Modeling and structuring of a bilingual French-Chinese phraseological dictionary: neural automatic approach for ontology and lexicography

Lian Chen 陈恋 [1, 2]

[1] LLL- University of Orleans, France
[2] CRLAO- CNRS-INALCO, FRANCE
E-mail: lian.chen@univ-orleans.fr

Abstract

This project builds a French–Chinese bilingual dictionary of idioms (e.g., *avoir la main heureuse* [to have a lucky hand]), which are considered phrasemes (Mel'čuk, 2011) or phraseological units (Mel'čuk, 2008; Mejri, 2011; Sułkowska, 2016; Chen, 2021). The idioms are grouped by themes such as the human body, animals, plants, or numbers. We focus on the concepts human body and animal, both in literal and metaphorical uses. For instance, the hand can mean a body part, a tool for work, or a symbol of power.
To link idioms to their meanings, we use an ontology-based method rather than manual tools like *Protégé*. Following the Ontology Layer Cake model (Despres & Szulman, 2008; Tiwari & Jain 2014), we apply a step-by-step automated process to a specialized corpus: 1) Idioms are extracted using statistical methods (TF-IDF, PMI, RAKE) and tagged via a *Streamlit* interface. 2) Co-occurring words help build a weighted graph of relations. 3) AI models (e.g., BERT) classify the links by meaning. 4) The interface supports sorting, export (OWL/RTF), graph viewing (PyVis), and timing. 5) Finally, the OntoLex-Lemon model is used to generate an RDF/OWL bilingual dictionary.
Keywords: auto-lexicography; ontological relations automation; knowledge engineering; natural language processing; e-lexicography

1. Introduction: Automating the creation of ontologies and phraseology

The creation of ontologies, long reserved for linguists and knowledge modeling specialists, is currently undergoing a major transformation thanks to advances in artificial intelligence and natural language processing (NLP). This development opens up new perspectives for phraseology, a field in which multi-word expressions (MWEs), often opaque and non-compositional, need to be identified, structured, and linked to abstract concepts or specific discourse situations (Constant, 2012: 6).

Tools like *Protégé*[1] have enabled the formalization of ontologies according to Semantic Web standards (the Web Ontology Language, OWL, and the Resource Description Framework, RDF), ensuring the interoperability and reusability of resources. However, their use remains largely manual, time-consuming, and poorly suited to the complexity of linguistic phenomena — particularly fixed or metaphorical expressions (Kapoor & Sharma, 2010).

[1] https://protege.stanford.edu/

Therefore, scholars have long been exploring ontology learning and the automatic creation of ontologies. With the advancement of research in this field, various models have been proposed, such as the Ontology Layer Cake (Després & Szulman, 2008; Tiwari & Jain, 2014) and, more recently, OLAF (Ontology Learning Applied Framework)[2], introduced in France in 2023 (Schaeffer, Sesboüé et al., 2023). OLAF is a modular framework that automates the creation of ontologies from unstructured corpora. The automation of ontology creation is currently based on a set of well-defined steps, according to many researchers (Marco, 2007; Elnagar et al., 2020; Amdouni et al., 2025, etc.): (1) the extraction of salient terms from a corpus (concepts, attributes, relations), which, in phraseology, includes the detection of multi-word expressions (MWEs) such as fixed expressions, collocations, or support verb constructions; (2) filtering and specialization, by comparison with reference corpora, in order to isolate domain-specific terms, using techniques such as contrastive analysis, LSA ou Latent Semantic Analysis, or subsumption; (3) the structuring of semantic relations (synonymy, hyperonymy, cause, agent, etc.), where MWEs play a crucial role—for example, by linking "crack his pipe" to the concept "die" via a paraphrase relation; (4) the hierarchization and formalization of concepts according to is-a or part-of relations, then translated into RDF or OWL; (5) the validation and dynamic evolution of the ontology, with logical verification and progressive enrichment.

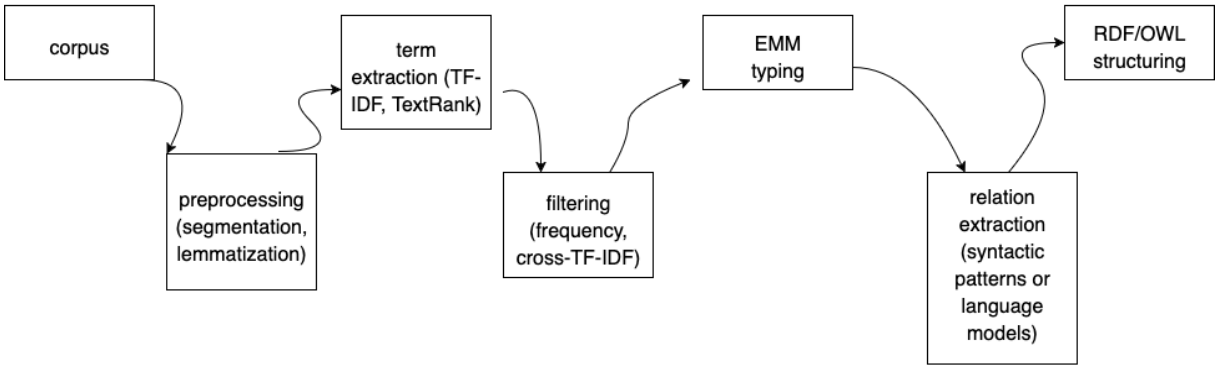

Figure 1: Typical computational lexicology and ontology-building pipeline

This process (see Figure 1) is based on a typical pipeline: corpus → preprocessing (segmentation, lemmatization) → term extraction (TF-IDF[3], TextRank[4]) → filtering

[2] https://github.com/wikit-ai/olaf

[3] TF-IDF (Term Frequency–Inverse Document Frequency): a weighting scheme that highlights terms frequent in a document but not frequent in the whole corpus.

[4] TextRank: a graph-based ranking algorithm (inspired by PageRank) that scores terms or keyphrases according to their co-occurrence relations.

(frequency, cross-TF-IDF) → EMM typing[5] → relation extraction (syntactic patterns or language models) → RDF/OWL structuring. The techniques used are diverse and complementary: (1) statistical (TF-IDF, PMI[6], C-value/NC-value[7], TextRank), (2) symbolic (syntactic patterns, grammars, lexicons), (3) machine learning (clustering, contextual embeddings, KeyBERT[8]), and (4) large-scale language models (LLMs), capable of directly generating RDF classes or triples from texts via prompting.

With the explosion of textual data available online, the automatic identification of relevant linguistic units has become a central issue in the field of natural language processing (NLP). Applications such as search engines, machine translation systems, and information retrieval systems rely on increasingly powerful linguistic analyzers. Phraseology, however, introduces specific challenges: idiomatic expressions are often non-compositional, subject to syntactic variation, and require several levels of analysis (lexical, syntactic, semantic, and pragmatic) (see the work of Gross, 1996; Mejri, 1997; Constant, 2012; Polguère, 2002; Chen, 2021, etc.). This category includes: collocations (e.g., *heavy weight*, *strong accent*), idiomatic expressions (*boire les paroles de quelqu'un* [to lap up what somebody says]), proverbs, named entities (e.g., *San Francisco*, *European Union*), specialized terms (e.g., *black hole* in astronomy), etc. To the machine, it is merely a sequence of words devoid of interpretation (Constant, 2012: 6). Their detection, classification, and integration into an ontology therefore require a hybrid approach, combining traditional linguistic tools, statistical methods, and recent advances in neural AI. Their adequate treatment would not only improve the syntactic and semantic analysis of texts, but also enhance the performance of downstream tasks such as translation or information retrieval.

