A lightweight interactive interface was developed using Streamlit, NetworkX, PyVis and RDFLib to facilitate qualitative exploration of the lexico-semantic resource.

The interface is built from the cleaned annotation dataset containing 440 annotated *chéngyǔ*, 108 lexical mediators and 18 semantic categories. Users can filter the graph by lexical keyword (e.g., 心 xīn "heart/mind", 口 kǒu "mouth", 马 mǎ "horse") or by semantic category (e.g., 情感 qínggǎn "emotion", 认知 rènzhī "cognition", 社会 shèhuì "society").

The visualization displays three node types: (i) *chéngyǔ*, (ii) lexical mediators, (iii) semantic categories.

The same relations used in the RDF/OWL formalization (hasKeyword, hasSemanticCategory and metaphoricallyRefersTo) are represented interactively through PyVis-based graph visualization. The resource can also be exported in Turtle (.ttl) and RDF/XML (.owl) formats, making it compatible with Semantic Web tools such as Protégé, RDFLib and SPARQL-based environments.

Figures 3 and 4 illustrate the interactive visualization environment developed for the lexico-semantic resource. Figure 3 presents the Streamlit-based exploration interface with summary statistics, filtering options and annotated *chéngyǔ* entries, while Figure 4 shows the PyVis graph visualization of the RDF-based lexico-semantic network. Figure 5 provides a Turtle (.ttl) excerpt illustrating how *chéngyǔ* instances are linked to lexical mediators and semantic categories through the relations hasKeyword, hasSemanticCategory and metaphoricallyRefersTo.

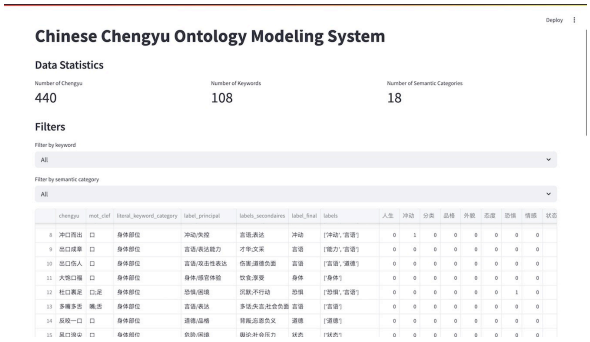

Figure 3: Interactive interface showing summary statistics, filtering options and annotated *chéngyǔ* entries.

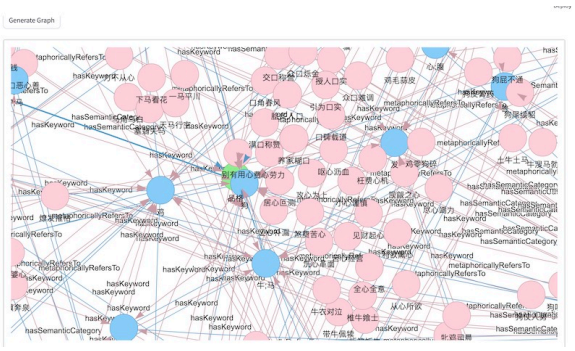

Figure 4: Interactive PyVis visualization of the RDF-based *chéngyǔ* lexico-semantic network.

In Figure 4, pink nodes represent *chéngyǔ* (MWEs), blue nodes represent lexical mediators (keywords), and green nodes represent semantic categories. Edges correspond to RDF-based semantic relations between nodes. Labels are displayed for all nodes to ensure readability independently of color.

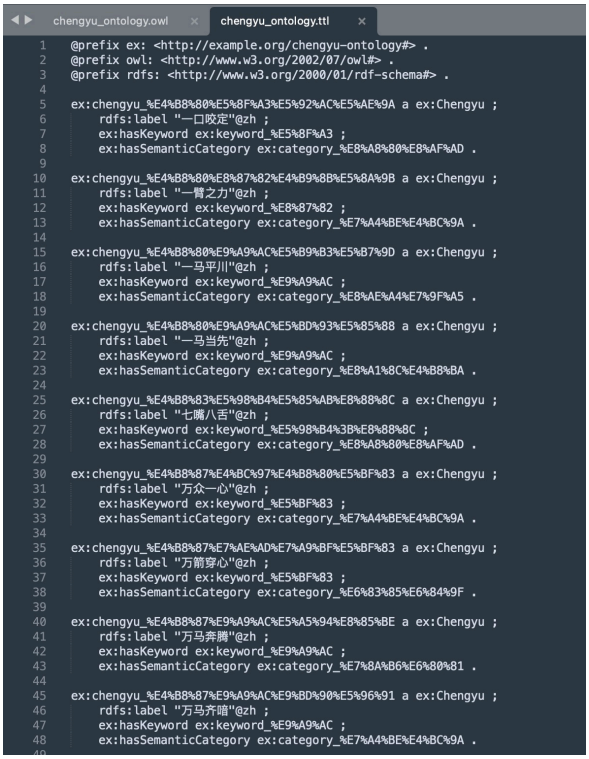

Figure 5: Turtle (.ttl) excerpt illustrating RDF relations between *chéngyǔ*, lexical mediators and semantic categories.