Colidioms: An Online Software for Phraseography and Paremiography

Elena Berthemet

Keywords: *collaborative*, *idiom*, *notion*, *semantics*, *translation*.

Abstract

This paper investigates the possibility of building a multilingual phraseological database. It presents the framework of a privately-funded online project called Colidioms. The goal of the Colidioms project is to build a public collaborative database. The software is designed for the full perception and reproduction of phrasemes. Combining tradition and innovation, Colidioms is based on recent technological advances. It is a web application that supports English, French, German and Russian and enables multi-directional search of phraseological equivalents in any of these four languages. Two types of search are available: semasiological and onomasiological. The central organizing principle of the software is based on the concept of 'notions'. Notions allow to create a bridge between phrasemes in different languages. It has been demonstrated that notions make it possible to carry out cross-lingual comparisons. Notions link all parts of the database and homogenize the corpus and are compatible with all studied languages.

1. Introduction

A great number of modern bilingual dictionaries propose equivalents of phrasemes. Anyone who has tried to search for a phraseme in such a dictionary knows how upsetting it can be. In fact, one may spend a lot of time searching for an equivalent. Besides, even if he finally finds the so-called equivalent, often the complementary information about its usage is not provided, so that there is a risk that the found equivalent will be used incorrectly.

In this paper, we will try to answer the following question: Is it possible to build a reliable multilingual phraseological database? In order to answer this question, the paper is divided into five sections. Section 2 gives an overview of Colidioms. Section 3 describes the central organizing principle, based on the concept of 'notions'. Section 4 presents a case study showing what the proposed software looks like. A demonstration of a search using a conventional word search and making complex queries is given in Section 5. Finally, concluding remarks are made in Section 6.

2. Brief presentation of Colidioms

Colidioms is an extensive template-based database which is used to record meanings, as well as semantic and syntactic combinatorial properties of phrasemes. Colidioms is based on recent technological advances. It combines tradition and innovation.

Colidioms is a web application: one does not need anything other than a web browser to use it. It has been tested with Internet Explorer, Firefox and Google Chrome. Colidioms is programmed in Java using Eclipse. While the server runs on Windows with MS SQL Server, it is possible to modify the application to run on any operating system with any database. The current Colidioms version is like a wiki in the sense that it is easy to use. However, unlike a wiki, the structure of the Colidioms article is rigid. Indeed, all articles have the same structure, independently of the language. This greatly helps readability. Also, all parts of each article are clearly identified, as further described in the section *Microstructure: A Case Study*.

The technologies used in Colidioms allow an unlimited number of lemmas. These technologies have already been used in another project with a database of over 100 million entries, which still exhibits a very fast behaviour. Every time an expression is modified, Colidioms updates it while keeping the original version. This makes it possible to compare all versions and revert to the previous one, if needed.

Colidioms supports English, French, German and Russian and enables multi-directional search of phraseological equivalents in any of these four languages. When Colidioms moves out of prototype phase, we will gradually start adding other languages to it, such as Arabic, Chinese, Italian or Japanese. As of today, however, it contains very few entries and still needs improvements, such as modifications to make it more user-friendly.

3. Central organizing principle, based on the concept of 'notions'

This study is based on the practical results of the author's PhD thesis 'Compared Lexicology of Phraseological Units (Zoomorphisms in French, English, German and Russian)'. Our objective is to achieve coherence of the database, as we want to put all phrasemes into a single space we need a method applicable to each phraseme of each language. Definitions are written in different languages and images differ from one language to another. Even if two phrasemes mean approximately the same thing and if they have the similar images, a computer is not able to automatically find the link between two phrasemes from different languages. Thus, we need to create a bridge between expressions in different languages. In order to link all parts of phraseme's patchwork, we use 'notions'.

Every entry in Colidioms is associated with several ideas called notions in the scope of the present work. For example, the English proverb *Dogs bark, but the caravan goes on* is associated with basic words *attack, critique, debate, jealousy* and *remark*. Namely, these basic words, or notions, belong to the domain of human behaviour, which describe physical, mental, and social activity of human beings. Notions in Colidioms are similar to descriptors in a thesaurus. The following two dictionaries have a similar structure: Baranov et al. (2007) and Urdang (1998).

To automate the search, notions are written in English. It would be time-consuming to write notions for each individual phraseme. For this reason, tagging process is partially automated. While definitions are written manually, we use templates which are an automating part of the annotation. The lexicographer can choose an existing notion or add a new one.

