(Meta)phraseography and phraseomatics: DiCoP, a computerized resource of phraseological units

Lian Chen

ASIALEX
The Asian Association for Lexicography

ASIALEX 2023 SEOUL

ASIALEX 2023

The Asian Association for Lexicography

"Lexicography,
Artificial Intelligence,
and Dictionary Users"

Date: **June 22-24**, 2023

Venue: The Commons, Yonsei University

Host: 한국사전학회 KOREALEX · 연세대학교 국어국문학과 · BK21 FOUR 한국 언어·문학·문화 미래선도 글로벌인재 양성 교육연구단

Organizer: 언어정보연구원 Institute of Language and Information Studies

Main Sponsors: KISO 한국인터넷자율정책기구 Korea Internet Self-governance Organization · 연세대학교 YONSEI UNIVERSITY

Sponsors: NAVER · Saltlux · IIR TECH (주)아르테크 · 한국문화사 · 동아출판 · Timbel Timeless Label

FOREWORD

The theme of ASIALEX 2023 is Lexicography, Artificial Intelligence, and Dictionary Users. While proposals on any other topics related to the study and use of dictionaries are also welcome, ASIALEX 2023 aims to provide opportunities to discuss the changes and challenges that go beyond the realms of traditional lexicography and seek new directions and prospectives for lexicography and dictionaries to cope with social problems and changes. Dictionaries, including their accompanying resources and tools, technologies, platforms, and publication formats, have been continuously developing according to changes in the trends and cultural contexts of the times. As lexicography undergoes periods of transition, researchers have questioned the future of dictionaries and dictionary-maker, and even the EURALEX 2010 roundtable discussion on the theme 'Will there be people who make dictionaries in 2020?'. Now that we are well beyond 2020, fortunately, the activities of many associations and researchers in the field of lexicography remain strong and ongoing. Although commercial models based on profit structures of print dictionaries no longer exist, the demand for refined language resources and the power of language information seem to have become even stronger. The questions we are faced with are thus related to what opportunities as well as crises dictionaries and lexicography face. With this in mind, we look forward to discussing the cultural roles of lexicography and lexicographers, the value of language information in the AI era, and dictionary users themselves as major topics. The following points detail our intention to propose the theme of Lexicography, Artificial Intelligence, and Dictionary Users for ASIALEX 2023.

Dictionaries in the Age of Artificial Intelligence

In the current era of AI, dictionaries exist not just for human beings, but also for machines, and this shift urges us to deepen the discussion of theoretical lexicography and to expand the scope of dictionaries more flexibly. While the word has long been considered the basic unit of dictionary entries, it is now necessary to consider how to better adopt typically unregistered categories, such as neologisms, non-standard forms, loanwords, hate speech, slang, and pragmatic or nonverbal information, which have often been neglected in traditional lexicography. As Sinclair et al. (2004) referred to an ideal dictionary as containing all semantic units, it is time to consider the useful extensions and forms of a dictionary containing all such semantic units used in everyday communication.

Implication and Significance for and of Dictionary Users

Not only have the boundaries of what is considered a dictionary expanded. but the definitions of dictionary users have expanded as well. As the term 'machine readable' shows, nowadays dictionary users include machines as well as humans. Nonetheless, even dictionaries designed for machines ultimately aim to represent human intuition. For a dictionary to properly function as a medium connecting human intuition and machines, it is necessary to think about how to represent knowledge of the world more precisely.

Popularization of Lexicography and the Role of Professionals

Finally, we hope that this conference will lead to discussions on popularizing dictionaries and fostering subsequent generations of lexicographers. Dictionaries are found all around us, and they are used everywhere in our daily lives, although we may not be aware of their presence. Despite the achievements of lexicography throughout human history and the relatively recent corpus revolution (Rundell and Stock 1992, Rundell 2008, Hanks 2012), the study of dictionaries does not seem to be widely appreciated by the public. For the public in general, dictionaries are still difficult to use, and lexicography is an unknown area. Scholars and professionals in lexicography thus need to seek out the desired identity of dictionaries as required in modern times by approaching and interacting with the public. We hope that ASIALEX 2023 will present opportunities to diagnosing modern social communication problems by gaining a better understanding of the public use of language, and listening to the needs of a new, modern era with a more flexible attitude toward the structures, forms, and boundaries of lexicography and dictionaries

Hanks, P., 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography*, 25(4), pp.398-436.

Rundell, M., 2008. The corpus revolution revisited. *English Today*, 24(1), pp. 23-27.

Rundell, M. and Stock, P., 1992. The corpus revolution. *English Today*, 8(4), pp.45-51.

Sinclair, J., Jones, S. and Daley, R., 2004. *English Lexical Studies: The OSTI Report*. London: Continuum.