A computer manipulates strings of characters without understanding their deeper meaning. Even embedding techniques, although effective in capturing contextual similarities, fail to represent the complex semantic relationships or figurative mechanisms specific to MWEs. This is where ontology plays a central role: by providing an explicit conceptual structure, it allows linguistic data to be linked to interpretable representations, thus facilitating reasoning, semantic annotation, or the inference of new knowledge.

[5] EMM typing (Entity–Mention Mapping/Typing): the step where candidate terms are normalized and disambiguated, then mapped to a canonical entity in a reference ontology/taxonomy and assigned a semantic type (e.g., Person, Organization, Event, or domain-specific classes). Typical methods include gazetteer/dictionary lookup, string/embedding similarity, and context-aware classifiers (NER + entity linking). Output: a stable ID/URI and an associated rdf:type for each mention.

[6] PMI (Pointwise Mutual Information): a statistical association measure indicating how strongly two words co-occur compared to chance.

[7] C-value / NC-value: methods for multiword term extraction; C-value favors longer, domain-specific terms, and NC-value refines this by considering context words.

[8] KeyBERT: a keyword extraction method using contextual embeddings from BERT to find terms semantically close to a document.

Phraseography (Murano, 2011; Chen, 2023), a branch of lexicography dedicated to the description and organization of idiomatic expressions and phraseological units (collocations, proverbs, fixed phrases, etc.), can now rely on advanced technologies from NLP. This evolution echoes that observed in general lexicography: since the 2010s, tools such as Sketch Engine, automatic semantic disambiguation techniques, or even semi-automatic models such as tickbox lexicography have made it possible to gradually transfer some of the tasks from the hands of lexicographers to machines. In this context, automatic phraseography plays a fundamental role. With the explosion of textual data available online, identifying complex linguistic units has become a priority for NLP systems. Often overlooked, multi-word expressions disrupt classic compositional processing and negatively affect the performance of tasks such as machine translation, information extraction, and document retrieval. These expressions—idiomatic, collocational, proverbial, terminological, or onomastic—function as stable lexical entities despite their internal complexity. Their explicit recognition not only simplifies syntactic and semantic analysis but also optimizes semantic alignment in multilingual contexts. Integrating phraseography into the ontological processing chain thus amounts to anchoring ontology in real linguistic usage while enriching knowledge structuring.

Thus, the objective of this project is to provide an interoperable and reusable resource that contributes to the ongoing DiCoP (Dictionary and Corpus of Phraseology) project (see Chen, 2023; 2024). The idiom ontology will be integrated into DiCoP as a module, ensuring open dissemination and interoperability with existing lexical infrastructures (OntoLex-Lemon lexicons, multilingual knowledge bases), and will be applicable to multilingual NLP tasks such as machine translation, semantic search, and cross-cultural teaching. By providing structured semantic links between idioms and their conceptual domains, this resource bridges phraseology, ontology, and AI, and offers a foundation for future work in digital lexicography and language technologies.

Building such an ontology therefore calls for a structured representation of idioms and their associated concepts, which in NLP is typically achieved through the extraction of relational triplets.

2. Understanding and automating relational triplets of phraseological units in NLP

In NLP, the representation of relationships between entities is essential for structuring information, modeling knowledge, and building semantic graphs (Dessi *et al.*, 2020). These relationships are often formalized in the form of triplets (subject, predicate, object), inspired by the RDF model. For example (see Figure 2), the sentence *"Frodo found the Ring"* can be translated into a triplet: *(Frodo, find, the Ring)*. This format allows the logical structure of texts to be extracted and made searchable in knowledge bases (OWL/RDF). The identification of these relationships can be based on several types of links: simple co-occurrences (presence of entities in the same sentence), explicit relationships expressed by a verb or phrase (e.g., *to be located in*, *to belong to*), or

implicit relationships such as apposition (*Bilbo, the hobbit*) or possession (*Gandalf's sword*).

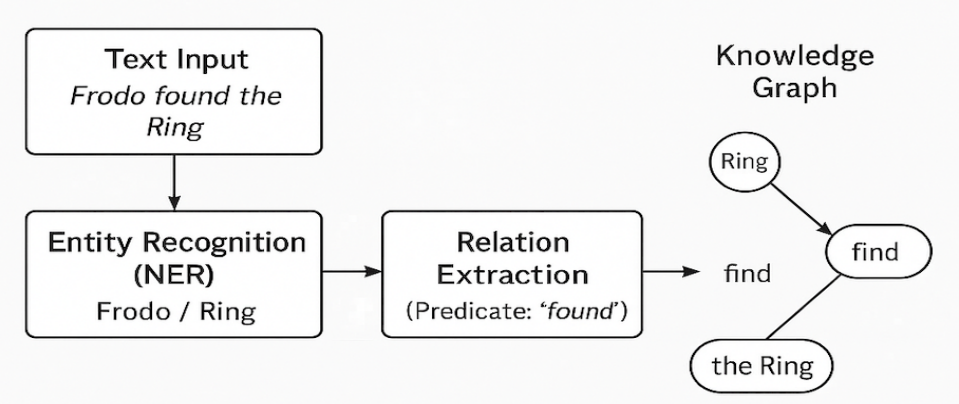

Figure 2: Triplet extraction pipeline

Automatic triplet extraction relies on several complementary approaches: (1) Open extraction (OpenIE), which identifies all possible triplets without a predefined schema —useful for exploring diverse corpora but subject to noise; (2) Schema-based extraction, which relies on a predefined inventory of relationships (e.g., *is_located_in, has_author*) and offers better accuracy, especially in specialized domains; (3) Complex event extraction, where triplets are enriched with temporal or causal metadata, such as *(Eowyn, kills, Witch King, during the battle)*. This process involves various components: named entity recognition (NER), parsing, relational pattern recognition, contextual embeddings (BERT, RoBERTa), and RDF/OWL graph structuring. Depending on the objectives (semantic web, expert systems, information extraction), the types of relationships targeted may range from generic to highly specialized.

How to apply the framework to idioms and semantic issues? In our project, applying this framework to idiomatic expressions poses specific challenges: it involves identifying, within a corpus, triplets of the type (idiom, keyword, semantic relation). For example, "give a hand" → (give a hand, hand, help). The difficulty lies in the non-compositionality of idioms, their syntactic variants, and their often context-dependent semantics. Their extraction requires a detailed analysis of nominal, adjectival, or verbal constructions, based on grammars such as that of Tesnière (1959). Integrating these triplets into an ontology involves combining linguistic recognition, semantic typing, and logical formalization. The challenge is to transform unstructured texts into usable representations—enriched by idioms—and capable of feeding interoperable knowledge bases in multilingual and multicultural contexts (Chen & Gasparini, 2025; Chen et al., 2025).

3. Our experience: automating the ontology of phraseological relations

The following five-phase roadmap (Table 1) outlines the key stages of our project, from the initial extraction of idiomatic expressions to their final integration into a usable and interoperable ontology.

