Innovation in phraseomatics : 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. Innovation in phraseomatics : DiCoP project and DiCoP-Text corpus for the enrichment of Language Models and Automatic Translation. EURALEX 2024 - 21st EURALEX International Congress Lexicography and Semantics, Kristina Štrkalj Despot; Ana Ostroški Anić; Ivana Brač, Oct 2024, Cavtat, Croatia. pp.227-234. hal-04664240

HAL Id: hal-04664240
https://hal.science/hal-04664240v1

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.

Kristina Š. Despot
Ana Ostroški Anić
Ivana Brač (Eds.)

Lexicography and Semantics

Proceedings of the
XXI EURALEX International Congress

8–12 October 2024
Cavtat, Croatia

Nakladnik: Institut za hrvatski jezik
Za nakladnika: Željko Jozić

Knjiga: Lexicography and Semantics,
Proceedings of the
XXI EURALEX International Congress

Urednice: Kristina Štrkalj Despot
Ana Ostroški Anić
Ivana Brač

Tehničko uređenje i računalni slog: Elena Vrbanić
Oblikovanje naslovnice: Elena Vrbanić
Tisak:

© This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License, 2024.

ISBN 978-953-7967-77-2

Kristina Š. Despot, Ana Ostroški Anić, Ivana Brač (Eds.)

Lexicography and Semantics

Proceedings of the

XXI EURALEX International Congress

8–12 October 2024
Cavtat, Croatia

TABLE OF CONTENTS

EURALEX XXI
CROATIA 2024

PART I: CONFERENCE OVERVIEW

Acknowledgements

We are deeply grateful to all those who generously supported the XXI EURALEX International Congress financially. Your contributions have been invaluable, and we sincerely appreciate your commitment to making this event a success.

Main Sponsors

Financira Europska unija
NextGenerationEU

Hornby
A. S. Hornby Educational Trust

LEXICAL COMPUTING

JANAF

Sponsors


ZAGREB
KRAŠ
1911


Jamnica
Since 1858


Republic
of Croatia
Ministry
of Culture
and Media
Republika
Hrvatska
Ministarstvo
kulture
i medija


REPUBLIKA HRVATSKA
Ministarstvo znanosti i
obrazovanja


HR
CLARIN

We extend our heartfelt thanks to everyone who dedicated their time and expertise to reviewing the submissions and papers. Your invaluable contributions have greatly enriched the quality of our conference, and we are truly appreciative of your efforts.

We wish to express our deepest gratitude to the members of the Local Organizing Committee for their remarkable dedication and hard work in making this event a reality. We truly appreciate all the time and energy you've invested in creating a memorable and smooth experience for everyone.

Programme Committee

Annette Klosa-Kückelhaus
Iztok Kosem
Robert Lew
Philipp Stöckle
Ivana Brač
Milica Mihaljević
Ana Ostroški Anić
Kristina Š. Despot

Local Organising Committee

Kristina Š. Despot
Ana Ostroški Anić
Ivana Brač
Željko Jozić
Ivana Filipović Petrović
Ivana Lovrić Jović
Maja Matijević
Martina Pavić
Siniša Runjaić
Lobel Filipić

Scientific Committee

Andrea Abel
Arleta Adamska-Sałaciak
Špela Arhar Holdt
Petra Bago
Martina Bajčić
Vincent Balnat
Vuk-Tadija Barbarić
Verginica Barbu Mititelu
Slobodan Beliga
Vladimír Benko
Goranka Blagus Bartolec
Tomislava Bošnjak Botica
Petar Božović
Ivana Brač
Maja Bratanić
Úna Bhreathnach
Thomas Burch
Valeria Caruso
Paul Cook
Rute Costa
Gilles-Maurice de Schryver
Janet Decesaris
Idalete Maria Silva Dias
María José Domínguez Vázquez
Anna Dziemianko
Pamela Faber
Ilse Feinauer
Ivana Filipović Petrović
Edward Finegan
Carolina Flinz
Thierry Fontenelle
Judit Freixa
Polona Gantar
Radovan Garabík
Zoe Gavriilidou
Alexander Geyken
Voula Giouli
Oddrun Grønvik
Volker Harm
Kris Heylen
Lana Hudeček
John Humbley
Max Ionov
Dubravka Ivšić Majić
Miloš Jakubíček
Jelena Kallas
Virna Karlić

