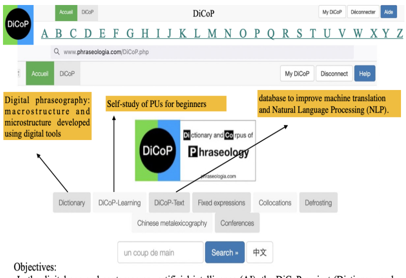
DiCoP project and DiCoP-Text corpus for the enrichment of Language Models and Automatic Translation

Lian Chen, Wenjun Sun, Flora Badin

EURALEX XXI CROATIA 2024 — INSTITUTE FOR THE CROATIAN LANGUAGE — EURALEX European Association for Lexicography

DiCoP project and DiCoP-Text corpus for the enrichment of Language Models and Automatic Translation

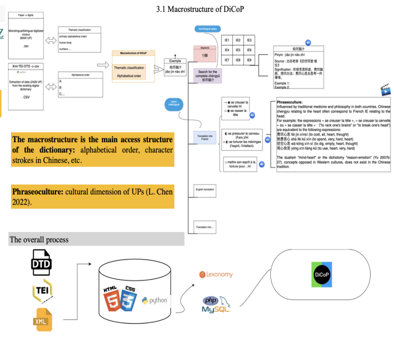
CHEN Lian 陈恋 loselychen@gmail.com
SUN Wenjun 孙文君 wenjun.sun@univ-lr.fr
BANDIN Flora flora.badin@univ-orleans.fr



Digital phraseography: macrostructure and microstructure developed using digital tools
Self-study of PUs for beginners
database to improve machine translation and Natural Language Processing (NLP).
Dictionary and Corpus of Phraseology

Objectives:
In the digital age and contemporary artificial intelligence (AI), the DiCoP project (Dictionary and Corpus of Phraseology (available on www.phraseologia.com, under development) **aims to develop a digital dictionary of multilingual phraseology** (initially French-Chinese/Chinese-French, and eventually multilingual), based on a corpus of phraseological units (PUs), in order to determine their frequency of use (in newspapers, literary works, school textbooks, etc.) in practice and therefore their vitality, to improve their automatic translation, and with the aim of facilitating access to phraseological units.

3.1 Macrostructure of DiCoP

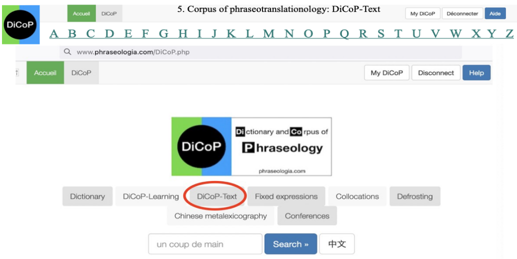


3.2 Computerized microstructure of DiCoP

Lexonomy
https://lexonomy.elex.is/
An online dictionary writing system for writing and publishing dictionaries.
elexis
https://elex.is/
Dictionary and Corpus of Phraseology (DiCoP)
PHP: Hypertext Preprocessor, better known by its acronym PHP, is a free programming language, mainly used to produce dynamic web pages via a web server.
MySQL: a relational database management system.

3.2 Computerized microstructure of DiCoP

5. Corpus of phraseotranslatology: DiCoP-Text



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, contrastive, multilingual corpus) to determine the frequency of use of UPs in translations, in order to verify their vitality in practice. It should make it possible to easily scrutinize the use of UPs in translations, for lexicometric studies and contribute to scientific research in the field of automatic phraseological translation and NLP (natural language processing).

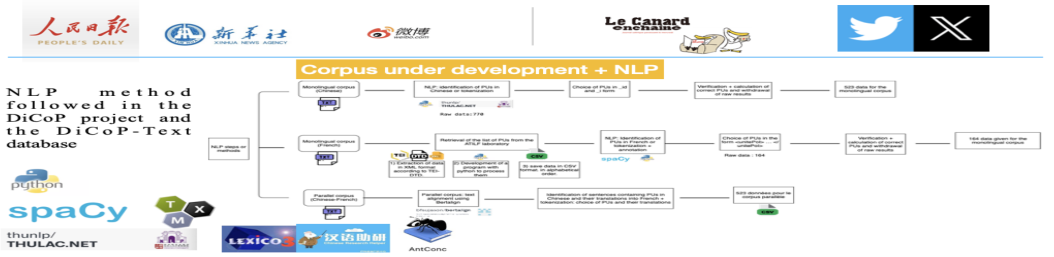
5. Corpus of phraseotranslatology: DiCoP-Text

The collection of texts and digital processing are currently underway. The choice of the corpus should reflect the linguistic diversity of the language and therefore be broad enough to ensure adequate representation of the PUs and improve the accuracy and reliability of the DiCoP.

- represent a variety of genres (literature, poetry, speeches, novels, science fiction, etc.) to find a balance between formal and informal language (official, academic texts, as well as dialogues, everyday conversations, etc.).
- modern resources (20th century and later)
- references already accessible in digital form (or already OCRed) : HTML, XML, TEI, etc.

Newspapers: "défigement" (活用 huóyòng in Chinese), a phenomenon that can be seen, for example, in wordplay in headlines, slogans and advertisements.