TABLE OF CONTENTS

Part I: Keynotes

Part II: Talks and Posters

Topic 1. Cultural and Societal Representation in Lexicography

Topic 2. Dictionary Use and User Studies

Topic 3. Semantic Representation in Lexicography

Topic 4. Bilingual and Multilingual Lexicography

Topic 5. Dialectal Representation in Lexicography

Topic 6. Dictionary-making Issues and Methods

Topic 7. Lexicology and Lexicography

Topic 8. Phraseology and Lexicography

Topic 9. Terminology and Specialised Dictionaries

Poster Session

Part III: Acknowledgements

Topic 8

Phraseology and Lexicography

(Meta)phraseography and phraseomatics: DiCoP, a computerized resource of phraseological units

Lian Chen

(Meta)phraseography and phraseomatics: DiCoP, a computerized resource of phraseological units

Lian CHEN 陈恋

PhD in Language Science at Paris-Cergy University (LT2D-Jean Pruvost Centre)

loselychen@gmail.com

Abstract

This article presents the Dictionary and Corpus of Phraseology (DiCoP) project, whose main objective is the development of a multilingual electronic dictionary of phraseology (currently French–Chinese and Chinese–French) relating to phraseological units. The current study focuses on Chinese to French, and among the different types of fixed expressions, more precisely on French idiomatic expressions and their Chinese correspondent, the *chéngyǔ*, both characterized by a high degree of fixedness.

This project comprises several innovative aspects: 1) digital (meta)phraseography; 2) phraseodidactics with DiCoP-Learning; 3) phraseotraductology and corpus with DiCoP-Text.

Keywords: DiCoP, digital dictionary, (meta)phraseography, phraseomatics

1. Introduction

The lexicon is all the words or sequences (Polguère 2002, Mortureux 2008), which include lexical phrases or polylexical units. Linguists call these phraseological units (PUs; González Rey 2002). In Chinese, the most widespread term to designate this fixedness is 熟语 *shúyǔ* (Sun 1989, Wang 2006, etc.). Phraseology is an essential linguistic and lexicultural phenomenon that "[carries] the idiosyncrasy of a culture, of a society, of a collective way of seeing things, of an idiomatic way of speaking" (González-Rey 2002, p. 40). The lexical units of a language, which include phraseology, are endowed with a strong cultural charge. The study of this specific idiomatic cultural phenomenon in phraseology is called "phraseoculturology" (Chen 2022a). A PU or 熟语 shúyǔ is a "… polylexical sequence consisting of two or more categorically related lexemes, whether contiguous or not" (Bolly, 2011: 28):

> Although they are different in length, content, scope of use, they have slowly become fixed in practice. Each *shúyǔ* has a specific meaning and its components cannot be taken literally. It has its own structural characteristics and cannot be changed at will (Cui 2005).

PUs are characterized linguistically by the following:

> (i) a certain degree of syntactic fixedness (blocking of transformational properties and unalterable constituent order); and/or (ii) a certain degree of semantic fixedness (at least partial non-compositionality); and/or (iii) a certain degree of lexical fixedness (paradigmatic restriction); and/or (iv) a constraint on use in a communication situation. (Bolly 2011, p. 28)

In both French and Chinese, PUs are "ubiquitous in everyday use" (Bolly 2011, p. 19). Thus, certain digital resources specific to PUs have emerged in France[1] and China[2], but these resources remain mostly monolingual, with an often fragmentary content and a scientific[3] character sometimes leaving something to be desired. Although bilingual paper dictionaries[4] exist, there is still no digital corpus (dictionary) on bilingual or even multilingual PUs that is sufficiently exhaustive.

It is in this context that this Dictionary and Corpus of Phraseology (DiCoP) project is developed[5], with the objective of creating an online electronic dictionary (initially French–Chinese/Chinese–French, and eventually multilingual), on the one hand, and the constitution of a corpus concerning the PUs and associated databases making it possible to know their

---

[1] Such as expressio.fr ;https://www.expressio.fr ; https://www.linternaute.com/, etc.

[2] Such as http://www.dffyw.com/cy/ ; http://www.hydcd.com/cy/ ; http://cy.5156edu.com/index.html

[3] For example: the common expressions "un regard d'aigle" (eagle-eyed perception), "une tête de cochon" (pigs head, a stubborn, obstinate person) do not always exist in these three online dictionaries.

[4] For example:

Chinese-French Dictionary: Doan, P. et Weng Z.-F. (1999); Sun, Q. (2012)

French-Chinese Dictionary: Yue, Y., Xiao, Z. (2000); Sun, Q. (2010); Cai, H.-B. (2014).

[5] Available on phraseologia.com, under development.

frequency of use (in newspapers, literary works, manuals, etc.), on the other. This is done to verify the vitality and use of PUs in practice, with the aim of providing easier access to them. These are "ready-made" expressions, such as collocations (peur bleue[6]; 骑自行车 *qízìxíngchē7*), proverbs (Pluie du matin n'arrête pas le pèlerin[8] ; 谋事在人，成事在天 *móushìzàirén, chéngshìzàitiān9*), IEs (un coup de main[10] ; 狐假虎威 *hújiǎhǔwēi11*), puns in Chinese (孔夫子搬家 – 净是书 [输] k*ǒngfūzībānjiā – jìngshìshū* [*shū*][12]*)*, and the "defrosting" of lexical frozenness (tout feu tout **femme,** defrosting of [être] tout feu tout **flamme,** 默默无蚊 *mòmòwú**wén***[13] defrosting of 默默无闻 *mòmòwú**wén***[14], etc.).

This project comprises several innovative aspects. First, from a (meta)phraseographic point of view, DiCoP's macrostructure is based on a query that is both mono- and interlingual. At the microstructure level, computerized dictionary data is enriched with a contrastive phraseoculture.