Boris Kern
Ilan Kernerman
Fahad Khan
Annette Klosa-Kückelhaus
Svetla Koeva
Veronika Kolářová
Kristina Koppel
Iztok Kosem
Simon Krek
Margit Langemets
Barbara Lewandowska-Tomaszczyk
Robert Lew
Marie-Claude L'Homme
Anja Lobenstein-Reichmann
Irina Lobzhanidze
Henrik Lorentzen
Magdalena Majdak
Tinatin Margalitadze
Saša Marijanović
Ivana Matas Ivanković
John P. McCrae
Peter Meyer
Milica Mihaljević
Fabio Mollica
Carolin Müller-Spitzer
Hilary Nesi
Henrik Nilsson
Pär Nilsson
Vincent Ooi
Christian-Emil Ore
Petya Osenova
Julia Ostanina-Olszewska
Ana Ostroški Anić
Gokhan Ozkan
Benedikt Perak
Sanja Perić Gavrančić
Ralf Plate
Boris Pritchard
Ida Raffaelli
Natascia Ralli
Geraint Paul Rees
Irene Renau
Ana Salgado
Thomas Schmidt
Hindrik Sijens
Ranka Stanković
Egon W. Stemle
Frieda Steurs
Philipp Stöckle
Kristina Štrkalj Despot
Giovanni Tallarico
Elsabe Taljard
Pius ten Hacken
Carole Tiberius
Lars Trap-Jensen
Anna Vacalopoulou
Carlos Valcárcel Riveiro
Tony Veale
Federica Vezzani
Craig Volker
Sabine Wahl
Ana Werkmann Horvat
Geoffrey Williams
Sascha Wolfer
Alexander Ziem
Tanara Zingano Kuhn
Mojca Žagar Karer
Slavko Žitnik

Lian Chen, Wenjun Sun, and Flora Badin

INNOVATION IN PHRASEOMATICS

DiCoP Project and DiCoP-Text Corpus for the Enrichment of Language Models and Automatic Translation

Abstract This article examines advances in phraseomatics (Chen, 2023) and digital phraseography through the DiCoP project and its DiCoP-Text corpus, aimed at enriching linguistic models and machine translation. The project evaluates the frequency of use of phraseological units (PUs) and improves their translation in different contexts, drawing on recent research in phraseotranslation (Sułkowska, 2022) and natural language processing (NLP). It emphasizes French-Chinese and Chinese-French language pairs. We integrated 549 PUs from the novel *The Three-Body Problem* by Liu Cixin for our tests. Various processes, such as tokenization, identification, alignment, and annotation, were used to improve the translation of PUs. DiCoP-Text, a comprehensive database including newspaper articles, literary works, and textbooks, aims to enhance the performance of language models (LMs).

Keywords DiCoP; DiCoP-Text; phraseomatics; phraseotranslation; digital phraseography; NLP; language models

1. Introduction: Digital (Meta)Phraseography and Phraseomatics

In the contemporary digital age, the field of computerized phraseography faces challenges, which highlights the emergence of a particular discipline: phraseomatics or computational phraseology (Chen, 2023). This article explores the intricacies between computer science and the analysis of phraseological units or PUs (Gonzalez-Rey, 2002; Bolly, 2011, p. 28) to understand the subtleties of semantics and the structure of expressions in a digital context. This emerging discipline highlights the challenges and opportunities inherent in the evolution of (meta)phraseography (Chen, 2022; 2023).

In this sense, we present the Dictionary and Corpus of Phraseology (DiCoP)[1] project we are currently developing. The main objective is to create an electronic dictionary of multilingual phraseology (currently French-Chinese and Chinese-French) concerning PUs. Thus, this article is concerned with Chinese into French and, more specifically, among the innovative aspects of this project, the DiCoP-Text, which will serve as a corpus for phraseotraduction.

