Towards End-to-End Multilingual Metaphor Processing: Integrating Detection, Translation, and Evaluation

Jiahui Liang[1], **Lifeng Han**[2,3]
[1]Centre for Linguistics, Humanities, Leiden University, NL
[2]LIACS, Leiden University, NL
[3]BDS, Leiden University Medical Centre, NL
j.h.l.jiahui@hum.leidenuniv.nl | l.han@lumc.nl

Abstract

Metaphorical language remains a major challenge for multilingual natural language processing because successful interpretation and translation require reasoning beyond literal lexical meaning. Existing research has largely investigated metaphor detection, machine translation, and translation evaluation as separate tasks, while little work has explored how these components can be integrated into a unified computational framework. This PhD proposal aims to develop an end-to-end framework for multilingual metaphor processing consisting of three complementary research directions: (1) robust metaphor detection across languages, (2) metaphor-oriented translation evaluation for both human assessment and automatic quality estimation, and (3) joint modelling that connects metaphor detection with translation evaluation. The proposed research will combine linguistic theory with recent advances in large language models to develop new datasets, annotation methodologies, evaluation benchmarks, and automatic evaluation approaches for metaphor-aware machine translation. The expected outcome is a unified framework that improves both the development and evaluation of multilingual NLP systems when processing figurative language.

1 Introduction

Metaphors are widespread in our daily usage and play a central role in human cognition and communication. Rather than being merely a rhetorical device used for stylistic effect, metaphor is regarded as a fundamental mechanism through which people conceptualise and understand the world (Lakoff and Johnson, 1980; Kövecses, 2010b). Specifically, the process is realized by conceptualising the abstract target domain through mappings from the concrete source domain (Lakoff and Johnson, 1980). For example, in the typical conceptual metaphor TIME IS MONEY, time is understood as a valuable resource.

These conceptual mappings are grounded in embodied physical and cultural experiences (Lakoff and Johnson, 1980; Gibbs, 2008), while their interpretation often depends on contextual information (Li and Chen, 2025a) and shared cultural knowledge (Lakoff and Johnson, 1980). Because embodied experiences and conceptualizations may differ across individuals and cultures (Kövecses, 2010b), metaphorical meaning is often non-compositional and cannot be derived from the literal meanings of individual words alone. Instead, it requires the integration of linguistic, conceptual, contextual, and cultural information. These challenges have motivated increasing research on computational metaphor processing, including metaphor detection, interpretation, generation and translation (Rai and Chakraverty, 2020; Ge et al., 2023). Among these tasks, metaphor **detection** and metaphor **translation** have developed largely as *independent* research directions (Liang et al., 2025; Liang and Han, 2026). Metaphor detection aims to distinguish metaphorical from literal language and serves as a prerequisite for many downstream applications (Shutova, 2010), whereas metaphor translation focuses on conveying metaphorical meaning across languages while accounting for linguistic and cultural differences (Karakanta et al., 2025). Despite substantial advances in neural language models, both tasks remain challenging. Successful metaphor detection requires models to recognize non-literal meaning beyond surface lexical patterns, while metaphor translation must further address variations in linguistic form, communicative function, and cultural embeddedness across languages (Karakanta et al., 2025). However, relatively *little work has connected these two areas*. Existing translation evaluation methods generally assume that metaphor instances have already been identified manually, while advances in metaphor detection have rarely been incorporated into translation evaluation or quality estimation.

Recent evidence also suggests that large language models may rely more heavily on superficial lexical cues than on genuine metaphor understanding, indicating that figurative language remains difficult even for state-of-the-art systems (Sanchez-Bayona and Agerri, 2025). Furthermore, progress in metaphor translation is constrained by the limited availability of high-quality parallel metaphor corpora and systematic evaluation resources (Wang et al., 2024; Han et al., 2026b).

Recent work has also highlighted shortcomings in current evaluation practice. General-purpose MT metrics, such as BLEU (Banerjee and Lavie, 2005) or COMET (Rei et al., 2020), are not designed to capture metaphor-specific phenomena, and human evaluation protocols rarely distinguish different types of metaphor translation errors. Emerging resources, including MMTE (Wang et al., 2024), have begun addressing this gap, while our MetaHOPE framework extends the human-oriented HOPE evaluation framework with metaphor-specific error categories and annotation guidelines (Liang and Han, 2026). Nevertheless, reliable metaphor-oriented evaluation and its integration with automatic metaphor processing remain open research problems.