Phase	Objective	Tools / Technologies
Corpus Construction and Preprocessing	Clean the corpus, perform tokenization, and syntactic analysis	spaCy / NLTK
Term and Idiom Extraction	Automatically extract PUs / idioms	TF-IDF, PMI, C-value, RAKE, RRF
Relationship Extraction and Graph Construction	Extract relationships approximately from sentences and construct a graph	Cooccurrence, graph / networkx
Deep Learning Analysis	Fine-tuning with BERT / S-BERT models, clustering	HuggingFace Transformers
Visualization and Export	Streamlit interface, export to OWL, graphical visualization	Streamlit + owlready2 + PyVis

Table 1: Overview of project phases, objectives, and tools

3.1 Presentation of the corpus

The corpus is composed of three complementary subsets:

1) Reference list of idioms

We first created a reference list of 750 French idiomatic expressions, focusing on two major semantic fields:

a) The human body: *en chair et en os* [in the flesh], *se croiser les bras* [to cross one's arms], *avoir bon coeur* [to have a good heart], *faire un pied de nez* [to thumb one's nose], etc.

b) Animals: *une mère de poule* [a mother hen], *avoir mangé du lion* [to have eaten a lion], *prendre la mouche* [to take the fly], etc.

This list was manually compiled as part of our DiCoP (Dictionary of Phraseological Concepts) project, by cross-referencing various lexicographic sources and authentic corpora. Early versions of this list have been presented and discussed in several previous works (Chen, 2023; 2024, etc.)

2) Context corpus

To analyze these expressions in actual usage and better understand the contexts in which they appear, we created a context corpus. This is a French corpus extracted from Wikipedia, distributed in Moses format by the OPUS collection (Wikipedia version v1.0, published on March 4, 2018: http://opus.nlpl.eu/Wikipedia-v1.0.php). This corpus comprises approximately 15.8 million words, which we segmented into 200 files

("chunks") of around 78,959 words each, in order to facilitate processing and automatic analysis—particularly for identification, co-occurrence studies, and the analysis of syntactic relationships between idioms.

3) Lists of thematic keywords

To enrich the semantic and thematic analysis of our corpus, we created two keyword lists associated with the areas under study:

a) 58 keywords related to the human body: *gorge, cœur, coude, jambe, main*, etc. [throat, heart, elbow, leg, hand, etc].

b) 82 keywords related to animals: *zibeline, chat, chien, cheval, dragon,* etc. [sable, cat, dog, horse, dragon, etc.]

These lists serve to guide the identification of fixed expressions and to explore the conceptual networks they activate in the texts.

This three-part corpus thus constitutes a solid foundation for the study of French phraseology by combining a lexicographic approach (reference list), an empirical approach (contextual corpus), and a semantic approach (thematic keywords). In this context, our objective is to extract, model, and visualize the relationships between these phraseological units using automatic language processing methods and ontological representation techniques.

3.2 Our experience in ontology automation

3.2.1 Concept extraction: Corpus preparation and statistical methods

In an initial, so-called statistical phase, we applied several classical techniques for extracting terms and relationships from the cleaned corpus. Processing was performed using the spaCy and NLTK libraries, utilizing methods such as TF-IDF, PMI, C-value/NC-value, and RAKE to identify fixed expressions and relevant multi-word units. The results obtained from these different approaches were consolidated using the Reciprocal Rank Fusion (RRF) algorithm, allowing us to unify the rankings generated by each method. Morphosyntactic filtering was then applied to eliminate irrelevant units. Finally, an interactive interface developed with Streamlit was implemented to automatically support the visualization, filtering, and enrichment of the extracted concepts, with no human annotation involved.

1) Detection of idioms / Phraseological Units (PUs) in the corpus

In this step, we focused on detecting idiomatic expressions in a real corpus, we compared each expression from the reference list to the corpus segments using literal matching. The objective was to identify the expressions actually present in the corpus

and to count their frequencies of occurrence. This method provides a detailed analysis of idioms' actual presence, usage frequency, and potential for future analyses of co-occurrence and syntactic relationships.

In total, 21 idiomatic expressions were identified along with their respective frequencies (e.g., *en chair et en os: 3* [in flesh and blood: 3]; *connaître par cœur: 1* [know by heart: 1] ; *une vie de chien: 5* [a dog's life]). This step therefore empirically validates the anchoring of phraseological units in real usage and serves as a basis for subsequent semantic and relational analyses, whether in terms of co-occurrences, syntagmatic dependencies, or ontological modelling.

2) Corpus preparation and structuring

We started from the *raw corpus (frechcorpora.txt)* and, in this phase, applied a series of operations to clean and structure it for subsequent processing. First, we assessed the presence and frequency of a set of 750 French idiomatic expressions across the raw data, in order to study their usage and distribution. To facilitate analysis, the corpus was then automatically divided into fifty equal segments, or "chunks," each containing a balanced number of lines from the original file, enabling parallel processing and batch analysis. Linguistic preprocessing was carried out using the spaCy library (*model fr_core_news_sm)*, including lexical tokenization (by words), sentence segmentation, token filtering (removal of punctuation marks, spaces, and stop words), morphosyntactic tagging (POS), and preparation of syntactic dependencies. At the end of this process, the corpus had been fully cleaned and structured, and two output files were generated: 1) *token_mots.txt*: a cleaned and tokenized corpus by words (one document per line); 2) *token_phrase.txt*: the corpus segmented into sentences (each line containing one or more sentences separated by vertical bars "|").

3) Automatic extraction of conceptual candidates

Based on this foundation, several statistical methods were implemented to extract multi-word expressions likely to correspond to idiomatic units. These methods include TF-IDF (Term Frequency–Inverse Document Frequency), PMI (Pointwise Mutual Information), C-value / NC-value, and the RAKE (Rapid Automatic Keyword Extraction) algorithm. Each method was finely parameterized—particularly by adjusting the size of the n-grams and the frequency thresholds—in order to optimize the relevance of the extracted expressions. The results were then combined using the Reciprocal Rank Fusion (RRF) algorithm, allowing the consolidation of a list of candidate expressions by taking into account the cross-rankings produced by each method.

4) Language filtering

The resulting n-grams were subjected to linguistic filtering based on structural rules,

such as frequent syntactic patterns (e.g., verb + noun, to have + noun, etc.), as well as criteria related to length and the proportion of function words. An interactive interface was developed using *Streamlit* to allow exploration, refinement, and annotation of candidate expressions. This step facilitated human validation of the most relevant units while discarding noisy or insignificant elements. The result of this process is a first consolidated lexical base of candidate idiomatic units, ready to be used in the subsequent steps of ontological modelling.

The final expressions were saved in tabular form (*ulps.csv*), with the following columns: *ngram*, *score*, and *method*. Simple visualizations (e.g., histograms, word clouds) were used to examine the distribution of the *scores* (e.g., TF-IDF, PMI, C-value) and to evaluate the impact of different threshold values applied to these scores on the selection of relevant expressions. This ensured that only candidates above a defined relevance threshold were retained.

This phase was structured around two main components: the preparation of the corpus and the automatic extraction of conceptual candidates. It lays the foundations of the project by ensuring both the linguistic robustness of the corpus and the reliability of the initial extraction of candidate idiomatic units. It serves as the basis upon which the subsequent phases of relational analysis, conceptual typing, and ontological modelling will be built.