Indeed, in French and Chinese, PUs are ubiquitous in daily use (Bolly, 2011), so their recognition poses a major challenge in natural language processing (NLP). The DiCoP project addresses this problem by creating a bilingual electronic dictionary

[1] To learn more about the architecture of the macrostructure and microstructure of this innovative DiCoP digital dictionary, we invite you to consult our work presented at the Asialex conference (Chen, 2023).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

(with multilingual expansion envisaged) and developing a corpus of common PUs (e.g., collocations, idiomatic expressions, and proverbs). However, identifying PUs in a monolingual source text and translating them correctly for a parallel (bilingual) corpus constitutes a real challenge for automatic phraseotranslation (Sułkowska, 2022; Chen, 2022), especially since these expressions are numerous and possess metaphorical and opaque semantics (Gross, 1996).

2. Corpus of Phraseotraductology: DiCoP-Text

It is possible to determine corpus types depending on the characteristics of a corpus and the origin of its data. Bowker and Pearson (2002) distinguished the following corpus types: general/specialized, synchronic/diachronic, written/oral, and monolingual/bilingual. Monolingual corpora are comprised of data from just one language, while multilingual corpora are comprised of data from several languages and are either parallel or comparable.

DiCoP-Text refers to a collection of texts used to study PUs. This corpus contains a comparative component, in each language separately, sourced from various media (e.g., literary works, poetry, magazines, and newspapers) to analyze the usage and check the vitality (e.g., type and frequency) of PUs in a specific language. It also includes a parallel component (i.e., bilingual, primarily sourced from literary works, as other media lack existing translations) to enhance bilingual translations.

The collection and digital processing of texts are currently undergoing expansion. To illustrate the application of NLP in the DiCoP-Text project and analyze PUs, more precisely, *chéngyǔ*[2] corresponding to French idiomatic expressions, we selected a parallel corpus: the novel *The Three-Body Problem (三体 Sāntǐ)* by Liu Cixin, with its French translation. The novel comprises 186,079 words in the original Chinese version and 141,279 in the French version.[3]

3. DiCoP-Text and NLP: Analysis of Parallel Corpus Results

To develop the corpus and NLP, we started by processing various electronic or digitized file formats, converting them to conventional formats such as .txt. The corpus is designed to ensure its quality and relevance. We then analyzed the data to extract morphosyntactic and syntactic information, and automatically selected PUs from the corpora. We will now present our method and the steps of processing the corpus in a systematic and rigorous manner.

[2] "*Chéngyǔ* are polylexical sequences, fixed phrases, or short sentences that function as monolexical units within a sentence. Semantically, they are endowed with a specific, non-compositional meaning that cannot be directly deduced from the individual characters. Syntactically, their basic form, which most often follows a fixed quaternary (quadrisyllabic) rhythm and is divided phonetically and/or syntactically into two hemistichs, is conventional and unchanged for generations; hence the name chéngyǔ, meaning "ready-made expressions." Culturally, they convey the idiosyncrasies of a culture. Most often derived from classical literary language, they reflect an elegant and concise style and frequently contain strong allusive content" (Chen, 2021, p. 129). For example: *佛口蛇心 fókǒu-shéxīn* (Buddha's mouth, serpent's heart): "the mouth of a Buddha, the heart of a snake" (Prov.).

[3] For example, in French, "avoir la tête dans les nuages" (have your head in the clouds).

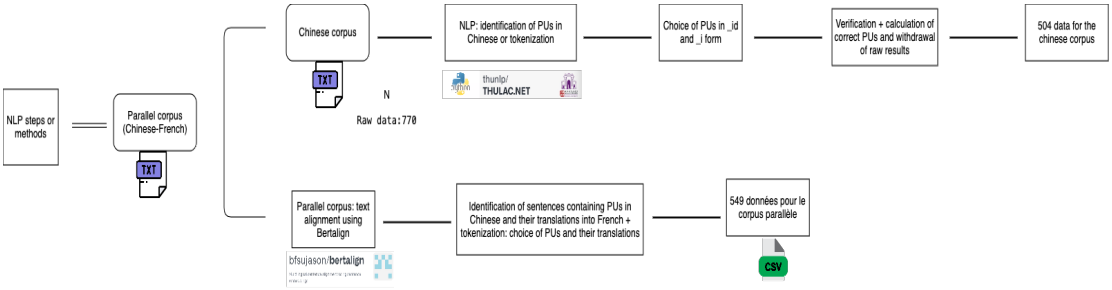

Fig. 1: NLP process followed for the parallel corpus

3.1 Performance and Evaluation of PU Identification in the Chinese Version with Thulac

For the Chinese analysis, we used the Thulac[4] tool with Python to identify *chéngyǔ*, which we marked as _i or _id to indicate they were idioms. Our results are described below.

