ASIALEX 2023
The Asian Association for Lexicography

“Lexicography, Artificial Intelligence, and Dictionary Users”

Date: June 22-24, 2023
Venue: The Commons, Yonsei University
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 perspectives 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-makers, 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.

# TABLE OF CONTENTS

## Part I: Keynotes

Automating the creation of dictionaries: are we nearly there?

**Michael Rundell**

The ROI of AI in Lexicography

**Erin McKean and Will Fitzgerald**

Research into dictionary use in an era of e-lexicography

**Yukio Tono**

The Development of Naver Dictionary’s User Participation - A case study of Open Dictionary PRO and Accentia

**Jonghwan Kim**

## Part II: Talks and Posters

### Topic 1. Cultural and Societal Representation in Lexicography

The dictionarization of the feminist lexicon

**Judit Freixa and Sabela Fernández-Silva**

What’s in a name? Onomastics, identities, and Philippine dictionaries

**Jesus Federico C. Hernandez**

User-Oriented Toponym Documentation: Lexicography Perspective

**Winda Luthfita and Adi Budiwiyanto**

### Topic 2. Dictionary Use and User Studies

Developing “Can do” descriptors for L2 dictionary use: A preliminary version

**Naho Kawamoto and Yukio Tono**

Inaccuracy of an E-Dictionary and Its Influence on Chinese Language Users

**Fanfei MENG, Xi WANG, Shiyang ZHANG, and Lan LI**

Don’t throw your paper dictionary away! Using different types of dictionaries for improving EFL vocabulary learning

**Pasqualina Sorrentino and Massimo Salgaro**

### Topic 3. Semantic Representation in Lexicography

Varieties of English and their inclusivity in the Naver English Dictionary

**Vincent B.Y. Ooi**

Thesauri and ontologies: What is their relationship?

**Maria Koliopoulou**
**Topic 4. Bilingual and Multilingual Lexicography**
Li Fei and Hansan Kim ................................................................. 150

Building a multilingual learners’ idiomatic dictionary
Elena Berthemet ................................................................. 161

**Topic 5. Dialectal Representation in Lexicography**
At the interface between “the Good Book” and “the wordbook” : Bible translations and lexicography
Mats-Peter Sundstrom and Marlene Nilsson ................................................................. 168

Jejueo talking dictionary: A collaborative online database for language revitalization
Moira Saltzman ................................................................. 173

**Topic 6. Dictionary-making Issues and Methods**
Lexicographic treatments of Bangla adjectival affixes
Syed Shahrier Rahman and Mithun Banerjee ................................................................. 179

The Treatment of Selected Function Words in Monolingual Filipino Dictionaries
Elsie Marie T. Or ................................................................. 184

Challenges of Compiling a Monolingual Dictionary in a Multilingual Setting: Reports from a Filipino Dictionary Project
Ma. Althea Enriquez ................................................................. 192

Building Project Marayum (marayum.ph) : Lexicographic Issues and Solutions
Samantha Jade Sadural ................................................................. 198

**Topic 7. Lexicology and Lexicography**
A Study on Lexicon Information and Learner Acceptance of Korean Loanwords
Qihui Fan and Sun-Woo Chang ................................................................. 208

Making known the what and the why - On foregrounding cultural information in bilingual lexicography
Cuilian Zhao ................................................................. 214

**Topic 8. Phraseology and Lexicography**
(Meta)phraseography and phraseomatics: DiCoP, a computerized resource of phraseological units
Lian Chen ................................................................. 224
**Topic 9. Terminology and Specialised Dictionaries**

What Academic Words Refer to in Specialized Texts and in Specialized Dictionaries?

**Ping-Yu Huang and Yueh-Tzu Chiang** ................................................................. 233

How many terminologies should be recorded in dictionaries? - A case study of the OED

**Yongwei Gao** ........................................................................................................ 240

Avoiding “A Certain Kind of Plant” : A Case Study in Multidisciplinary Approaches to Lexicography

**Eric G. Englert and Sadaf Munshi** ................................................................... 245

Learner’s LSP Dictionary for Medical Coordinators : a Lexicographical Concept

**Elizaveta Krivetskaya and Alexey Matyushin** .................................................... 250

A Study on practice of North and South Korean infectious disease glossary compilation

**Juwon Park, Sukjeong Kim, Shin Ha, Wonyoung Do, and Young Hoon Kim** ........ 257

**Poster Session**

A Preliminary Access Structure Comparison-Description of the KWF Diksiyonaryo ng Wikang Filipino (online version) and diksiyonaryo.ph

**April J. Perez** ........................................................................................................ 263

A Study on Biased Expressions in Learners’ Dictionaries: Based on Examples from the Korean Basic Dictionary

**Chaerin Jang and Jong Won Yoon** ................................................................. 269

Revisit Agent Focus in Sakizaya: from a corpus-based approach

**Chihkai Lin** ........................................................................................................ 275

Ethno-cultural realia in the dictionary of a minority language (the case of Ket)

**Elizaveta Kotorova and Andrey Nefedov** .......................................................... 279

Classifying Korean Unregistered Words According to Registers and Exploring Their Potential As Dictionary Headwords