2. (Meta)phraseography and phraseomatics: The DiCoP lexicographic database

Phraseography is a branch of applied phraseology whose object of study is the development of collections, glossaries, and dictionaries of PUs. Metaphraseography (Murano 2011, Chen 2022a) is a (sub-)discipline whose objective is the study of the types of PUs dictionaries and the methods by which they are constructed, as well as an object of reflection and research on the phraseographic design. The theoretical and applied development of phraseology, phraseodidactics, phraseography, phraseoculturology, and new digital technologies mark the birth of a new interdisciplinary branch: phraseomatics (computer phraseography).

The first step is in creating the DiCoP to digitize the corpus of existing monolingual[15] and bilingual paper dictionaries of PUs whose lexicographic relevance is recognized. The architecture of the DiCoP is as follows:

2.1. Macrostructure

The electronic dictionary is classified not only in alphabetical order, but fixed expressions are also grouped by theme (animals, human bodies, etc...) for easier and more diversified queries:

a) Thematic classification: The lexical database currently includes the descriptions of approximately 2,400 PUs (currently mainly French idiomatic expressions and their Chinese equivalent *chéngyǔ*[16]) concerning animals and the human body. I am continuing this work and enriching the lexicographic corpus with additional areas, including colors, numbers, and plants.

b) Alphabetical classification: in each language or characters in Chinese according to pinyin.

The expression 绞尽脑汁 jiǎojìnnǎozhī is provided as an example:

---

[6] blue scare : terrified.

[7] ride a bike.

[8] Morning rain does not stop the pilgrim.

[9] The planning lies with man, the outcome with Heaven.; Man proposes, but God disposes.

[10] give a helping.

[11] the fox assuming the majesty of the tiger -- borrowing power to do evil.

[12] Confucius moves – books (chess) only/always be defeated.

[13] The character "蚊 wén" (mosquito) replaced "闻 wén" (hear), in an advertisement for incense against mosquitoes.

[14] unknown to the public; be completely unknown or unrecognized.

[15] For example: Rey, A., Chantreau, S. (2003) *Dictionnaire des expressions et locutions.* Paris : Le Robert.

Centre de recherche « Explication des lexies (词 cí) et des sinogrammes (字 zì) » (2014) *Grand dictionnaire des chéngyǔ chinois [*中华成语大词典 *zhōnghuá chéngyǔ dà cídiǎn].* Presse d'enseignement de la langue chinoise.

[16] Both constitute polylexical sequences, semantically non-compositional (they have an overall conventional meaning, which is generally not deducible from the meaning of the elements that compose them), syntactically characterized by their structure presenting a high degree of fixedness, and then by the fact that they do not always submit to the combinatorial rules that govern free syntax. Culturally, they are loaded with cultural implicits. For example in French: "avoir la tête dans les nuages" (to have the head in the clouds, to be distracted), etc.

In Chinese, syntactically, their basic form, which most often follows a fixed quaternary (quadrisyllabic) rhythm, phonetically and/or syntactically divided into two hemistiches, is conventional and unchanged for generations, hence the name *chéngyǔ*, "ready-made expressions". For example: 佛口蛇心 *fókǒu-shéxīn* (Buddha+mouth+snake+heart): Buddha's mouth, serpent's heart/honeyed words but evil intent.

French idiomatic expressions come from common usage or even popular tradition, whereas the *chéngyǔ* have their origin in literary, and are of a bookish or scholarly character, frequently with a strong allusive content.

c) Monolingual query: Two search functions exist, by keyword 脑 nǎo (brain/head) ①[17] or by full *chéngyǔ* ②.

b) Interlingual query ③: This function allows the user to find a literal translation (→), partial equivalent (◐), perfect equivalent (◉), metaphor (□), or finally, a quick explanation of the phraseoculture of the keywords.