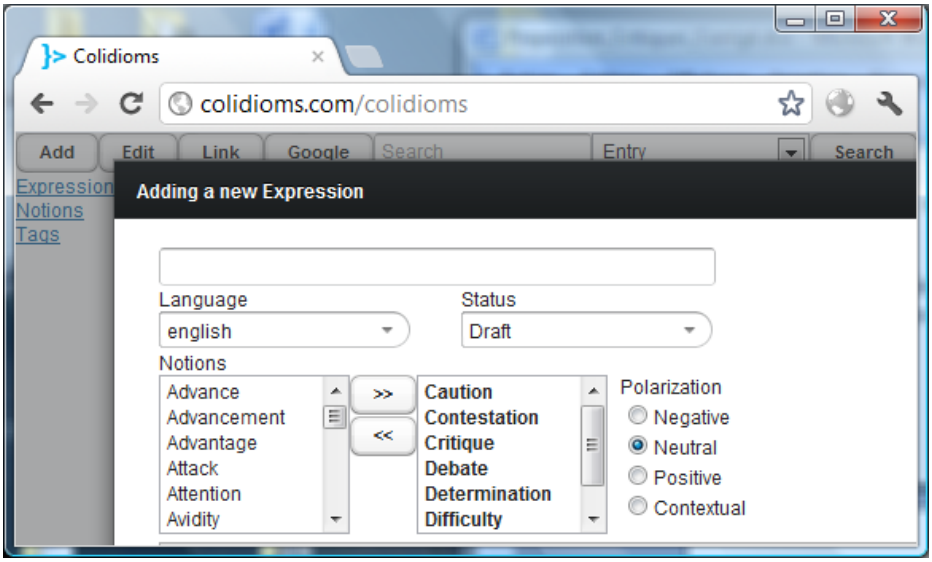

Figure 1. Annotations as they appear in the annotation template.

All notions form a system that has no hierarchical structure. As a given phraseme is associated with several notions, its semantic classification does not have strict boundaries and is not homogeneous. Rather, it is a multi-criteria system that represents a complex graph of relations with no subordinate values.

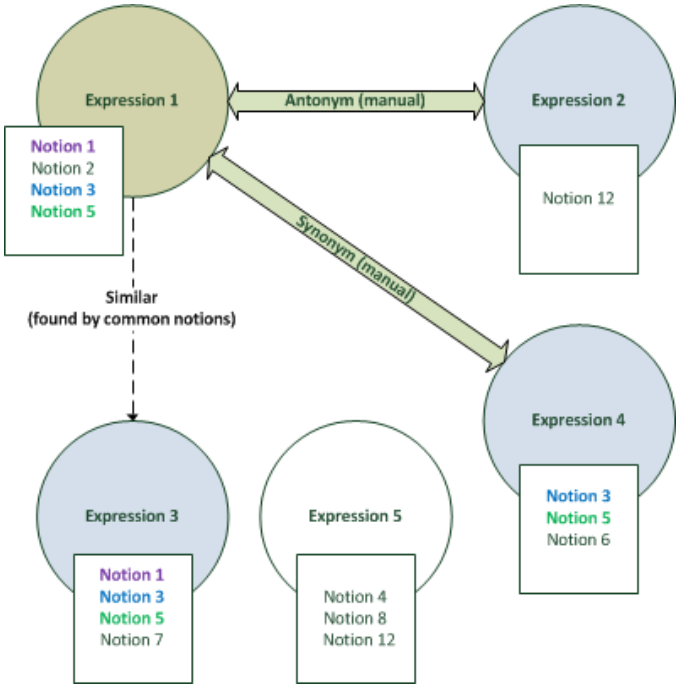

Figure 2. Complex graph of relations.

It is not easy to attribute notions to phrasemes. Being parts of definitions, they derive from the meaning of each particular phraseme. While definitions have a socio-linguistic aspect, notions do not. Notions are attributed intuitively and inductively, they are as neutral as possible. It is true that the attribution of notions is a fastidious, time-consuming and subjective task. As Ahmanova (1974) suggests, *'... we can assume that all human beings and all human societies have more or less (or basically) the same conceptual taxonomic structure... On the other hand, ... there still exist societies which do not distinguish between rich and poor, stupid*

and clever.' Indeed, it seems that notions have no scientific basis as it has not been proven that all languages have the same concepts. Moreover, the lexicographer and the user may have different notions in mind.

However, the use of notions homogenizes the corpus, where they can be viewed as labels allowing to easily find phrasemes. Notions do not replace the traditional description of a phraseme, but rather complement it. Thus, notions, founded on the definition of a phraseme, take into account its semantics and allow to do the primary selection when searching for a phraseme. To compensate for this scientific imprecision, each unit has a definition and a large corpus. Two sources can be used to produce reliable lexicographical data: existing dictionaries and corpora. Regrettably, the scope of this paper does not allow a thorough presentation of corpora exploration.

Let us take a look at a Colidioms article to see how it describes phrasemes.