**Jinsan An, Yelin Go, Minkyu Sung, and Kilim Nam** ............................................ 283

On the Cognitive Linguistic Description of Polysemy in the 17th- and 18th-century English Lexicography

**Masaaki Ogura** .................................................................................................... 288

A study on the aspects of real phonetic variations in messenger conversations and its phonetic notation scheme in dictionary

**Miae Ahn, Sion Park, Nina Marie Victoire Constance Lee, and Juhyeon Ahn** ........ 292

Issues and Challenges in Lexicography: A Comparative Study on Burushaski, Kashmiri and Mankiyali(work in progress)

**Sadaf Munshi and Eric Englert** ................................................................. 298
Instagram Terminology: The Creation of a Project
Saghar Sharifi .................................................................................................................. 304

What Vocabulary Expresses Emotionally in Korean? - Focusing on Building a Korean Emotion Dictionary -
Yeonji Jang, Yejee Kang, Seoyoon Park and Hansaem Kim ............................................. 313

Compilation of Korean Dialect Dictionary Based on User Participation: Focusing on ‘Yanbian Malmoi’
Yinxia Huang, Jingri Cui, Meihua An, and Xian Piao ................................................... 317

Part III: Acknowledgements

<table>
<thead>
<tr>
<th>Host</th>
<th>325</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizer</td>
<td>325</td>
</tr>
<tr>
<td>Main Sponsors</td>
<td>325</td>
</tr>
<tr>
<td>Sponsors</td>
<td>325</td>
</tr>
<tr>
<td>General Chairs</td>
<td>326</td>
</tr>
<tr>
<td>Local Organizers</td>
<td>326</td>
</tr>
<tr>
<td>Scientific Committee</td>
<td>327</td>
</tr>
<tr>
<td>Student Committee</td>
<td>328</td>
</tr>
</tbody>
</table>
(Meta)phraseography and phraseomatics: DiCoP, a computerized resource of phraseological units

Lian Chen
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, phraseomastics

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 "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 2022). A PU or 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 France1 and China2, but these resources remain mostly monolingual, with an often fragmentary content and a scientific character sometimes leaving something to be desired. Although bilingual paper dictionaries3 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 developed4, 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.
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:
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; 骑自行车 qìxìngchē), proverbs (Pluie du matin n'arrête pas le pèlerin; 常用在天 móushìzàirén, chéngshìzàitiān), IEs (un coup de main; 狐假虎威 hújiáhǔwéi), puns in Chinese (孔夫子搬家 – 净是书 [輸] kǒngfūzǐbājiā – jīngshìshū [shū]; and, the “defrosting” of lexical frozenness (tout feu tout femmelain, defrosting of [être] tout feu tout flamme, 默默无闻 mòmòwúwén) defrosting of 默默无闻 mòmòwúwén, 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 acontrastive 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 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 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 jiǎojǐnlóuzhǐ is provided as an example:

---

6 blue: scared.
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.


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'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 腦 (brain/head) ① or by full 頭痛 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:

```xml
<!ELEMENT woxinchangdan DEF_CONTENU>
<!ATTLIST woxinchangdan TYPE_OBLIGATION VALEUR_DEFAULT>
<!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).
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
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE product SYSTEM "product.dtd">
<entry>
  <word>
    <simplified><b>嶎薪晄胆</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/>
    <pinyin font="italic">wò xīn cháng dǎn</pinyin><audio src="audio.mp3"/>
    <definition>
      <literal>
        un homme qui couche sur du bois de chauffage (sur de la paille) goûte une vésicule biliaire (du fiel) tous les jours
      </literal>
      <implicit>entretenir le ressentiment et préparer la vengeance</implicit>
    </definition>
    <source>Selon l'ouvrage «史記 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>
    <history>Vaincu par l'empereur de l'État de Wu (吳國), 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'affirmer 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>
    <example>
      越王嶎薪晄後, 息客生息, 几年后, 一举将吴灭亡。
    </example>
    <example>
      我国历史上有过多发愤图强的故事, 最著名的是, 莫过于“卧薪尝胆”了。
    </example>
  </word>
</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:

![Dictionary and Corpus of Phraseology (DiCoP)](image)

As the database grows, it will be migrated to a dedicated website developed under PHP+MySQL. This will be a simple (Meta)phraseography and phraseomatics: DiCoP, a computerized resource of phraseological units.
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” (García, 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 (*) and partial equivalent (○) relating to the human body:

<table>
<thead>
<tr>
<th>Chinese</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>耳听八方 [耳聾聨方] (oreille, écouter, huit, directions)</td>
<td>une oreille écoute dans huit l'oreille à tout</td>
</tr>
<tr>
<td>多嘴多舌 (beaucoup, bouches, beaucoup, langues)</td>
<td>beaucoup de bouches, beaucoup de langues</td>
</tr>
</tbody>
</table>

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-forward 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.

Lian CHEN 陈恋
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](image)

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.

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

---

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

The contrastive and translational 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 phraseons 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” (Orena-Ottaiano, 2017).

References


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