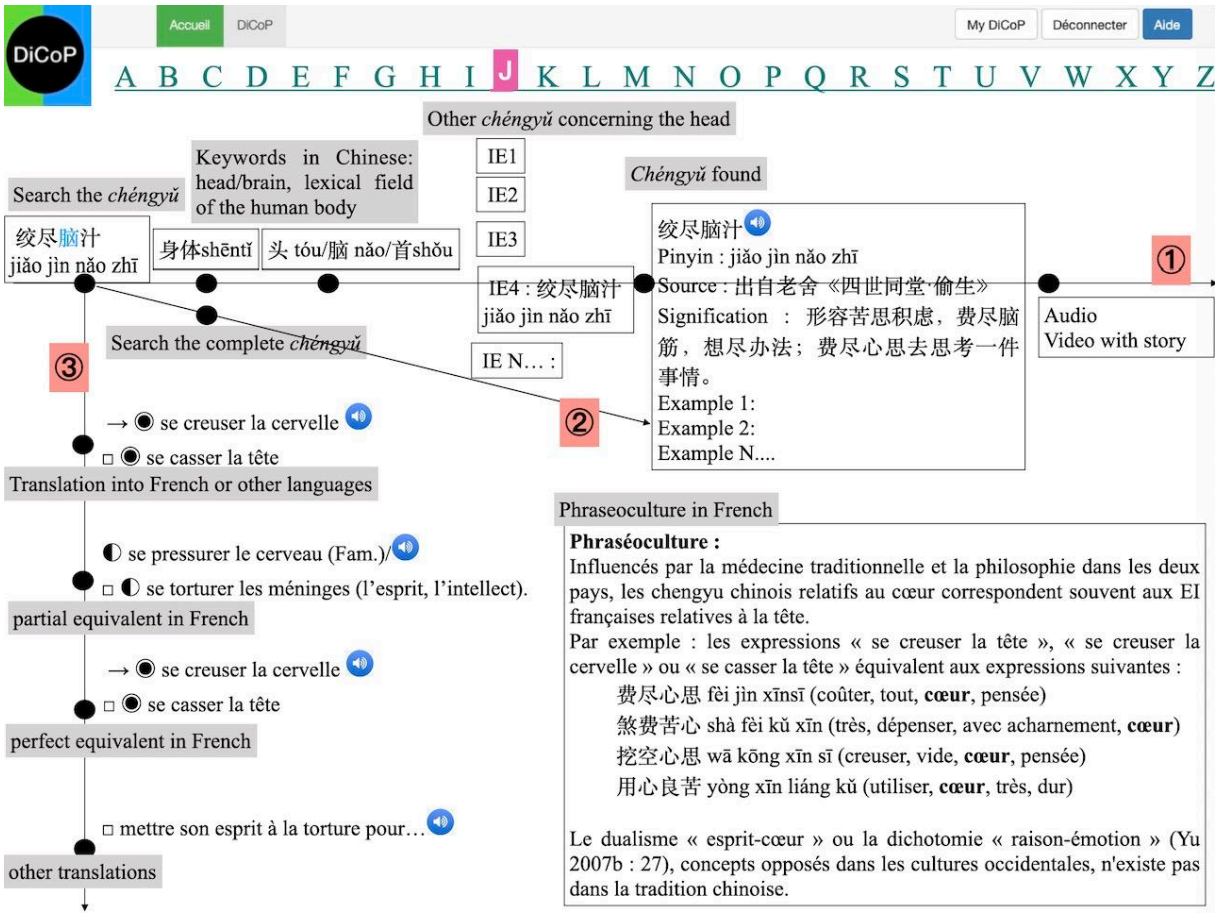

Figure 1. Phraseographic design of phraseological unit queries

2.2. Computerized microstructure

The DiCoP pays particular attention to phraseoculture (Chen 2022a), which is essential for applied phraseology, particularly for PUs falling under partial or non-equivalence. Indeed, the search for PUs carrying strong allusions (history, fable, mythology, culture, etc.) in bilingual (French–Chinese and Chinese–French) dictionaries of fixed expressions reveals that phraseoculture is not satisfactorily treated in specialized dictionaries (Chen 2022b, p. 19). The DiCoP offers a detailed microstructure by adding examples of uses, as well as a phraseocultural dimension, by providing quick information on the origin or concise history of these expressions, with the etymology and phraseoculture from one language to another.

The Extensible Markup Language (XML) standard based on the Document Type definition - Text Encoding Initiative (DTD-TEI) in XML for dictionaries has been retained, particularly with a view toward ensuring the sustainability of the resource. Below is an example for Chinese to French dictionaries of this DTD adapted to DiCoP.

```
<!ELEMENT woxinchangdan DEF_CONTENU>
<!ATTLIST woxinchangdan TYPE OBLIGATION VALEUR_DEFAUT>
<!ELEMENT entry (word, pinyin, audio, definition, source, history, video, examples)>
<!ELEMENT word (simplified, traditional, translation)>
<!ELEMENT definition (literal, implicit)>
<!ELEMENT examples (example+)>
<!ELEMENT simplified (#PCDATA)>
<!ELEMENT traditional (#PCDATA)>
<!ELEMENT translation  (#PCDATA)>
<!ELEMENT pinyin  (#PCDATA)>
```

[17] This query method is particularly well suited to French, given the problem of pronouns, determiners, etc. In the example "se casser la tête/casser la tête" (breaking your head), it is relevant to search for the keyword "tête" (head).

```
<!ELEMENT audio EMPTY>
<!ELEMENT literal (#PCDATA)>
<!ELEMENT implicit  (#PCDATA)>
<!ELEMENT source  (#PCDATA)>
<!ELEMENT history (#PCDATA)>
<!ELEMENT video  EMPTY>
<!ELEMENT example  (#PCDATA)>
<!ATTLIST audio src CDATA #REQUIRED>
<!ATTLIST video src CDATA #REQUIRED>
```

