



HAL
open science

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

Lian Chen, Wenjun Sun, Flora Badin

► **To cite this version:**

Lian Chen, Wenjun Sun, Flora Badin. DiCoP project and DiCoP-Text corpus for the enrichment of Language Models and Automatic Translation. XXI EURALEX International Congress, Oct 2024, Cavtat, Croatia. . hal-04725787

HAL Id: hal-04725787

<https://hal.science/hal-04725787v1>

Submitted on 8 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

The screenshot shows the DiCoP website with a search bar at the top. Below it, there are several sections: 'Digital phraseology in macrostructure', '3.1 Macrostructure of DiCoP', and '3.2 Computerized microstructure of DiCoP'. The website also features a 'Photo' section and a 'Video' section.

5. Corpus of phraseology: 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.

The screenshot shows a flowchart titled 'Corpus under development + NLP'. It details the process from identifying the corpus to NLP processing. Key steps include: 'Corpus identification', 'NLP processing', and 'Integration into the DiCoP-Text database'. The flowchart also mentions 'spaCy', 'thunlp/THULAC.NET', and 'LEXICO'.

The screenshot shows the NLP interface for identifying PUs. It includes a code editor with Python code for tokenization and PU identification. Below the code, there is a table showing the results for 'Liu Cixin's Three Body Problem'.

Total		Chinese PUs identified by Thulac	
Success rates	771	65.43% (504/771)	94.87% (687/771)
Error rate			
Form_1 (567 PUs)	Form_1 (204 PUs)	Correct	Error
321	246	163	21
Percentage	56.1%	49.5%	89.71%

The table displays 549 sentences in Chinese and their corresponding translations in English. The sentences are organized into columns, with Chinese text on the left and English translations on the right. The table is titled 'Table: Sentences containing the 549 PUs, and their translation'.



The figure shows the pipeline of model fine-tuning. It starts with 'Tokenizer' and 'Model'. The process involves 'Training' and 'Evaluation'. The final output is a 'Translation Task'.

Basic language models: Mbart (Tang et al., 2020), M2m100 (Fan et al., 2021), Nllb (Costajuss-Ja Marta et al., 2022), and Mirebel (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.

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 capacity. Thus, our future research will focus on expanding the corpus and improving the model's Chinese-French translation capability.