3.2.2 Extraction and structuring of idiomatic relations

The second phase, focused on relationships, began with an approximate strategy based on co-occurrence within the same sentence, particularly between subjects and other entities. The extracted relationships were represented as a weighted graph, facilitating the identification of central nodes (e.g., *donner un coup de main* → *main* → *aide* [*give a helping hand* → *hand* → *help*, *authority*, *work*]).

After detecting the idiomatic expressions in the corpus, the next step in our project consisted of automatically extracting the semantic relationships between these idioms and the keywords they contain, and then structuring these relationships in the form of interpretable triplets (idiom → word → concept).

We developed a Python script based on the spaCy library to analyze each sentence in the tokenized corpus (token_phrase.txt) and identify, for each detected idiom: (1) The keyword (often a noun related to the body or animals) it contains; (2) The verb or associated syntactic relation, when the idiom functions as the subject; (3) The semantic concept corresponding to the keyword, based on a manually defined correspondence map.

This step relies on two lexical files:

(1) mots_corps.txt: a list of nouns related to body parts

(2) mots_animaux.txt: a list of animal names

These keywords (see Figure 1) were then grouped according to a semantic map linking each word to an abstract concept (for example: *main* → *aide*, *rat* → *ruse*, *pied* → *stabilité*, etc. [*hand* → *help*, *rat* → *cunning*, *foot* → *stability*, etc.]). In the code, this mapping was implemented by adding each detected relation to a list of triplets using the following command: *triplets.append((idiom, word, semantic_map[word])).* This line simply records, for each idiom, the keyword it contains and the corresponding abstract concept into a triplet structure. For instance, the idiom *avoir le bras long* (to have a long arm) would generate the triplet (*avoir le bras long*, *bras*, *pouvoir*) = (*have long arm*, *arm*, *power*).

```
semantic_map = {
    "bras": "accueil",              # à bras ouverts
    "chair": "corps",               # en chair et en os
    "cœur": "mémoire",              # connaître par cœur
    "visage": "identité",           # à visage découvert
    "main": "aide",                 # avoir la main, de main de maître, etc.
    "œil": "perception",            # à l'œil, voir à l'œil nu
    "yeux": "confiance",            # les yeux fermés
    "pied": "stabilité",            # de pied ferme, mettre sur pied
    "talon": "faiblesse",           # talon d'Achille
    "tête": "intelligence",         # perdre la tête, tête à tête
    "cochon": "obstination",        # tête de cochon
    "chien": "malheur",             # un temps de chien, une vie de chien
    "poule": "peur",                # poule mouillée
    "fourmi": "travail",            # une vraie fourmi
    "loup": "appétit",              # une faim de loup
    "bouc": "culpabilité",          # bouc émissaire
    "brebis": "déviance",           # brebis galeuse
    "oiseau": "rareté",             # oiseau rare
    "ailes": "autonomie",           # voler de ses propres ailes
    "poisson": "inadéquation",      # comme un poisson hors de l'eau
    "rat": "ruse",                  # être rat, face de rat
}
```

Figure 3: Semantic mapping of idiomatic anchors to core concepts

This processing step generated a structured set of semantic triplets of the form: *<idiom → keyword → semantic category>*.

Below (Table 2) are some typical triplets extracted from the corpus:

Idiome	**Mot-clé [key word]**	**Catégorie sémantique [Semantic category]**
en chair et en os	chair [flesh]	corps [body]
connaître par cœur	cœur [heart]	mémoire [memory]
avoir la main	main [hand]	aide [help]
de pied ferme	pied [foot]	stabilité [stability]
bouc émissaire	bouc [goat]	culpabilité [guilt]
brebis galeuse	brebis [sheep]	déviance [deviance]
comme un poisson hors de l'eau	poisson [fish]	inadéquation [inadequacy]
une faim de loup	loup [wolf]	appétit [appetite]
les yeux fermés	yeux [eyes]	confiance [confidence]
perdre la tête	tête [head]	intelligence [intelligence]

Table 2: Mapping of french idioms to keywords and semantic categories

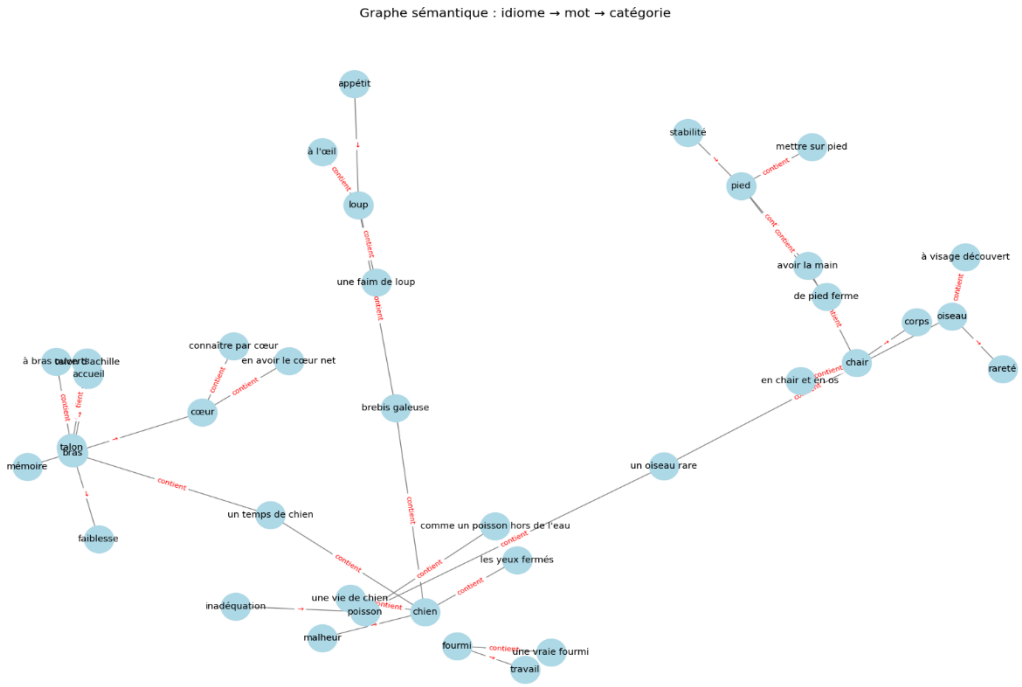

Figure 4: Semantic and relational graphs of French idioms

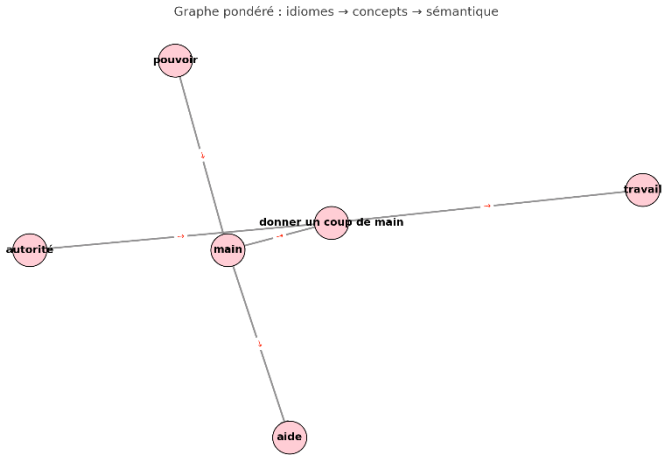

Figure 5: Weighted semantic graph of idioms related to *hand*

In this step, we built directed graphs to visually represent the relationships between idiomatic expressions, their component keywords, and their associated semantic categories. Two types of visualizations were generated from the <idiom, keyword, semantic category> triplets extracted from a CSV file.