This PhD proposes an end-to-end computational framework for multilingual metaphor processing that unifies three complementary research directions: (1) metaphor detection, (2) metaphor-oriented translation evaluation integrating MetaHOPE, and (3) joint modelling of metaphor detection and translation evaluation. Rather than treating these components as isolated tasks, the proposed research investigates how they can mutually reinforce one another to improve both the evaluation and the development of MT and LLM systems. The long-term objective is to establish reliable methodologies, resources and benchmarks for metaphor-aware multilingual NLP.

2 Background and Related Work

2.1 Cognitive Metaphor

Metaphor has traditionally been regarded as a rhetorical device used to enrich literary expression. However, this view was fundamentally challenged by Conceptual Metaphor Theory (CMT), which argues that metaphor is not merely a property of language but a fundamental mechanism of human thought and cognition (Lakoff and Johnson, 1980). According to CMT, abstract concepts are systematically understood through more concrete source domains, giving rise to conceptual mappings such as *ARGUMENT IS WAR* and *TIME IS MONEY*. Linguistic metaphorical expressions are therefore surface manifestations of deeper conceptual structures rather than isolated stylistic devices.

Subsequent research has further demonstrated that metaphor is a multidimensional phenomenon involving cognition, language, communication, and culture. While conceptual mappings provide the cognitive basis for metaphorical expressions, their linguistic realisations are influenced by discourse context and communicative intention (Semino, 2008). Deliberate Metaphor Theory (DMT), for example, distinguishes between metaphors that are intentionally employed to direct readers' attention towards the source domain and those that have become conventionalised in everyday language (Steen, 2017). This distinction is particularly relevant for computational processing because not all metaphorical expressions require the same level of semantic interpretation.

Metaphor is especially prevalent in news discourse, political communication and other public-facing genres, where it serves multiple communicative functions beyond stylistic ornamentation. Previous studies have shown that metaphor can simplify complex events, frame public opinion, evoke emotional responses, and communicate ideological positions (Steen et al., 2010b). Consequently, successful metaphor processing requires models to capture not only lexical meaning but also discourse-level and pragmatic information.

Cross-linguistic variation further increases the complexity of metaphor processing. Although some conceptual metaphors appear to be widely shared across cultures, their linguistic realisations often differ considerably because of language-specific conventions, cultural experiences, and socio-historical backgrounds (Kövecses, 2010b). As a result, multilingual NLP systems must account for both universal conceptual mappings and culture-specific metaphorical expressions (Han et al., 2026a). These challenges motivate the need for computational approaches capable of recognising, interpreting, translating and evaluating metaphorical language across languages.

2.2 Computational Metaphor Detection

Automatic metaphor detection has become one of the most active research topics in figurative language processing (Shutova, 2015; Rai and

Chakraverty, 2020; Han et al., 2026a). Early computational approaches mainly relied on hand-coded knowledge, lexical resources, and selectional preferences to distinguish metaphorical from literal language (Shutova, 2015). However, these rule-based methods were often difficult to scale and generalise across domains and languages, motivating the development of data-driven approaches.

With the development of corpus-based metaphor research, manual metaphor identification procedures also became increasingly standardised. The Metaphor Identification Procedure (MIP) (Group, 2007) and its extended version, MIPVU (Steen et al., 2010b), provide systematic dictionary-based guidelines for determining whether a lexical unit is used metaphorically by comparing its contextual meaning with a more basic meaning . Based on this procedure, the VU Amsterdam Metaphor Corpus (VUAMC) (Steen et al., 2010a) was constructed as one of the largest manually annotated metaphor corpora and has since become a widely adopted benchmark for metaphor detection research. The availability of MIPVU-annotated corpora has enabled the development and evaluation of supervised metaphor detection models under consistent annotation standards.