4. Microstructure: A Case Study

The microstructure is determined by a set of rules. The example represents the proverb *Dogs bark, but the caravan goes on*. Each article is rigorously structured as follows: *entry*, *notions*, *links*, *variants*, *definition*, *etymology*, *folk etymology*, and *examples*.

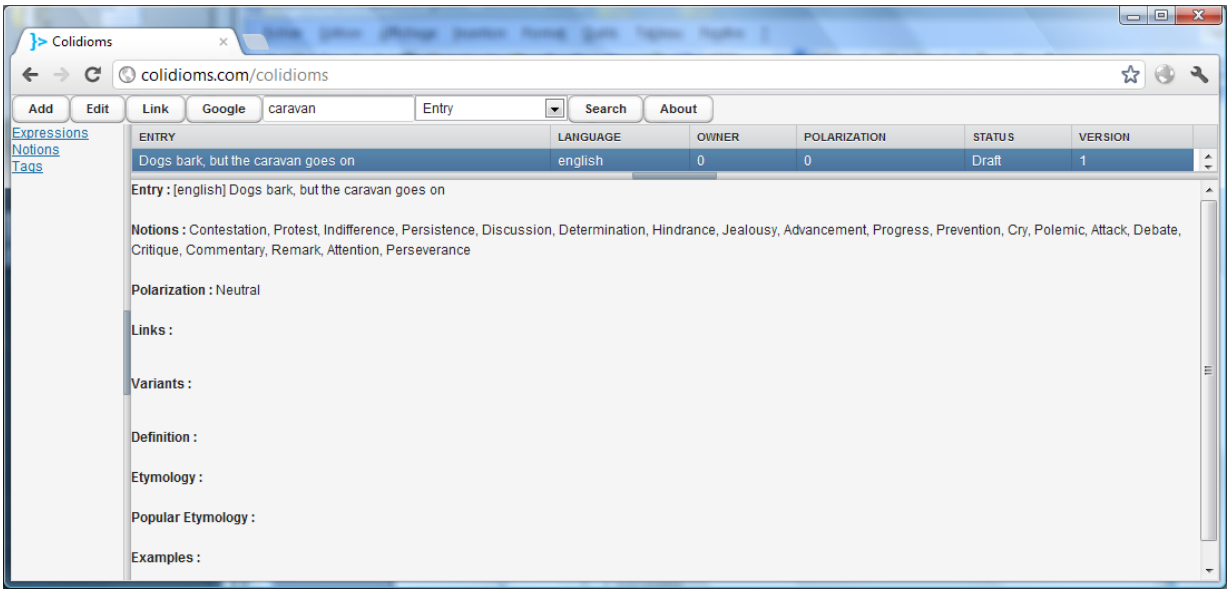

Figure 3. Colidiom's microstructure.

Apart from the notions, the entire article is written in the original language of the phraseme. For example, one can see that in the Figure 4, the definition of the French phraseme *le chien du jardinier* 'a dog in the manger' is written in French.

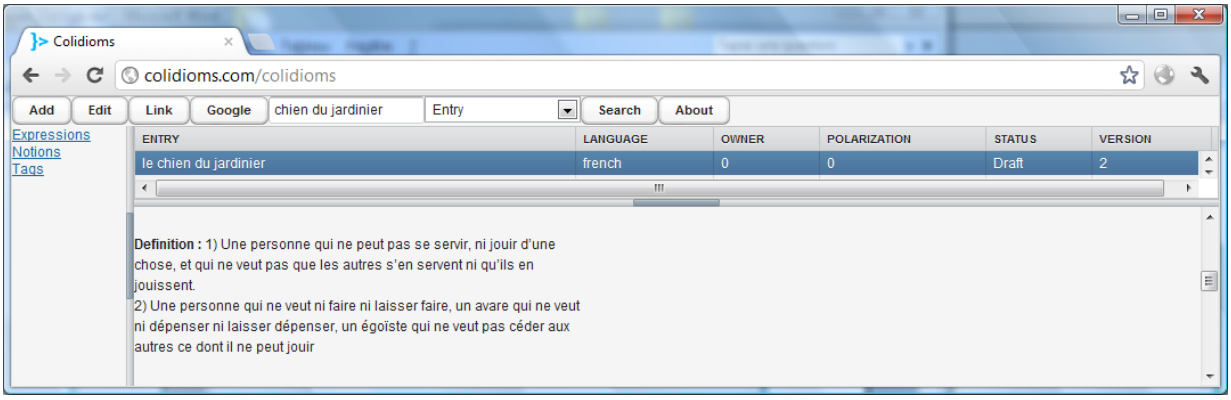

Figure 4. Definition. A Case Study of the French phraseme *le chien du jardinier*.

5. Search

Colidioms allows two types of search – semasiological and onomasiological.