NLP method followed in the DiCoP project and the DiCoP-Text database

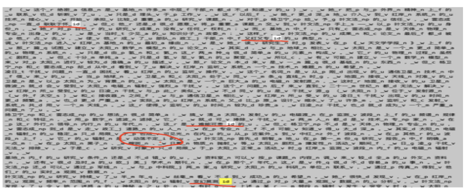


NLP: Identify PUs in Chinese or token


```
##终端命令: cd ~/Documents。 然后python3 import\ thulac.py



import thulac
thu1 = thulac.thulac()
#t = '''"红色联合"对"四·二八兵团"总部大楼的攻击已持续了两天, 他们的旗帜在
#大楼顶上出现了一个娇小的身影, 那个美丽的女孩子挥动着一面"四·二八"的大旗,
#"红色联合"的战士们欢呼起来, 几个人冲到楼下, 掀开四. 二八的旗帜, 抬起下面
#'''

f=open("texte_ch.txt","r")
f_out=open("text_ch_tok.txt","w")
for line in f:
    text = thu1.cut(line, text=True)
    f_out.write(text+"\n")
    #print(text)

f.close()
f_out.close()
```

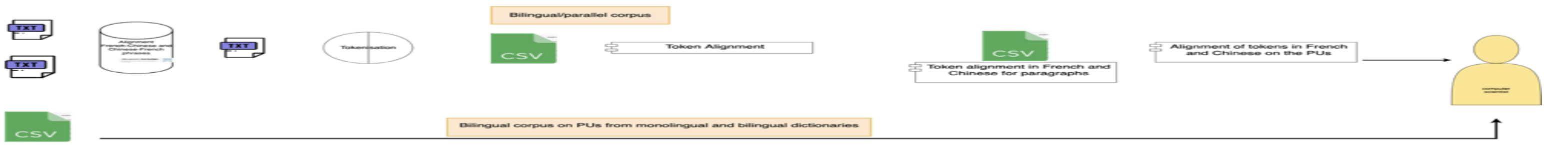
Liu Cixin's *Three Body Problem*. That's 186,079 words.

python
thunlp/ THULAC.NET



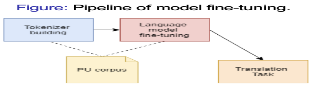
	Chinese PUs identified by Thulac			
Total	771			
Success rates	65.43% (504/771)			
Error rate	34.57% (267/771)			
	Form _id (567 PUs)	Form _i (204 PUs)		
	Correct	Error	Correct	Error
Number of IEs	321	246	183	21
Percentage	56.61 %	43.39%	89.71%	10.29%

Table: Sentences containing the 549 PUs, and their translation



DiCoP-Text for the enrichment of language models and automatic trai

Figure: Pipeline of model fine-tuning.



- Update the vocabulary of the tokenizer of the pre-trained model.
- Renew the model's embedding layer.
- Fine-tune the model with all PUs and sentence training set.
- Validate the model with sentence test set.

DiCoP-Text for the enrichment of language models and automatic translation

- Basic language models[1]: Mbart (Tang et al., 2020), M2m100 (Fan et al., 2021), Nllb (Costa-juss`a Marta R. et al., 2022), and Mrebel (Cabot et al., 2023).
- Metrics: SacreBLEU (Post, Matt. 2018).
- Experiment corpus: 549 data items, divided the corpus into training, validation, and test sets at a ratio of 6:2:2.
- Tokenizers: 409 new tokens were added for each to tokenizer.

[1]Model weights are from: https://huggingface.co/Babelscape/mrebel-large, https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt, https://huggingface.co/facebook/nllb-200-distilled-600M, and https://huggingface.co/facebook/m2m100 418M.

DiCoP-Text for the enrichment of language models and automatic translation

Table: The experimental results.

Model	Original Model	Fine-tuned Model	Improved Value
Mbart	2,0686	16,7191	14,6505
M2m100	4,9078	16,2120	11,3042
Nllb	4,6345	17,1385	12,5040
Mrebel	0,0131	7,1798	7,1667

- Introduction of PUs corpus can help translation models better translate PUs.
- Although each fine-tuned model outperformed its original model, they all scored low:
  - Volume of the training data.
  - Shortcomings of the original model.
- Improvement methods:
  - Expanding corpus size.
  - Improving the Chinese-French translation ability of models with the Chinese-French corpus of ordinary texts.

Conclusion

The first evaluation of the NLP tools in the DiCoP-Text project provided a detailed overview of the effectiveness of the DiCoP-Text corpus and the improved LMs. Our proposal aimed to improve LMs by integrating more fixed expressions and refining linguistic models for more accurate identification and translation of PUs. However, room for improvement exists:
1) In the future, we envision broader applicability of our DiCoP project. Indeed, expansion to other languages would strengthen its relevance and applicability across the IT language community. We also aim to provide more information, including details concerning user interfaces, accessibility, and integrating user feedback into ongoing development.
2) We studied the effect of fine-tuning the translation model using a PU corpus. This approach involved updating the tokenizer, training the model to integrate these PUs, refining the model from sentences containing them, and testing based on sentence-level data. The results indicated that fine-tuning the model with PUs could improve its translation capacity. However, given the limitations of the model and the corpus volume, additional efforts are necessary to refine its translation capability. Thus, our future research will focus on expanding the corpus and improving the model's Chinese-French translation capability.