Advances in deep learning have substantially improved metaphor detection performance. Contextual language models, including BERT and subsequent Transformer architectures, enable richer semantic representations by modelling contextual information surrounding metaphorical expressions. These models consistently outperform earlier statistical approaches on standard benchmark datasets and demonstrate improved robustness across multiple domains. This transition was reflected in the VUA Metaphor Detection Shared Tasks. While the top-performing systems in 2018 were mainly based on recurrent neural networks and feature engineering, the 2020 shared task was dominated by BERT-based models, with the best F1 score improving from 0.651 to 0.769 (Leong et al., 2018, 2020).

More recently, researchers have begun to explore large language models (LLMs) for automatic metaphor detection. Instead of relying exclusively on task-specific supervised training, LLMs can identify metaphorical expressions through zero-shot and few-shot prompting. Recent studies have investigated the use of Llama, Qwen, Mistral, DeepSeek, Gemma, GPT-4, and other LLMs for metaphor identification, demonstrating their potential for detecting conventional metaphorical expressions while also revealing limitations in achieving robust metaphor understanding (Liang et al., 2025; Reimann and Scheffler, 2025; Han et al., 2026a). Building on these advances, recent work has explored task-specific prompt design, MIPVU-inspired reasoning, chain-of-thought prompting, and retrieval-augmented generation (RAG) to further improve metaphor detection (Reimann and Scheffler, 2025; Fuoli et al., 2026). Related studies have also extended these approaches to multilingual, cross-lingual, and multimodal metaphor detection (Lai et al., 2023; Hülsing and Im Walde, 2024; Xu et al., 2024). Nevertheless, current results remain mixed, with the effectiveness of LLM-based approaches depending strongly on the task, prompting strategy, and evaluation dataset (Reimann and Scheffler, 2025; Fuoli et al., 2026; Han et al., 2026a).

Despite these advances, several challenges remain. Current systems often struggle with conventional metaphors, culturally specific metaphorical expressions, and implicit figurative meanings that require discourse-level reasoning or common-sense knowledge. Moreover, most existing work evaluates metaphor detection as an isolated classification task, while relatively little attention has been paid to how detected metaphorical expressions can support downstream applications such as machine translation and translation quality evaluation.

Building upon our previous work on GPT-4-based metaphor detection in English news texts (Liang et al., 2025), this limitation motivates the *first work package* of this PhD, which extends metaphor identification towards *multilingual* settings, improved prompting strategies, and more robust detection methods that can support downstream translation and evaluation.

2.3 Metaphor Translation

Metaphor translation has long been recognised as one of the most challenging problems in translation studies because metaphorical expressions often convey meaning beyond their literal lexical content. Unlike literal translation, successful metaphor translation requires preserving not only semantic equivalence but also conceptual mappings, communicative functions, stylistic effects, and cultural connotations. Classical translation scholars have therefore argued that metaphor translation should be viewed as a process of conceptual transfer rather than simple lexical substitution. Newmark (1988), for example, proposed a taxonomy of metaphor translation procedures ranging from reproducing the same metaphor in the target language to replacing or paraphrasing metaphorical expressions according to communicative needs. Similarly, van den Broeck (1981) and Schäffner (2004) emphasised that translation strategies should consider both linguistic form and underlying conceptual structures, highlighting the importance of cultural adaptation in cross-linguistic metaphor transfer.

The rapid development of neural machine translation (NMT) has substantially improved translation quality for general-purpose texts (Johnson et al., 2017; Kocmi et al., 2025). Transformer-based systems and multilingual large language models (LLMs) are increasingly capable of generating fluent and contextually appropriate translations without explicit linguistic rules. Nevertheless, metaphorical language remains a persistent challenge because successful translation often requires pragmatic reasoning, cultural knowledge, and an understanding of implicit conceptual mappings (Han et al., 2026b). Existing MT systems may produce literal translations that distort metaphorical meaning, omit metaphorical imagery, or replace metaphors with more conventional expressions, thereby reducing their communicative impact (Liang and Han, 2026).