Each article includes a PU in simplified and traditional characters, the translation of each character, the literal translation, the implicit and metaphorical meaning, the source, the story, and pragmatic examples according to the following model:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE product SYSTEM "product.dtd">
<entry>
 <word>
  <simplified> <b> 卧薪尝胆 </b></simplified><br style="display:block;"/>
  <traditional><b> 臥薪嘗膽 </b></traditional><br style="display:block;"/>
  <translation>coucher, bois de chauffage, goûter, vésicule biliaire</translation><br/>
 </word><br/>
 <pinyin font="italic">wò xīn cháng dǎn</pinyin> <audio src="audio.mp3" /><br/>
 <definition><br/>
  <literal> un homme qui couche sur du bois de chauffage (sur de la paille) goûte une vésicule biliaire (du fiel) tous les jours</literal><br/>
  <implicit>entretenir le ressentiment et préparer la vengeance</implicit><br/>
 </definition><br/>
 <source> Selon l'ouvrage « 史记 Shiji » (Mémoires du Grand Historien ou Mémoires historiques) (109 - 91 av. J.-C.), cette histoire se déroule dans l'État de Yue (越国), en 494 av. J.-C. </source><br/>
 <history>Vaincu par l'empereur de l'État de Wu (吴国), l'empereur Gou-Jian de l'État de Yue (越) décida de prendre sa revanche. Afin de ne pas oublier l'opprobre qui avait couvert son pays déchu et de s'affermir dans sa résolution de se venger, il couchait la nuit sur de la paille et goûtait souvent la bile d'une vésicule biliaire suspendue au mur de sa chambre. Cet exercice de mortification le rendit plus fort et il finit par vaincre l'État de Wu.</history><br/>
 <video src="video.mp4" /> <br/>
 <example> 越王归国后，这才得以有机会“卧薪尝胆 ”、休养生息，几年后，一举将吴灭亡。</example><br/>
 <example> 我国历史上有过许多发愤图强的故事。最著名的，莫过于“卧薪尝胆 ”了。</example><br/>
</entry>
```

The DiCoP is currently being developed using the dictionary editor Lexonomy, recommended by EURALEX (the European Association for Lexicography), as shown in the figure below:

Figure 2. Example from the Lexonomy dictionary editor

As the database grows, it will be migrated to a dedicated website developed under PHP+MySQL. This will be a simple

change of format, with the actual content of the PU records remaining the same. Later, the DiCoP-Text corpus (under development) will make it possible to highlight the most frequently used PUs (as discussed below). We hope to develop an easy to use application in the future.

For each PU in a given language, the system will display different equivalents in the other languages, "ranked by a confidence value" (Garcia, García-Salido, Alonso-Ramos 2019, p.747). In addition to providing a multilingual perspective, the resulting dictionary can also serve as a monolingual resource, "thus being a useful application for both native speakers and language learners" (*Ibid.*).

In addition, the DiCoP will include two primary tools that linked to the maximization of resources: DiCoP-Learning (phraseodidactics) and DiCoP-Text (phraseotraductology).

3. Dictionary for the (self-)learning of PUs in a foreign language: DiCoP-Learning

Language is a vector of culture, and PUs are quintessential in their rich and colorful cultural content. PU learning should be part of a long-term process in foreign language and culture teaching (Gonzalez-Rey 2007, Sułkowska 2016, Chen 2021). DiCoP-Learning is thus a didactic tool, more precisely an electronic (self-)learning dictionary for PUs in foreign language (currently French–Chinese and Chinese–French). Its objective is to develop an improved program of self-learning and intra- and interlingual teaching of these PUs, according to their degree of opacity, through a progressive phraseodidactic and phraseographic methodology, from the beginner levels.

3.1. Presentation of DiCoP-Learning

The analysis and the comparison lead to the distinguishing three types of relationships from one language to another for the IEs. Based on nature identity (i.e. idiomatic expression), semantic identity, lexical identity, and structural similarity between the IEs of the two languages, the following typology was established: perfect equivalents, partial equivalents, and non-equivalents. In DiCoP-Learning, the chosen *chéngyǔ* primarily involve perfect equivalence and partial equivalence according to themes (human body, animals, colors, numbers, etc.) and linguistic and cultural correspondence, to adapt to the level of the students.

A literal translation (→), as faithful and close as possible to the often imagined meaning of the characters, is proposed first. Indeed, it is this that most often makes it possible to apprehend the metaphorical meaning characteristic of most *chéngyǔ*. The implicit and/or figurative meanings (□) are then indicated, and finally, if possible, an equivalent. Below are examples concerning the perfect equivalent (⦿)[18] and partial equivalent (◐)[19]relating to the human body:

耳听八方 [耳聽捌方] (**oreille**, écouter, huit, directions)	ěr tīng bā fāng	→ une oreille écoute dans huit	□ ⦿ tendre l'**oreille** à tout	
多嘴多舌 (beaucoup, **bouches**, beaucoup, **langues**)	duō zuǐ duō shé	→ beaucoup de bouches, beaucoup de langues	□ ◐ ne pas savoir tenir sa **langue**	□ ◐ avoir la **langue** bien pendue

Table 1. Examples of *chéngyǔ* in the DiCoP-learning

Unlike traditional teaching, non-equivalent *chéngyǔ* (such as 塞翁失马[20]) are not integrated at the beginner level. Although they are often more interesting and contain the deep culture and stories of the target language, they are the most difficult to learn.

Based on the premise that difficult words should not be included in DiCoP-Learning and that the expressions should always be presented in context, the advantages of computers allow us to select the most common and simple examples among the corpora to avoid "the keyword in an example sentence [carrying] a different meaning from what it aims to illustrate" (Tian, J. Huang, F. Huang 2018).

For partial equivalence PUs, the absence of phraseoculture can confuse students, due to the difference between word-for-word or global translation or equivalence between the two languages with terms that are nonetheless different (heart in Chinese and head in French, for example, linked to different philosophical and medical substrates). Thus, a short historical

[18]i.e when there is an identity of nature (EI), the same metaphorical meaning, the same keyword(s) concerning the thematic such as the human body, animals, etc. between the expressions in the two languages.

[19] i.e. when the expressions have the same identity of nature (proverbs, idiomatic expressions, ..) in both languages or only in one of them, provided that it is a fixed expression in the other. We will consider as partial equivalents expressions that are idiomatic in both languages but present differences on other "secondary" criteria (semantics, keywords or structure).

[20] 塞翁失马 *sàiwēngshīmǎ* : the old man of frontier lost his horse; misfortune might be a blessing in disguise.

Story: From a tale inserted in the "Huainanzi 淮南子", a collection of chapters dealing with various subjects, written in the 2nd century BC, under the Western Han at the initiative of Liu An (-179~-122): An old man, who lived near the border, lost his horse one day. A few months later, the animal returned, bringing with it another steed. The son of this old man, while riding this horse, lost his leg. But this accident turned into a piece of luck, as it allowed him to escape the war.

or cultural addition for expressions with strong allusive content would facilitate the understanding and memorization of these expressions and reinforce the role of the dictionary in terms of cultural transmission. Figure 3 presents an example:

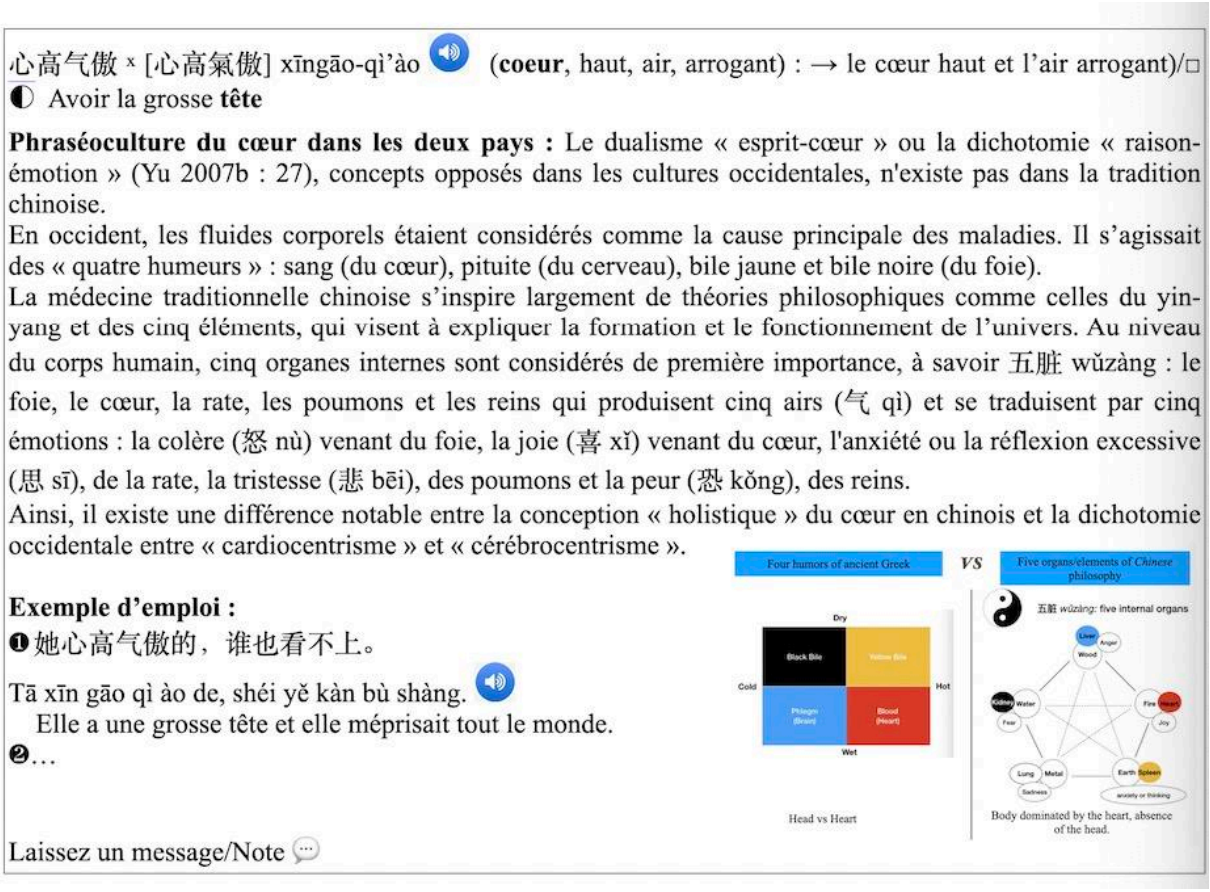

Figure 3. Illustration of phraseoculture in DiCoP-Learning

Furthermore, "corpus linguistics has played a crucial role in lexicography studies and provides valuable data for dictionary making and research." (Shen 2010, p.1) It is possible to create large linguistic databases quickly and inexpensively and to analyze the data they contain in an efficient and highly sophisticated way (Rundell 2010, p. 368). Thus, frequency information will be provided in DiCoP-Learning, which will include an online "library of examples" through the DiCoP-Text tool, allowing users to tap into a pool of additional examples. They will be translated into the target language and enriched with contextual information to ensure accurate understanding and appropriate use "to avoid ambiguity or even misunderstanding possibly caused by cultural gaps…" (Wang, Shen, Guo, 2018). In terms of use cases, it is important to teach these PUs in context, with a view that is both receptive and productive of repetition and of multiplying the use of this lexicon to facilitate learning. As Loock (2018, p. 21) states, the aim is to make learners aware of "the usefulness of a (quality) linguistic database, in other words a corpus, making it possible to verify the relevance of the equivalents found and their use in context."

3.2. Didactic and self-learning of DiCoP-Learning: Integration into digital education

Finally, the objective is to combine theory and practice to allow students to better understand the content and implement independent learning. DiCoP-Learning can provide fun exercises and games on digital educational platforms (Moodle+H5P, Chamilo, etc.), with ergonomic tools, to allow students to memorize and practice PUs from the beginner level. Figure 4 presents some examples.

Figure 4. Examples of playful exercises on *chengyǔ* using Moodle+H5P as part of DiCoP-Learning

4. Corpus of phraseotraductology: DiCoP-Text