Semasiological search, which is searching by lemma, is done by choosing *entry* from the list of search options. The example in the Figure 5 shows all the expressions that contain the word *chien* 'dog' in the entry.

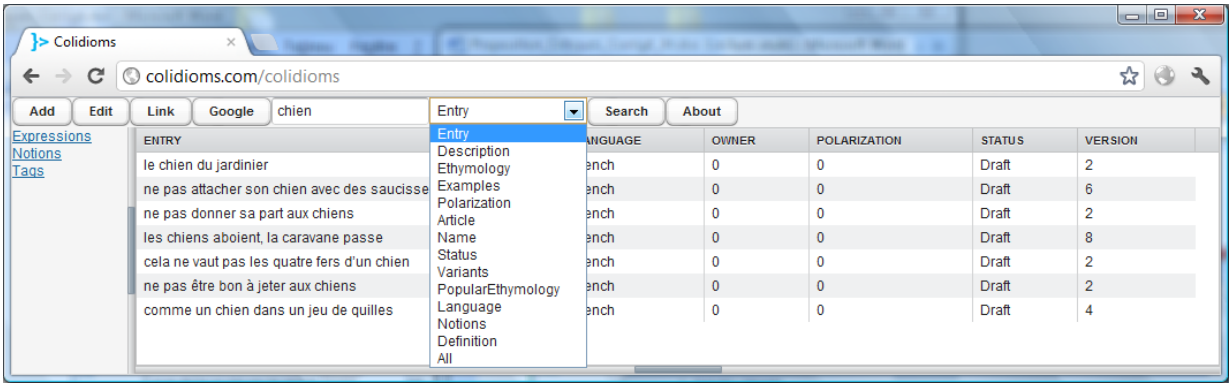

Figure 5. Semasiological search.

It is also possible to base the search on *etymology*, *examples*, *variants*, *definitions* or the combination of all these fields.

Onomasiological search is useful when the user cannot remember exactly the expression. It is about making complex queries based on the notions. For example, the user remembers only that there exists an expression speaking about protests and jealousy. He can type *protest* and *jealousy* and the software will return a list of results containing these notions (Figure 6).

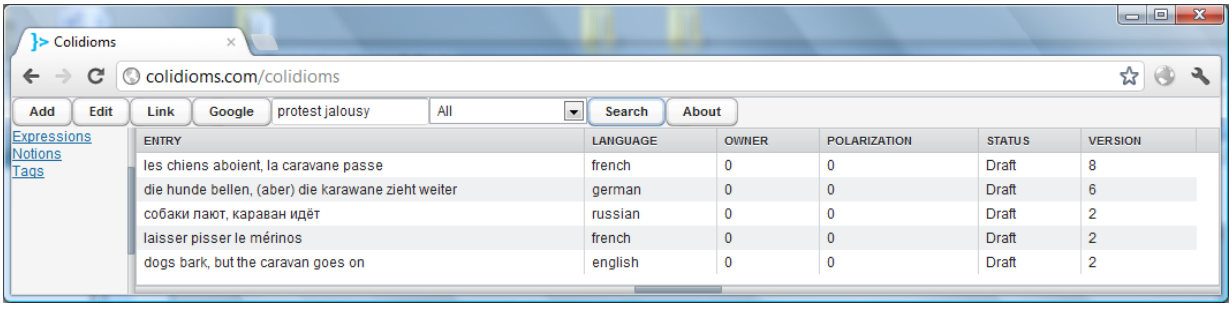

Figure 6. Onomasiological search.

6. Conclusion

Notions seem to be the most important component of Colidioms as they make it possible to carry out cross-lingual comparisons. They can be used to find a phraseme and compare it to other phrasemes, even in another language. The use of notions enables users to find expressions without knowing any of their key-words. By providing language agnostic templates, the application allows systematic classification of phraseological units. This facilitates the comparison of similar phrasemes across different languages.

In this respect, notions are particularly useful as they seem to apply to all four languages used in Colidioms. We do not claim that the current proposition has no drawbacks. Nonetheless, at this stage, notions seem to work well for this type of comparison. We don't know for sure if this principle will work if applied to a wide range of phrasemes in languages belonging to different language families. The greater the number of languages compared, the more complex the task.

References

A. Dictionaries

Баранов, А., Д. Добровольский, К. Киселева and А. Козеренко 2007. *Словарь-тезаурус современной русской идиоматики*. Москва: Мир энциклопедий Аванта +.

Urdang, L., W. W. Hunsinger and N. LaRoche 1998. *Picturesque expressions. A thematic dictionary*. USA: Verbatim Books.

B. Other literature

Ahmanova, O. 1974. *Word-combination: Theory and Method*. Moskva: MGU.