Recent research has begun to investigate metaphor translation from a computational perspective. Wang et al. (2024) introduced MMTE, a multilingual benchmark specifically designed for evaluating machine translation of metaphorical language across several language pairs. Their study demonstrated that although modern LLM-based systems outperform earlier neural MT models, substantial performance gaps remain, particularly for culturally specific and conceptually complex metaphors. Other recent studies have similarly compared human translators with NMT systems and LLMs, showing that although modern LLMs often produce fluent and semantically adequate translations, they remain less reliable in preserving metaphorical expressions, conceptual mappings, and rhetorical effects, particularly for culturally specific or novel metaphors (Wang et al., 2024; Karakanta et al., 2025; Li and Chen, 2025b). These findings suggest that improvements in general machine translation quality do not necessarily translate into reliable metaphor translation.

Although the quality of MT and LLM systems continues to improve rapidly, relatively little work has investigated how metaphor translation errors should be analysed systematically. Existing studies typically report overall translation quality or metaphor preservation rates, providing limited diagnostic information about the specific linguistic and conceptual problems underlying translation failures. Consequently, metaphor translation research increasingly requires evaluation frameworks capable of identifying fine-grained error types beyond conventional MT quality metrics. This observation motivates *the second work package* of this PhD, which investigates metaphor-oriented translation evaluation and analysis.

2.4 Translation Evaluation

Automatic evaluation has become an indispensable component of modern machine translation research. Traditional corpus-level metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), and LEPOR (Han et al., 2012, 2013) primarily measure lexical or n-gram overlap between system outputs and reference translations. Although these metrics remain useful for benchmarking general translation quality, they are often insensitive to semantic equivalence and perform poorly when evaluating figurative language, where multiple valid translations may differ substantially in lexical realisation.

Recent evaluation methods have increasingly shifted towards semantic and learned metrics. COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020), for example, employ pretrained language models to estimate translation quality based on semantic similarity rather than surface-form matching. These approaches generally correlate more strongly with human judgement than traditional automatic metrics and have become standard evaluation tools in contemporary MT research. Nevertheless, they remain general-purpose metrics and do not explicitly assess whether metaphorical meaning, conceptual mappings, or rhetorical effects have been preserved during translation.

To address the limitations of automatic metrics, human-centred evaluation frameworks have received growing attention. The Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2024, 2014; Gladkoff et al., 2025) provides a hierarchical taxonomy for classifying translation errors according to dimensions such as accuracy, fluency, terminology, and style. HOPE (Gladkoff and Han, 2022) further simplified human evaluation by introducing a penalty-based annotation scheme that improves annotation efficiency and inter-annotator consistency while retaining diagnostic capability. However, neither MQM nor HOPE explicitly considers metaphor as an independent evaluation dimension.

Recent metaphor-specific evaluation research has therefore begun to bridge this gap. MMTE (Wang et al., 2024) proposed benchmark datasets and dedicated evaluation metrics for metaphor translation, demonstrating that conventional automatic metrics frequently overestimate translation quality for figurative language. However, MMTE corpus has isolated sentences without context and it only contains verbal metaphors. Building upon these developments, our previous work (Liang and Han, 2026) introduced MetaHOPE, a metaphor-oriented extension of the HOPE framework that incorporates *five fine-grained error categories* to capture different aspects of metaphor translation quality, including communicative impact (IMP), adaptation to target language (RAM), semantic accuracy (MIS), stylistic preservation (STL), and naturalness (PRF)· MetaHOPE used context-aware translation before extracting the metaphor-containing segments for evaluation. It also produces a open-source parallel corpus during the post-editing phase that can be useful for future metaphor studies and benchmarks. Preliminary pilot studies further showed that refined annotation guidelines substantially improve inter-annotator agreement, supporting the feasibility of reliable human evaluation for metaphor translation.

Despite these advances, metaphor evaluation remains largely dependent on manual annotation, making large-scale evaluation both expensive and time-consuming. Furthermore, current automatic metrics rarely incorporate information obtained during metaphor detection, and human evaluation frameworks generally assume that metaphorical expressions have already been identified. These limitations motivate the *final work package* of this PhD, which investigates automatic metaphor translation quality estimation by integrating metaphor detection, human evaluation, and LLM-based evaluation into a unified computational framework.