Phraseology is particularly difficult in translation, as it is influenced by linguistics, culture, and stylistics and strongly reflects the translator's choices and translation technique. DiCoP-Text will be a database (monolingual, parallel, multilingual corpus) making it possible to know the frequency of use of PUs to verify their vitality in practice. It should make it possible to easily examine the use of PUs in translations for lexicometric studies and scientific research in the field of automatic phraseotranslation and natural language processing (NLP).

Text collection and digital processing is currently underway. The choice of corpus should reflect the linguistic diversity of the language and therefore be large enough to ensure adequate representation of the PUs and improve the accuracy and reliability of the DiCoP. It is thus necessary to represent a variety of genres (literature, poetry, newspapers, speeches, etc.) to find a balance between formal and informal language (official, academic texts, as well as dialogues, daily conversations, etc.) and represent a variety of speakers (native or not, regional, etc.) in order to cover the dialectal variations and regional uses of PUs, for example. Most of the corpus will comprise modern resources (20th century and later) to verify the liveliness and frequency of PU usage. Nevertheless, an exception will be made for the Chinese corpus with the incorporation of the four great classics, which are already translated into many languages and are full of PUs. Furthermore, in French will include great authors of the 19th century (Victor Hugo, Honoré de Balzac, etc.) to reflect the evolution of PUs since that time. We favor references already accessible in digital form (or already ocerized) according to the principle recalled by Nelson (2010, p. 61): "the best source of texts for corpus usage are others, pre-existing corpora". When compiling the list of works, the Frantext and Wikisource textual databases for French and the Beijing corpus centre for Chinese, for example, will be consulted.

We are also interested in newspapers in both countries, such as the *People's Daily* in Chinese, *Xinhuashe,* and *Canard enchaîné* in French, as well as social networks such as Twitter and Weibo. One of the interests of newspapers is the regular presence of "defrosting" (défigement in French; 活用 *huóyòng* in Chinese), a phenomenon that one encounters, for example, through puns contained in titles, slogans, and advertisements .

Large corpora in each language (initially Chinese and French) will be compiled and analysed using NLP tools to obtain morphosyntactic and syntactic information. Next, we will apply different measures to automatically select the candidate PUs from the corpora. We will then use cross-linguistic models able to identify the equivalents of a given PU in other languages. Instead of using whole parallel corpora, only sentences concerning PUs to provide context for improving automatic phraseotranslation will be retained, "thus avoiding the need of obtaining large parallel texts for each language pair" (Artetxe and al., 2018)[21]. For each PU in a source language, the resulting dictionary will provide a set of equivalents in the target languages, ranked by a confidence value that represents the translation probabilities.

5. Conclusion

Paper dictionaries only meet the needs of foreign language learners to a limited extent and are no longer in step with the digital age. With the theoretical and applied development of phraseology, phraseodidactics, phraseography, phraseoculturology and new digital technologies, we must pay more attention to the teaching and learning functions of PUs, as well as the translation of dictionaries. This also marks the birth of a new generation of PU dictionaries, and even of a new interdisciplinary branch: phraseomatics (computer science of phraseology or computer science of phraseography). Thanks to computers, the lexicographer's work is significantly facilitated; they can surf from one dictionary to another with a click, provided that these dictionaries are digitized.

The DiCoP database will be accessible online via a dedicated internet portal, benefiting from a query module responding to the purposes of the general public and researchers. It aims to become a two-way, free, and open interactive platform on the internet where users can obtain information, interact with it through comments (via "My DiCoP"), and even disseminate it when and where they want. By developing this new type of digital dictionary in the digital age, we wish to contribute to

[21] Quoted by Garcia, García-Salido, Alonso-Ramos (2019 : 748).

the “paradigm shift that is needed in the field of lexicography” (Binon and Verlinde, 2008, p. 17), centered on PUs.