The first type of graph (Figures 4–5) represents the relations as chains: idiom → keyword → semantic concept. Figure 2 shows a general example, while *Figure 3 illustrates a simplified weighted graph focused on the idiom 'donner un coup de main', showing how it connects to semantic concepts like 'aide', 'travail', or 'autorité'.* The relationships are indicated by annotated and weighted edges, which highlight both the

internal structure of the expressions and their conceptual anchors.

Figure 6: Visualization of relationships between idioms, keywords, and semantic categories

The second type of graph (Figure 6) enriches this representation by clearly distinguishing the nature of the links: *contains* between idiom and keyword, and a typological relation (→) between keyword and semantic category. Figure 4 shows the full relational graph in its raw form, which illustrates the density of the network, while Figure 5 provides a clearer representation with explicit labels.

This relational representation provides a synthetic view of the conceptual configurations underlying phraseological units and enables the identification of recurring semantic cores (e.g., *main* [hand] associated with *aide* [help], *cœur* [heart] with *mémoire* [memory], *tête* [head] with *intelligence*). These graphs serve as a foundation for further analyses (e.g., clustering, centrality, RDF/OWL modeling), while also

facilitating the pedagogical and lexicographic exploration of idiomatic expressions.

This step allowed us to link idiomatic expressions to semantic concepts through the lexical entities they mobilize. It constitutes a fundamental foundation for ontological modeling in the subsequent steps (RDF/OWL) and will contribute to enriching both an interactive interface and a semantically oriented phraseological dictionary.

3.2.3 Fine-tuning the BERT model for the semantic classification of idioms

In the third phase of our project, we used pre-trained neural networks (LLMs) to analyze the semantic relationships between idiomatic expressions and their associated keywords. The goal was to train a multilingual BERT-type model—specifically bert-base-multilingual-cased (hereafter referred to as BilBERT)—to automatically predict the most plausible semantic relation between a French idiom and one of its key components.

These relationships are expressed in the form of semantic triplets:

<idiom> — <semantic relation> — <keyword> For example: *donner un coup de main — expresses — help*, or *avoir le bras long — evokes — power.*T

To refine this classification task, we explored multiple strategies:

(1) the analysis of frequent verbs to define candidate meta-relations;

(2) supervised fine-tuning of BERT to recognize these semantic links;

(3) and alternative approaches based on semantic similarity using Sentence-BERT, combined with clustering techniques (e.g., Kmeans).

The input data file (triplets_semantiques.csv) was preprocessed to remove missing or empty entries. Semantic labels were then encoded using a LabelEncoder, resulting in the identification of 10 distinct semantic classes (e.g., body, memory, misfortune, work, etc.).

Model performance was evaluated using the F1-score and manual validation, confirming the potential of this approach to identify meaningful conceptual relations between idioms and their core lexical components.

Next, the data were split into a training set (80%) and a validation set (20%), and then tokenized using the BertTokenizer tokenizer with a format combining idiom, keyword, and expected relationship. The datasets were prepared in a format compatible with the Hugging Face datasets library, and the labels were explicitly cast to *int64* to ensure compatibility with the model.

The *BertForSequenceClassification* model was initialized with 10 outputs corresponding to the 10 identified semantic relationships. Since the classifier weights are randomly

initialized, training was required to allow the model to learn the specific task.

Training was performed on CPU due to environmental constraints. We fine-tuned the *bert-base-multilingual-cased* model using the following hyperparameters: a learning rate of 0.00002, 4 training epochs, a batch size of 16 for training and 64 for evaluation, a weight decay of 0.01, and an evaluation/save strategy applied at each epoch.

The model was successfully trained on the 17 examples in the training set and evaluated on 5 validation examples. Although the dataset remains modest, the initial results (loss $\approx$ 2.09) are encouraging and indicate that the model is beginning to learn meaningful correspondences. Evaluation was performed using accuracy as the main metric.

3.2.4 Evaluation of semantic relationships generated by LLMs: BLEU, METEOR, ROUGE, and BERTScore

Finally, we evaluated the quality of the semantic relationships automatically generated between idioms and keywords using large language models (LLMs), such as BilBERT (our fine-tuned model) and GPT (in generation mode). The objective was to assess the semantic relevance of the generated triplets using a set of automatic text evaluation metrics.

a) Classical text evaluation metrics (via Hugging Face's *evaluate* library): We computed the following scores: BLEU (Bilingual Evaluation Understudy), to assess n-gram overlap (precision); METEOR, which considers synonymy and grammatical variations; and ROUGE (Recall-Oriented Understudy for Gisting Evaluation), commonly used in summarization tasks, focusing on n-gram recall and longest common subsequences (LCS).

To apply these metrics, we reformulated the predicted semantic triplets into simple textual statements. For example, a model prediction might be expressed as *"main is power"* or *"main represents authority"*. Human references for the same relation were phrased as *"main is a symbol of power"* or *"main signifies authority"*. This transformation into short sentences allows us to compute overlap-based scores (BLEU, ROUGE) and synonym-aware scores (METEOR) in a comparable way. In other words, these phrases are not the main results themselves but serve as the textual representation of semantic relations, making it possible to quantitatively evaluate the closeness between machine-generated relations and human interpretations.

b) Semantic Similarity with BERTScore

We also used BERTScore, a metric based on BERT (or RoBERTa) models, which compares the vector representations (embeddings) of sentences to assess their semantic similarity at a deeper level. This approach allows us to detect paraphrastic relationships, even when the surface forms differ.

4. Results, analysis, and outlook

Below (table 3) is a table presenting the evaluation results obtained using various machine translation metrics:

Metric	Value / Detail
BLEU	0.00
1-gram accuracy	0.8333
2-gram accuracy	0.25
3-gram accuracy	0.0
4-gram accuracy	0.0
Brevity Penalty	0.6065
Length Ratio	0.6667
Translation length	6
Reference length	9
METEOR	0.3908
ROUGE-1	0.6667
ROUGE-2	0.1429
ROUGE-L	0.6667
ROUGE-Lsum	0.6667
BERTScore (F1)	0.9443

Table 3: Evaluation metrics for translation quality

The analysis of the results highlights several key findings. The BLEU score is zero (0.00), which can be explained by the absence of strictly identical n-grams between the predictions and the references—since the former are often paraphrastic. In contrast, the METEOR score, which is more tolerant of lexical variation, reaches a moderate level (≈ 0.39), while the ROUGE-L score indicates good structural similarity with the references (0.66). The BERTScore F1 score, meanwhile, is very high (0.9443), confirming that the generated sentences are semantically close to the reference sentences, even when they differ lexically.

These results suggest that large language models (LLMs) are capable of producing relevant semantic relationships between idioms and keywords, but that traditional metrics such as BLEU are inadequate for effectively evaluating this type of task. Conversely, BERTScore appears to be particularly well-suited, as it captures fine-grained semantic similarities beyond surface forms. For the next phases of the project—focusing on automatic extraction and large-scale evaluation—BERTScore or embedding-based approaches should be prioritized, while incorporating partial human validation. It would also be worthwhile to explore more advanced metrics such as COMET or BLEURT, particularly in the context of fine-tuning on multilingual or domain-specific corpora.