2.5 Research Questions from the Gap

Research Questions

- RQ1: How can metaphorical expressions be reliably detected across languages?
- RQ2: How should metaphor translation quality be evaluated?
- RQ3: Can metaphor detection and translation evaluation be integrated into a unified computational framework?

3 Proposed Research Framework

We design the thesis framework in Figure 1, which includes three components 1) metaphor detection, 2) metaphor translation evaluation, and 3) joint modelling on detection and translation evaluation.

Step I detection task: we plan to investigate rule-based, simple LLM-prompting with examples and chain-of-thoughts (CoTs), and RAG.

Step-II translation and evaluation task: we divide it into reference-free evaluation and reference-based. The reference-free evaluation includes 1) automatic quality estimation (QE) using source text and system output, 2) human native-speaker assessing the translation quality, and 3) post-editing based metrics. The reference-based evaluation needs corpus with multiple-references for better mapping between hypothesis translation and the references. For this, we can use the created references from post-editing based evaluation from the reference-free step.

Step-III joint task: we investigate the possibility of joint modelling on both metaphor detection task and metaphor translation evaluation, which can be reference-based or reference-free quality estimation.

4 WPs

WP1 includes Detection, datasets, prompting, fine-tuning, and multilingual models.

WP2 includes: metaphor translation, MT/LLM, and Translation Benchmark.

WP3 is on Human Evaluation, Integrating Meta-HOPE (Liang and Han, 2026), annotation, corpus, guidelines, and reliability discussion.

WP4 is of Automatic Eval, auto MQE, llm-as-a-judge, and end-to-end Joint framework (Detection, Translation, Evaluation /Automatic QE, LLM judge).

5 Methodology

In our earlier work on LLM-based metaphor detection for English news texts (Liang et al., 2025), we discovered that GPT-4 achieved substantially lower

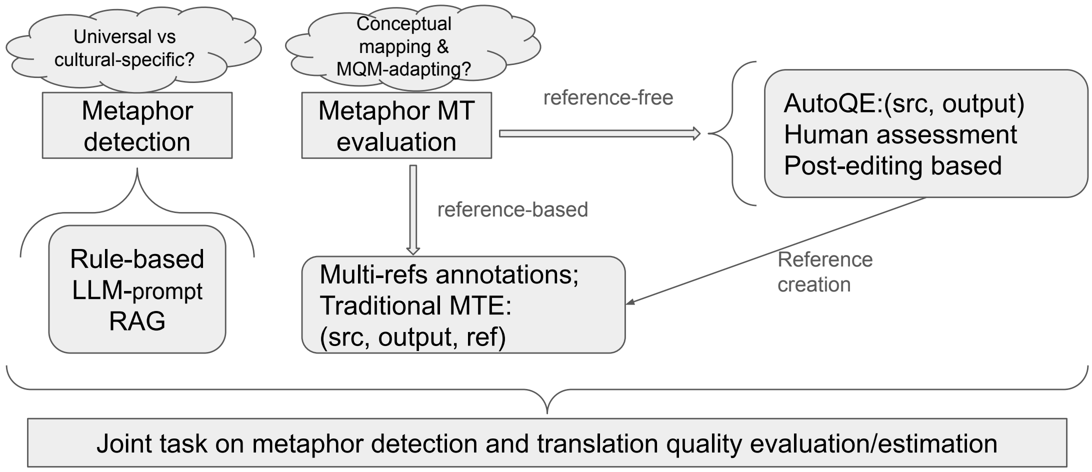

Figure 1: Thesis Proposal Framework on Metaphor Detection, Translation Evaluation, and Joint Modelling.