The contrastive and translatological approach to IEs is still not very thorough, and substantive work in bilingual lexicography remains rare or is devoid of linguistic approaches. Indeed, most existing work fails to integrate a textual approach or an analysis that considers the linguistic context and the variants of the phrases. The constitution of a bilingual contextual dictionary of phrasemes is therefore innovative and original in the current scientific context. Multilingual resources for PUs and other multi-word expressions, such as IEs, are rare, despite some usefulness. In this respect, building multilingual dictionaries of PUs “is a hard task which requires a huge effort from expert lexicographers in different languages” (Orenha-Ottaiano, 2017).

References

Binon, J., Verlinde, S. (2008) Lexicographie pédagogique, des principes théoriques à la pratique. In *Les Cahiers de la Recherche,* UER-Langues de HEC-ULg, n° 2, pp. 8-19.

Bolly, C. (2011) *Phraséologie et collocations. Approche sur corpus en français L1 et L2*. Bruxelles, New-York: Peter Lang.

Cai, H.-B. (2014) *Dictionnaire explicatif des expressions et locutions françaises [*法语成语解析词典 *fǎyǔ chéngyǔ jiěxī cídiǎn],* Beijing Commercial Press.

Chen, L. (2022b) Phraseoculture in the construction of the corpus of the DiCoP: The treatment of the phraseographic microstructure. In *Proceedings of the International Conference EUROPHRAS: 4th International Conference ‘Computational and Corpus-based Phraseology, 28-30 September 2022*. Malaga, Spain, pp. 17-25.

Chen, L. (2022a) Phraséoculturologie: une sous-discipline moderne indispensable de la phraséologie. In *SHS Web of Conferences*, n° 138, 04011, pp. 1-18.

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.

Cui, X.-L. (2005) *Chinese Phraseological Units and the Representation of Humanity in Chinese* [汉语熟语与中国人文世界 Hànyǔ shúyǔ yǔ zhōngguó rénwén shìjiè]. Beijing: University language and culture press.

Doan, P. et Weng Z.-F. (1999) *Dictionnaire de chéngyǔ : idiotismes quadrisyllabiques de la langue chinoise*. Paris : Librairie You-Feng.

Garcia, M. García-Salido, M. Alonso-Ramos, M. (2019) Towards the Automatic Construction of a Multilingual Dictionary of Collocations using Distributional Semantics. In *Proceedings of eLex 2019, The sixth biennial conference on electronic lexicography, eLex 2019, 1–3 October 2019*. Sintra, Portugal, pp. 747-762.

Gonzalez Rey, M.-I. (2002) *La phraséologie du français*. Toulouse: Mirail University Press.

Gonzalez Rey, M.-I. (2007) *La didactique du français idiomatique*. E.M.E.

Loock, R. (2018) Les traducteurs sont-ils des linguistes comme les autres ? L’intégration des outils de corpus dans la formation des futurs traducteurs. In *Myriades*, pp. 18-34.

Mortureux, M.-F. (2008) *La lexicologie entre langue et discours*. Armand Colin.

Murano, M. (2011) Le traitement des Séquences Figées dans les dictionnaires bilingues français-italien, italien-français. Edición en Francés.

Nelson, M. (2010). Building a written corpus. In O’Keeffe, A. & Mc Carthy, M. (eds.) *The Routledge Handbook of Corpus Linguistics, Routledge,* pp. 53-65.

Orenha-Ottaiano, A. (2017) The compilation of an online Corpus-Based bilingual Collocations Dictionary: motivations, obstacles and achievements. In Proceedings of eLex 2017, Electronic lexicography in the 21st century: Lexicography from Scratch, 19-21 September 2017. The Dutch Language Institute, pp. 458–473.

Polguère, A. (2002) *Notions de base en lexicologie*. Observatoire de Linguistique Sens-Texte.

Rundell, M. (2010) Taking Corpus Lexicography to the next level: Explicit use of corpus data in dictionaires for language learners. In Zhang, Y.-H. (eds.) *Learner's Lexicography and Second Language Teaching*. Shanghai: Foreign language education press, pp. 367-386.

Shen, Y.-Y. (2010), EFL Learners Synonymous Errors: A Case Study of Glad and Happy. In *Journal of Language Teaching and Research*, Vol. 1, No. 1, Academy publisher Manufactured in Finland, pp. 1-7.

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

Sun, Q. (2012) *Nouveau dictionnaire chinois-français des locutions et proverbes*. Xia Men: University Press.

Sun, Q. (2010) *New French-Chinese Dictionary of Phrases and Proverbs [Nouveau Dictionnaire Français-Chinois des Locutions et Proverbes]*. Xiamen: University Press.

Sun, W.-Z. (1989) *Chinese phraseology [*汉语熟语学 *hànyǔshúyǔxué]*. Jilin: Changchun Educational Press.

Tian, B., Huang, J., and Huang, J.-H. (2021) Huang, J. ’s ”Grand contemporary chinese-french dictionary (2014)” and the story behind it. In Cybèle Berk (eds.) *Dictionnaires et apprentissage des langues*. France: Editions des archives contemporaines, pp. 103-119.

Wang, Q. (2006) *Chinese phraseology [*汉语熟语论 *Hànyǔ shúyǔ lùn]*. Shangdong: Educational Press.

Wang, X., Shen H. and Guo, Y. (2021) La métaphore dans les dictionnaires bilingues d’apprentissage - L’exemple des dictionnaires français-chinois. In Berk, B. (eds.) *Dictionnaires et apprentissage des langues*. France: Editions des archives contemporaines, pp. 79-88.

Yue, Y., Xiao, Z. (2000) *French-Chinese Dictionary of Phrases and Proverbs [Dictionnaire Français-Chinois des Locutions et Proverbes]*. Shanghai:Translation Publishing House.