This final stage of the project transformed the Streamlit-based interface into a true ontological exploration environment dedicated to idiomatic expressions. Initially designed to display <idiom, keyword, semantic concept> triplets automatically

extracted from the corpus, the interface has been enhanced with several major features that significantly improve both usability and ergonomics.

1) Dynamic sorting of concepts and relationships

Users can now dynamically sort idioms, either alphabetically or by type of semantic relationship. This interactive sorting facilitates navigation through the dataset and enables the rapid identification of recurring linguistic or conceptual patterns.

2) Interactive visualization of relationships in graph form

Thanks to the integration of the PyVis library, a dynamic graphical visualization has been implemented. In this directed graph: a) Nodes represent idioms and their associated keywords (often body parts, animals, etc.); b) Edges encode the detected semantic relationships (such as aide, mémoire, stabilité, etc. [help, memory, stability, etc.]).

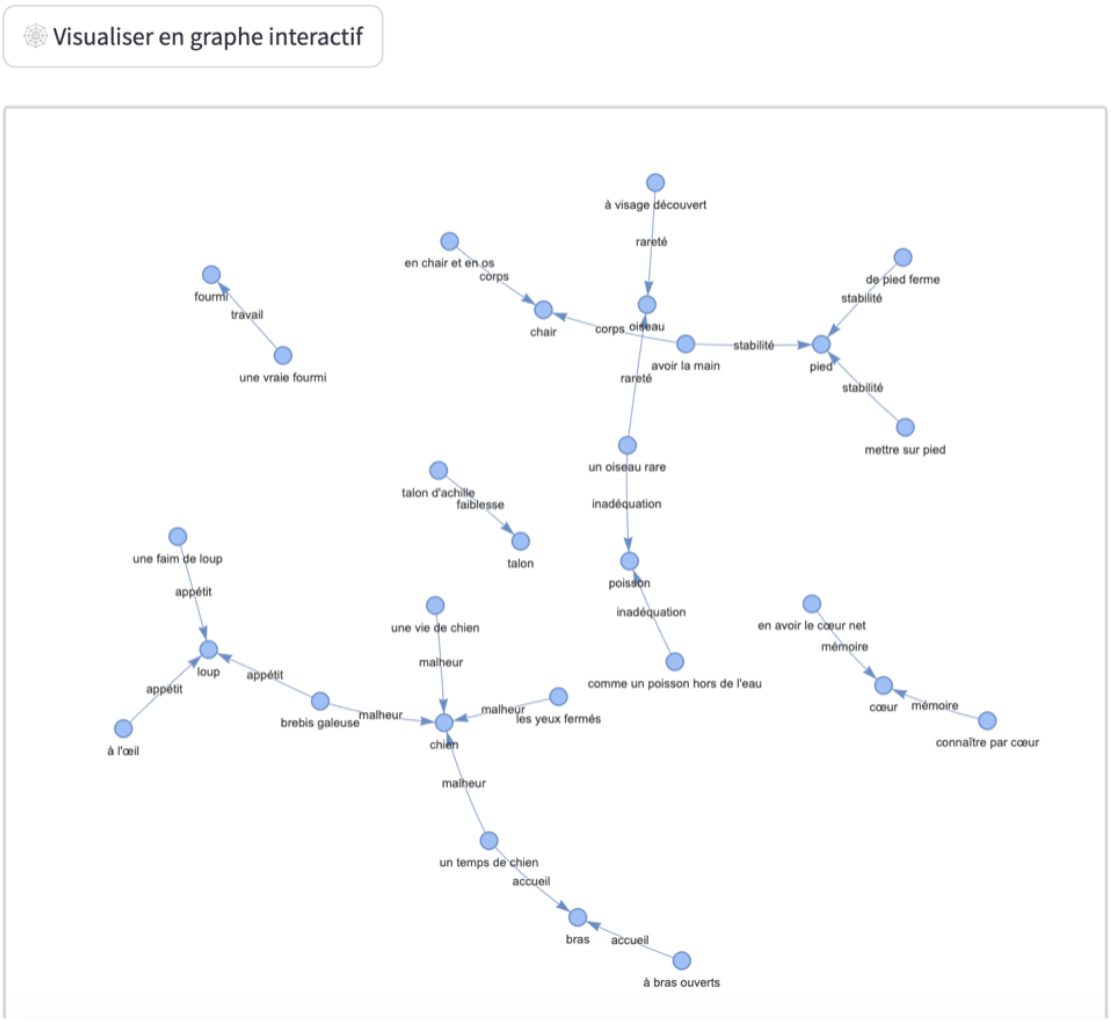

Figure 7: Interactive semantic graph of French idioms

This mode of visual representation (see Figure 7) greatly enhances the readability of relational structures and provides an intuitive view of the semantic networks underlying the idiomatic lexicon.

3) Export to OWL format (RDF/XML)

Finally, the interface now includes a feature for automatically exporting data as an ontology compliant with the OWL standard (Figure 8). This option not only enables idiomatic knowledge to be structured in an interoperable format, but also allows it to be used in semantic reasoning environments such as Protégé or accessed via SPARQL queries. Each triplet is converted into a valid RDF structure:n a) idioms are typed as instances of the class ex:Idiome; b) keywords as ex:Concept; c) semantic relationships are modelled as object properties.

The export generates an OWL file (ontology_idiomes.owl) (Figure 9), ready to be imported into Protégé, compared with other resources (e.g., OLAF), or used for advanced semantic processing.