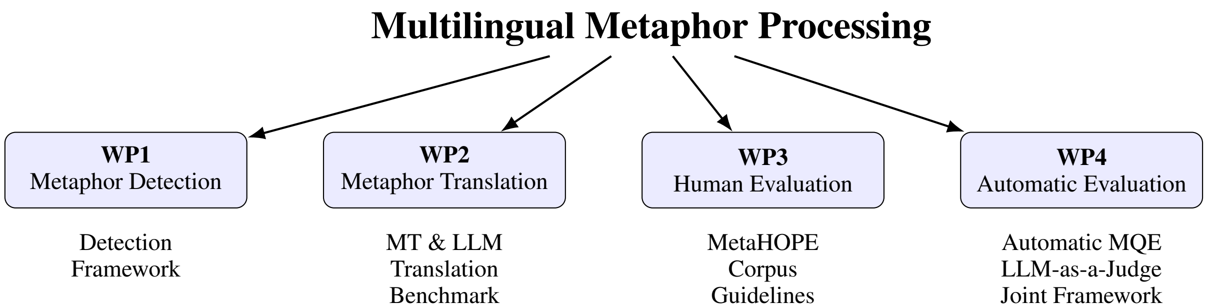

Figure 2: Overview of the proposed PhD research programme. The four work packages form an end-to-end framework for multilingual metaphor processing, spanning metaphor detection, translation, human evaluation, and automatic evaluation.

performance than SOTA supervised metaphor detection models on conventional metaphor detection. We also observed that model performance was highly sensitive to prompt formulation. While one-shot prompting consistently improved performance over zero-shot prompting, providing additional examples did not lead to further improvements. Our error analysis further showed that prepositional metaphors (e.g., under in under strain) were particularly difficult to detect, with many prompts failing to identify such cases in the zero- to five-shot settings. Although detection improved in the ten-shot setting, our findings indicate that the observed gains cannot be straightforwardly interpreted as improved metaphor understanding and require further investigation. These findings motivate further research on prompt optimization, alternative output formats, fine-tuning open-source LLMs, and RAG. Building upon our earlier work, the first work package (**WP1**) extends metaphor identification towards *multilingual* and more *robust* detection methods. In particular, we will explore newer and more diverse open-source LLMs such as GPT -5, DeepSeek, Huanyuan, and Gemma 4 from 2025 onwards. Furthermore, we will investigate chain of thoughts (CoTs) and RAG methods on this task, naming it *MetaDetect*.

Regarding Metaphor Translation and Evaluation (WP2 and WP3), we are conducting the roll-out of larger size testing set annotation based on the MetaHOPE framework (Liang and Han, 2026) using 200+ segments with their context using 5 annotators for both translation directions on English-Chinese. The resulting outputs from this will be a statistically robust analysis of the current state-of-the-art LLMs regarding metaphor-related word translations. In addition, it will create a bilingual corpus to be shared publicly for research on English-Chinese pairs with multiple references (from multiple annotations). We believe the multi-reference parallel corpus will also serve as an valuable resource to the metaphor study field.

For WP4 on automatic evaluation, we are attending the WMT2026 shared task on Test Suites[1], where we used the AlphaMWE (Han et al., 2026b, 2020) multilingual parallel corpus with MWE annotations including metaphors and idioms. We have received 30+ MT system submissions on the translation of this corpus that are being analysed using both automatic metrics and human evaluations. The automatic metrics being deployed include BLEU, CharF++, LEPOR, and BERTscore, covering both lexical-based and word-embedding metrics.

Our next step on the thesis moving forward is about end-to-end metaphor detection and translation evaluation modelling, i.e., part of WP4. For the end-to-end models, we will investigate "LLMs as Agents" (Plaat et al., 2025), applying different LLMs for detection and translation evaluation from two joint phases in a collective manner. LLM-as-a-judge will also be explored as the evaluation phase with human-in-the-loop, as suggested from the literature (Wit et al., 2026).

Overall, the four WPs are coherently connected: outputs of WP1 feed into WP2; WP2 provides resources for WP3; and WP3 supervises WP4.

6 Research Plans

The Thesis Research Plan is listed in Figure 3.

7 Expected Contributions

Expected outputs include:

- Resources
  - multilingual corpus
  - benchmark
  - annotation guidelines
- Methods
  - MetaHOPE
  - automatic evaluator
  - joint framework
- Scientific contributions
  - end-to-end metaphor processing paradigm

8 Conclusion

This PhD proposal presents a research agenda towards end-to-end multilingual metaphor processing by integrating metaphor detection, metaphor translation, human evaluation, and automatic translation evaluation into a unified computational framework. While these research areas have each advanced considerably in recent years, they have largely been investigated independently. As a result, current multilingual NLP systems still face substantial challenges when processing metaphorical language, particularly across languages and cultures.