Table 1: Chinese monolingual corpus performance and evaluation results

	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%

We identified 771 instances marked as _id or _i. Of these, 504 were confirmed correct, with 267 errors not constituting PUs. Even with a success rate of approximately 65.43%, this tool greatly facilitated identification and tokenization at this stage. We also checked 66 idiomatic expressions (IEs) that the Thulac tool did not identify as PUs.

Furthermore, we observed repetitions in the IEs listed in this table. For example, the expression 一动不动 *yīdòngbùdòng* appeared seven times but was translated differently depending on the case, such as "ne bougeait plus" (no longer moved), "être immobiles" (to be still), and "resta immobile" (remained still), or it could be omitted in the translation depending on the specific context.

3.2 Data Alignment in Chinese and French

Data alignment in Chinese and French was conducted using the Bertalign[5] tool, which is known for its effectiveness in multilingual alignment. Bertalign

[4] http://thulac.thunlp.org/

[5] https://github.com/bfsujason/bertalign

achieves more accurate results than the traditional length-, dictionary-, or MT-based alignment methods (Thompson & Koehn, 2019). Bertalign identifies correspondences between words in both languages and facilitates the subsequent extraction of Chinese PUs and their French translations. Hence, the tool helps avoid literal translations of compositional expressions, which is a key challenge in machine translation.

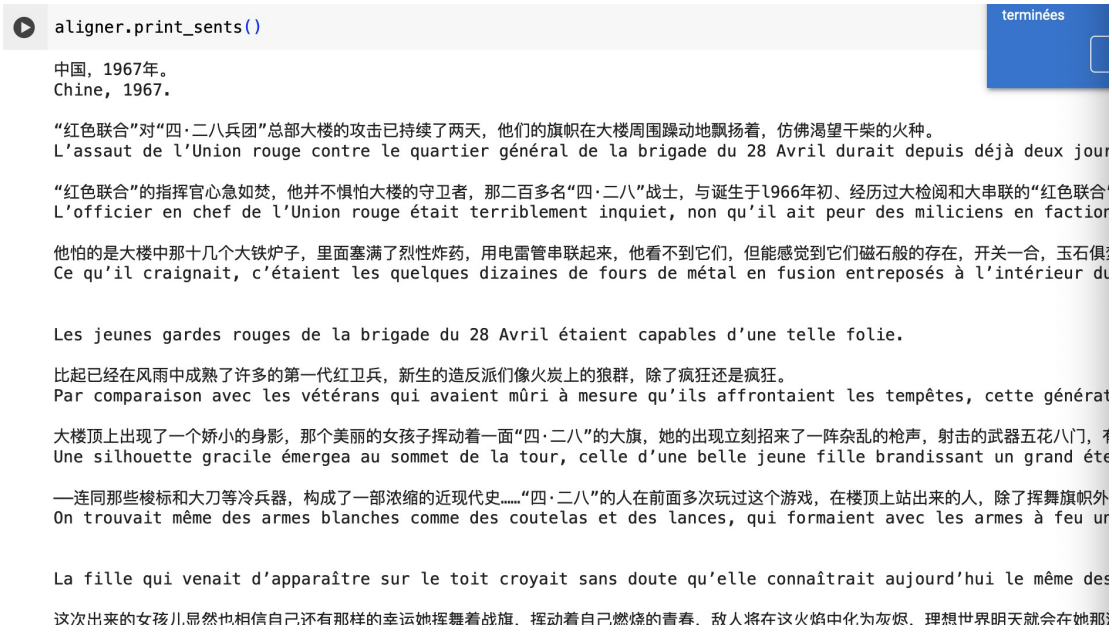

Fig. 2: Each sentence in Chinese and its French translation after alignment

After alignment, the results were evaluated for accuracy by analyzing selected sentences to guarantee the quality and reliability of the aligned data. Sentences containing these idioms and their translations were compiled in a CSV file for detailed analysis. By focusing on sentences containing PUs, we provided essential enriched context to enhance the efficiency of automatic translation. This targeted approach, recommended by Artetxe et al. (2018), eliminated the need for large parallel corpora. The PUs identified were associated with equivalents in the target languages and were classified by a confidence value.