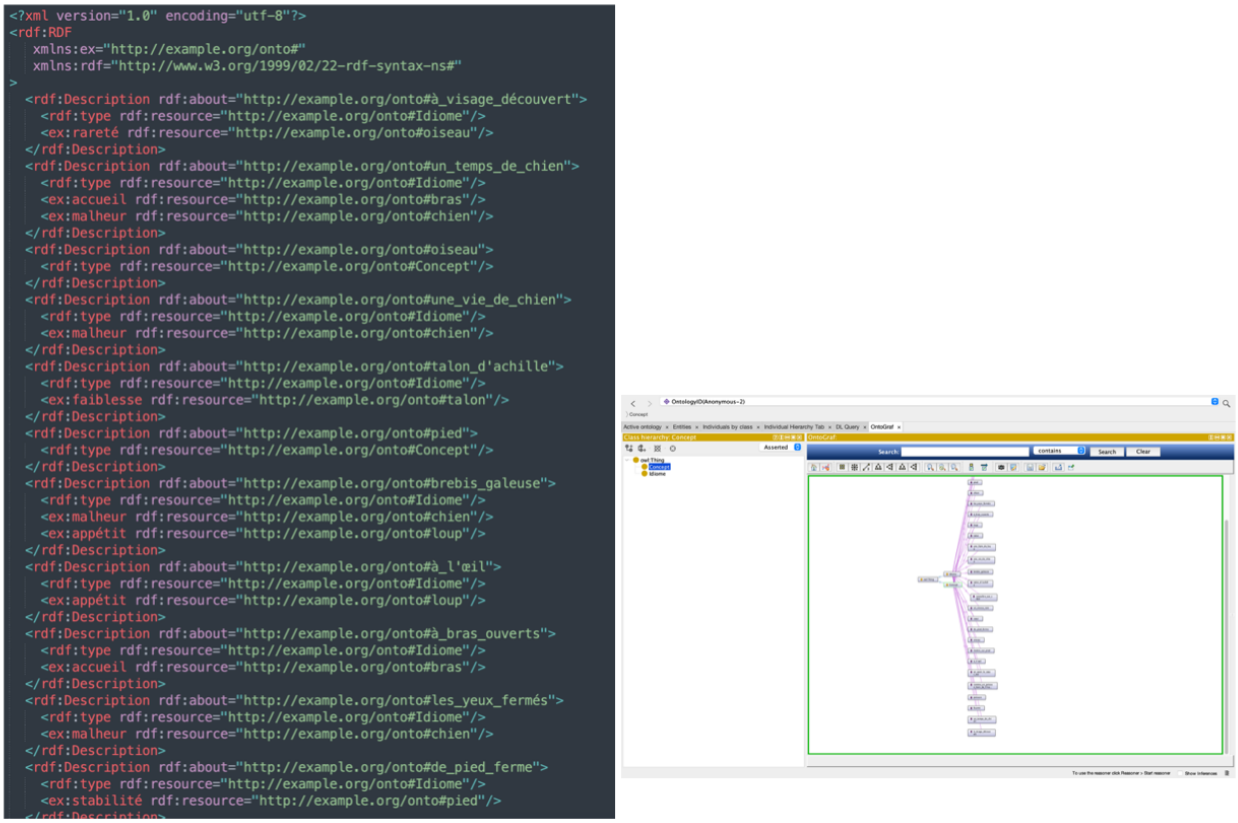

Figure 8: RDF integrated into *Protégé*

The finalized system is accessible locally at http://localhost:8503/. It provides a user-friendly and dynamic interface for exploring phraseological relationships, conducting

qualitative analyses, visualizing semantic structures, and exporting data in the standardized OWL format. This module marks the completion of Phase 5, establishing the foundation for an interoperable, multilingual idiomatic ontology—ready to be enriched through machine learning or collaborative annotation.

5. Auto-lexicography and multilingual phraseological modeling with OntoLex

This project is part of a phraseological auto-lexicography approach aimed at modeling bilingual (French–Chinese) idiomatic expressions in RDF/OWL format, based on the OntoLex-Lemon model and its extensions, including VarTrans, LexInfo, and SKOS. In this fifth and final phase of the project, we implemented a multilingual semantic auto-lexicography module grounded in the OntoLex-Lemon standard.

This project is part of a phraseological auto-lexicography approach aimed at modeling bilingual (French–Chinese) idiomatic expressions in RDF/OWL format, based on the OntoLex-Lemon model and its extensions, including VarTrans, LexInfo, and SKOS. In this fifth and final phase of the project, we implemented a multilingual semantic auto-lexicography module grounded in the OntoLex-Lemon standard. The objective was to formally represent a set of French and Chinese idiomatic expressions, along with their conceptual and translational relationships, in an interoperable format compliant with Semantic Web standards.

To achieve this, we used the RDFLib library to build an RDF graph comprising:

1) Hierarchical concepts (e.g., *goat* and *fish* as subclasses of *animals*; *heart* as a subclass of *human body*), modeled using SKOS;

2) Bilingual phraseological units (French idioms and their Chinese equivalents), modeled as `ontolex:LexicalEntry`;

3) Lexical senses (`ontolex:LexicalSense`) associated with concepts, with usage examples and pronunciations where appropriate;

4) Translation relationships aligned with the VarTrans module, linking idioms across languages.

Each idiomatic triple is encoded as an RDF structure. For example, the French expression *bouc émissaire* is linked to the concept *bouc* (in the *animals* domain) and to its Chinese equivalent 替罪羊 (*tì zuì yáng*), through a `vartrans:Translation` relationship. Similarly, *savoir par cœur* is aligned with 熟记于心 (*shú jì yú xīn*), both linked to the concept *cœur* (*heart*).

1) Ontological structure and conceptual modeling

The first step consisted in defining a hierarchy of phraseological concepts using SKOS,

allowing idioms to be grouped by semantic theme. For example: Heart is a sub-concept of HumanBody; Fish and Goat are related to the broader concept Animals. These concepts are instantiated as RDF nodes of type skos:Concept, with skos:broader relationships reflecting their taxonomic organization.

2) Representation of idioms (Lexical entries)

Each idiomatic expression is modeled as a lexical entry (ontolex:LexicalEntry), comprising: 1) a canonical form (ontolex:canonicalForm → ontolex:writtenRep), in French or Chinese; 2) an optional pronunciation (ontolex:pronunciation) for the Chinese forms; 3) a lexical meaning (ontolex:sense), linked to a SKOS concept via ontolex:concept; 4) a usage note (rdfs:comment) illustrating the expression in a real-life context.

French and Chinese expressions are linked through translation relationships using the vartrans:Translation module. Each relationship (vartrans:Translation) associates a source meaning (vartrans:source) with a target meaning (vartrans:target), thereby modeling idiomatic equivalence between languages. Examples of modeled alignments include: know by heart ↔ 熟记于心 → concept: Heart; scapegoat ↔ 替罪羊 → concept: Goat (a subclass of Animals); like a fish out of water ↔ 如鱼得水 → concept: Fish.

The generated RDF export file (Figure 9) can be opened with semantic tools such as Protégé (Figure 10), WebVOWL, etc., for manual inspection or enrichment, or integrated into SPARQL applications for querying and inference. This file includes all idioms, their linguistic forms, associated concepts, cross-lingual translations, and conceptual hierarchies.

Figure 9: RDF ontology code

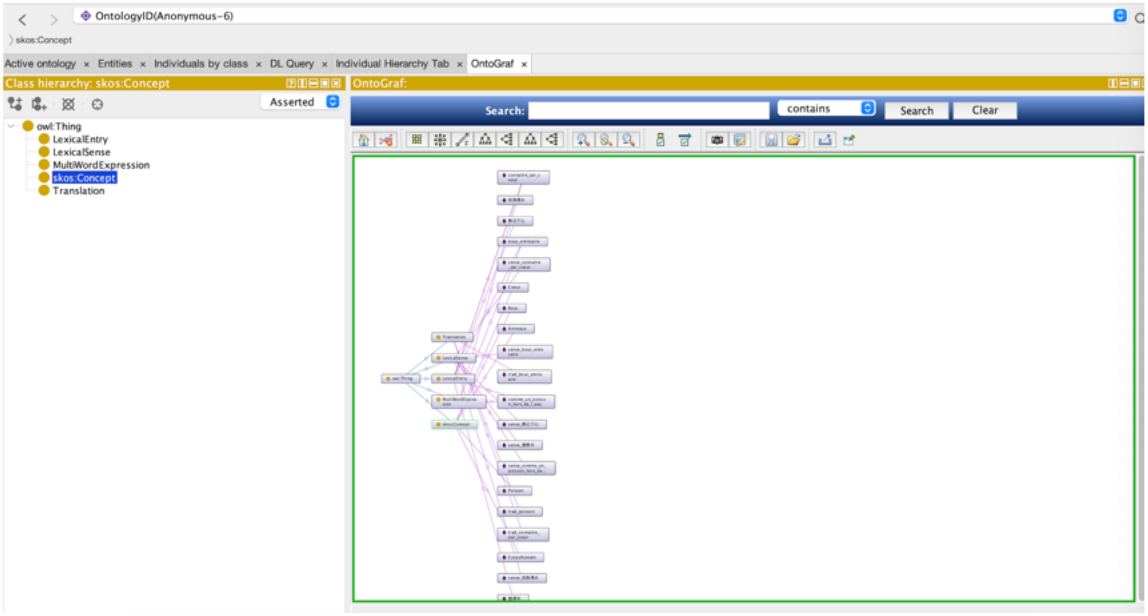

Figure 10: Visualization in protégé (OntoGraf View)