These data were then passed on to our engineers for technical integration and optimization, ensuring efficient linguistic data management in a multilingual context.

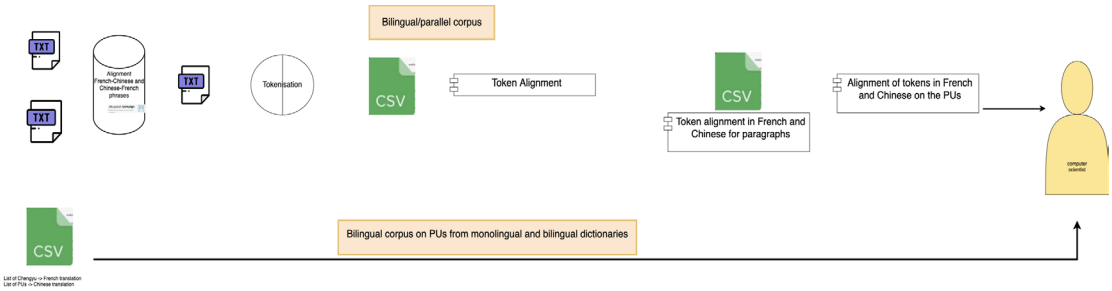

Fig. 3: Processes from NLP to IT

4. DiCoP-Text for the Enrichment of Language Models and Automatic Translation: Phraséo – Tratuctologie et Informatique

Today, neural network technology is widely used in machine translation task, especially pre-trained models methods. Pre-trained models are trained with large-scale datasets and then fine-tuned for specific subtasks to achieve better performance. However, because of the dependency on pre-trained models for data and the lack of specific PU training data, the current pre-trained translation models faced challenges when the sentences contained PUs.

Thus, we proposed an approach to improve the translation of machine translation models' PUs.[6] First, we updated the vocabulary of the tokenizer of the pre-trained model based on PUs and then renewed the model's embedding layer so that the model could recognize PUs from the input context and produce the new token embedding. Then, we fine-tuned the model with individual PU data to improve its ability to embed the PUs into a semantic vector.

We proceeded to sentence-level training after the model could distinguish and embed individual PUs. We split the sentence data into training, validation, and test sets. With this idea, we tested if fine-tuning with the PU corpus could improve the machine translation model's ability to recognize and translate PUs at the sentence level.

Experiment, Results, and Analysis

We first extracted Chinese PUs from all datasets and compared each PU with the tokenizer of the language model. We added it if there was no corresponding item in the tokenizer's vocabulary while expanding the language model's embedding layer to initialize the weights of the added Chinese PUs. In the training phase, the model learned the embeddings of all Chinese PUs and the training sentence set. Then the model was tested on the test sentence set.

We selected these language models:[7] Mbart (Tang et al., 2020), M2m100 (Fan et al., 2021), Nllb (Costa-jussà et al., 2022), and Mrebel (Cabot et al., 2023). For metrics, we chose SacreBLEU (Post, 2018). The experiment corpus had 549 data items. We divided the corpus into training, validation, and test sets at a ratio of 6:2:2. Subsequently, 409 new tokens were added for each of the four language models to the tokenizer. In the training phase, we set learning to 4e-5 and the batch size to 8. For the translation task, we selected Chinese to French.

The results are shown in Table 2. The "Original model" and the "Fine-tuned model" designate the performance of the original and fine-tuned language models, respectively.

[6] The relevant code is available at the anonymous code repository: https://anonymous.4open.science/r/DiCo-C7BA

[7] The weights files of all the language models can be downloaded at this address separately: 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.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Table 2: 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

Based on the experimental results, we noted that after fine-tuning, all language improved in translation performance, indicating that the introduction of PUs can help translation models better understand PUs. Although each fine-tuned model outperformed its original model, they all scored low. This phenomenon was related to the volume of the training data, as the pre-trained models needed a large corpus to enable them to learn as much as possible about the semantics of individual tokens under different contents. However, the dataset used in this experiment had only 549 items, which was insufficient. Therefore, the fine-tuning performance of the model will be further improved after the PUs corpus is expanded. Nonetheless, fine-tuning could not completely make up for the shortcomings of the original model. In addition to expanding the PU corpus, using the Chinese-French translation corpus of ordinary texts to improve the Chinese-French translation ability of the model can also improve the performance. We will use methods such as prompt engineering to explore the effects of corpus improvement in future work.

4. 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.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

References

Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In I. Gurevych, & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers)* (pp. 789–798). Association for Computational Linguistics.

Bolly C. (2011). *Phraséologie et collocations. Approche sur corpus en français L1 et L2*. Peter Lang.

Bowker L., & Pearson, J. (2002). *Working with Specialized Language – A Practical Guide to Using Corpora.* Routledge.

Cabot, P.-L. H. et al. (2003). REDFM: A Filtered and Multilingual Relation Extraction Dataset. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4326–4343). Association for Computational Linguistics.

Chen L. (2023). Meta)phraseography and phraseomatics: DiCoP, a computerized resource of phraseological units. *Proceeding of ASIALEX 2023: Lexicography, Artificial Intelligence, and Dictionary Users – The 16th International Conference of the Asian Association for Lexicography* (pp. 224–231). Asian Association for Lexicography.

Chen L. (2022). Phraséoculturologie: une sous-discipline moderne indispensable de la phraséologie. *SHS Web of Conferences – 8e Congrès Mondial de Linguistique Française – CMLF 138* (04011), 1–18. https://doi.org/10.1051/shsconf/202213804011

Chen L. (2021). *Analyse comparative des expressions idiomatiques en chinois et en français (relatives au corps humain et aux animaux)*. Doctoral dissertation, University of Cergy Paris]. HAL.

Costa-jussà M.-R. et al. (2022). *No language left behind: Scaling human-centered machine translation*. Retrieved July 30, 2024, from https://arxiv.org/pdf/2207.04672

Fan, A. et al. (2021). Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, *22* (107), 1–48.

Gross, G. (1996). *Les expressions figées en français: noms composés et autres locutions*. Ophrys.

González-Rey M. I. (2002). *La phraséologie du français*. Presses Universitaires du Mirail.

Jha, A., & Patil, H. Y. (2023). A review of machine transliteration, translation, evaluation metrics and datasets in Indian Languages. *Multimedia Tools and Applications*, *82*(15), 23509–23540.

Lu, Y., Zeng, J., Zhang, J., Wu, S., & Li, M. (2021). Attention calibration for transformer in neural machine translation. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1288–1298). Association for Computational Linguistics.

Mejri S. (2013). Figement et défigement: problématique théorique. In L. Perrin (Eds.), *Pratiques: Le figement en débat*, n° *159–160*, 79–97.

Mel'čuk I., & Polguère A. (2007). *Lexique actif du français*. De Boeck.

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

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores[C]. In O. Bojar et al (Eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics.

Sułkowska M. (2022). Phraseotranslation: Problems, Methods, Concepts. *Romanica Cracoviensia*, *1*, 29–41. doi: 10.4467/20843917RC.22.003.15635

Tang, Y. et al. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Thompson B., & Koehn P. (2019). Vecalign: Improved Sentence Alignment in Linear Time and Space. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp.1342–1348). Association for Computational Linguistics.

Acknowledgements

We extend our sincere thanks to Mrs. Anne-Lyse Minard, a specialist in Natural Language Processing (NLP) and Associate Professor at the Laboratoire Ligérien de Linguistique (LLL) at the University of Orléans, for her invaluable assistance in guiding us through computing techniques.

Contact information

Lian Chen
LLL, Universiét d'Orléans
LT2D, Cergy Paris Université
lian.chen@univ-orleans.fr

Wenjun Sun
L3i, Université de La Rochelle
wenjun.sun@univ-lr.fr

Flora Badin
CNRS-LLL
flora.badin@univ-orleans.fr