By structuring idiom data in this way, this step marks the transition from a simple annotated corpus to an interoperable multilingual phraseological ontology capable of powering NLP applications (semantic extraction, assisted translation, bilingual example generation, etc.) and paving the way for auto-lexicography enriched by AI and collaborative annotation. This modeling not only allows for the representation of the multilingual and cultural dimensions of idioms, but also enables the preparation of semantic queries via SPARQL, interactive visualizations, and extensions to machine translation, teaching, or AI-assisted generation.

6. Conclusion

This work lays the foundations for a bilingual French-Chinese phraseological dictionary modeled as an interoperable ontology, combining statistical, symbolic, and neural approaches derived from NLP. By automating the extraction, structuring, and semantic linking of idiomatic expressions from authentic corpora, our method overcomes the limitations of traditional lexicography, such as manual compilation, slow update cycles, and reliance on expert intuition. The integration of pre-trained language models (LLMs), such as BERT, into a multilingual ontological pipeline based on the OntoLex-Lemon standard, demonstrates that it is possible to link idioms to abstract concepts and their translations, while preserving their cultural roots. Through automatic annotation, interactive visualization, and RDF/OWL export, we transform a simple linguistic inventory into a knowledge base suitable for multilingual and multicultural applications. This project paves the way for AI-enriched auto-lexicography, where phraseology becomes a preferred vector for semantic modeling, assisted translation and the generation of contextualized bilingual examples, particularly in the fields of education, the Semantic Web and language technologies.

7. References

Amdouni, E., Belfadel, A., Gagnant, M., Renault, I., Kierszbaum, S., Carrion, J., Dussartre, M. & Tmar, S. (2025). Semi-Automatic Building of Ontologies from Unstructured French Texts: Industrial Case Study. *Data Science and Engineering*. Published 19 June 2025. Available at: https://doi.org/10.1007/s41019-025-00284-z

Chen, L., Gasparini, N., Dao, H.-L., & Do-Hurinvillle, D.-T. (2025). Toward a trilingual ontology of phraseological units: Lexicographic and computational modeling in Chinese, French, and Vietnamese. AsiaLex 2025: The 18th International Conference of the Asian Association for Lexicography, pp. 13–21.

Chen, L. & Gasparrini, N. (2025, May). Modélisation et structuration d'un dictionnaire bilingue français-chinois des expressions idiomatiques : approche lexicographique et ontologique. In *Proceedings of the Sixième Colloque international « Dictionnaire et polylexicalité »*, Université de Bari (Italie), Université Sorbonne Paris Nord (France), Université de Silésie à Katowice (Pologne).

Chen, L. (2024). Traitement de la traduction et de la transmission culturelle de la microstructure dans les dictionnaires bilingues des UP: étude et analyse contrastive de corpus métalexicographique. *SHS Web of Conferences*, 139, 11001, pp. 1–18.

Chen, L. (2023). (Meta)phraseography and phraseomatics: DiCoP, a computerized resource of phraseological units. In *Conference Proceedings of ASIALEX 2023: Lexicography, Artificial Intelligence, and Dictionary Users – The 16th International Conference of the Asian Association for Lexicography*, pp. 224–231.

Chen, L. (2021). *Analyse comparative des expressions idiomatiques en chinois et en français (relatives au corps humain et aux animaux)* [PhD thesis, Cergy Paris Université].

Constant, M. (2012). *Mettre les expressions multi-mots au cœur de l'analyse automatique de textes : sur l'exploitation de ressources symboliques externes*. Traitement du texte et du document. Université Paris-Est. ⟨tel-00841556⟩

Despres, S. & Szulman, S. (2008). Réseau terminologique versus Ontologie. In *Proceedings of Toth 2008*, France, pp. 1–19.

Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D. & Motta, E. (2020). Generating Knowledge Graphs by Employing Natural Language Processing and Machine Learning Techniques within the Scholarly Domain. Available at: https://arxiv.org/pdf/2011.01103

Elnagar, S., Yoon, V. & Thomas, M.A. (2020). An Automatic Ontology Generation Framework with An Organizational Perspective. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, pp. 4860–4869.

EUROPHRAS. (2023). 4th International Conference on Computational and Corpus-based Phraseology. In *Proceedings of EUROPHRAS 2023*, pp. 17–25.

González-Rey, M.I. (2002). *La phraséologie du français*. Toulouse: Presses Universitaires du Mirail.

Kapoor, B. & Sharma, S. (2010). A Comparative Study of Ontology Building Tools for Semantic Web Applications. *International Journal of Web & Semantic Technology*

(IJWesT), 1(3), pp. 1–13.

OLAF: An Ontology Learning Applied Framework. (2023). In *Proceedings of the 27th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES 2023)*, Athens, Greece, pp. 2106–2115.

Mejri, S. (2011). Phraséologie et traduction. *Équivalence*, 38(1–2), pp. 111–133.

Mejri, S. (1997). *Le figement lexical: descriptions linguistiques et structuration sémantique* (Série Notions de base en lexicologie). Manouba: Publications de la Faculté des Lettres de la Manouba.

Mel'čuk, I. (2008). La phraséologie et son rôle dans l'enseignement/apprentissage d'une langue étrangère. *Études de Linguistique Appliquée*, 92, pp. 82–117.

Mel'čuk, I. (2011). Phrasèmes dans le dictionnaire. In J.-C. Anscombre & S. Mejri (eds.) *Le figement linguistique: la parole entravée*, Paris: Honoré Champion, pp. 41–61.

Mel'čuk, I. (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais…. *Cahiers de Lexicologie*, 102, pp. 129–149.

Murano, M. (2011). *Le traitement des séquences figées dans les dictionnaires bilingues français-italien, italien-français*. [Édition en français].

Musen, M.A. (2015). The Protégé project: A look back and a look forward. *AI Matters*, 1(4), pp. 4–12.

Polguère, A. (2008). *Lexicologie et sémantique lexicale*. Montréal: Les Presses de l'Université de Montréal.

Polguère, A. (2002). *Notions de base en lexicologie*. Paris: Ophrys.

Sułkowska, M. (2016). Phraséodidactique et phraséotraduction: quelques remarques sur les nouvelles disciplines de la phraséologie appliquée. *Yearbook of Phraseology*, 7, pp. 35–54.

Tiwari, S.M. & Jain, S. (2014). Automatic Ontology Acquisition and Learning. *IJRET: International Journal of Research in Engineering and Technology*, pp. 38–43.

Varone, M. (2011). Method and System for Automatically Extracting Relations Between Concepts Included in Electronic Text. *U.S. Patent* 7,899,666 B2, issued March 1, 2011. Assignee: Expert System S.p.A. (Modena, Italy). Application filed May 4, 2007.