The proposed research addresses this gap through four complementary work packages. The first investigates robust multilingual metaphor detection using both linguistic knowledge and recent large language models. The second examines metaphor translation and develops benchmark resources for analysing metaphor translation strategies across machine translation and LLM systems. The third develops a reliable human evaluation methodology for metaphor translation, including annotation guidelines, corpora, and empirical studies of annotation reliability. Finally, the fourth explores automatic metaphor translation evaluation by integrating metaphor detection, human evaluation, and LLM-based quality estimation into a unified framework.

Beyond developing individual methods and datasets, this research aims to establish stronger connections between metaphor detection and translation evaluation. Information produced during metaphor detection has the potential to support more accurate translation quality estimation, while human evaluation can provide supervision for developing metaphor-aware automatic evaluation models. By investigating these interactions, the proposed research seeks to move beyond isolated task-specific solutions towards a coherent framework for multilingual metaphor processing.

The expected outcomes include new multilingual resources, reproducible annotation methodologies, benchmark datasets, and automatic evaluation techniques that facilitate future research on figurative language processing. More broadly, the proposed work contributes to the development of multilingual NLP systems that better capture the linguistic, conceptual, and cultural dimensions of metaphor, thereby supporting more reliable machine translation and evaluation of figurative language.

Current Progress

- Completed: GPT-4 metaphor detection study; MetaHOPE framework.

[1] https://www2.statmt.org/wmt26/testsuite-subtask.html

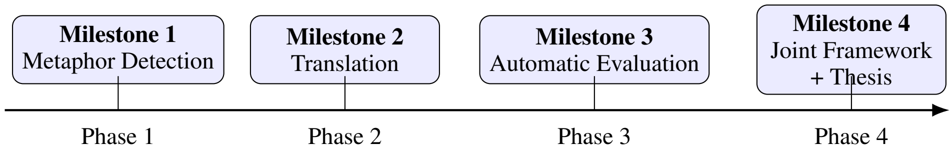

Figure 3: Proposed research timeline and milestones for the PhD project.

- In progress: Large-scale MetaHOPE annotation; WMT 2026 Test Suites evaluation.
- Planned: Joint metaphor detection and evaluation models; LLM-as-a-judge framework.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Matteo Fuoli, Weihang Huang, Jeannette Littlemore, Sarah Turner, and Ellen Wilding. 2026. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning. *Applied Corpus Linguistics*, page 100204.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(Suppl 2):1829–1895.

Raymond W. Gibbs. 2008. Metaphor and thought: The state of the art. In Jr. Gibbs, Raymond W., editor, *The Cambridge Handbook of Metaphor and Thought*, Cambridge Handbooks in Psychology, pages 3–14. Cambridge University Press, Cambridge.

Serge Gladkoff and Lifeng Han. 2022. Hope: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective mt evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21.

Serge Gladkoff, Lifeng Han, and Katerina Gasova. 2025. Non-linear scoring model for translation quality evaluation. *arXiv preprint arXiv:2511.13467*.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. 2012. LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING 2012: Posters*, pages 441–450, Mumbai, India. The COLING 2012 Organizing Committee.

Aaron L.-F. Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013. Language-independent model for machine translation evaluation with reinforced factors. In *Machine Translation Summit XIV*, pages 215–222. International Association for Machine Translation.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, David Lindevelt, Sander Puts, Erik van Mulligen, and Suzan Verberne. 2026a. Dutch metaphor extraction from cancer patients' interviews and forum data using llms and human in the loop. *CL4Health WS at LREC2026, Palma, Spain*.

Lifeng Han, Najet Hadj Mohamed, Malak Rassem, Gareth JF Jones, Alan F Smeaton, and Goran Nenadic. 2026b. Towards a resource for multilingual lexicons: an mt assisted and human-in-the-loop multilingual parallel corpus with multi-word expression annotation. *Language Resources and Evaluation*, 60(2):33.

Anna Hülsing and Sabine Schulte Im Walde. 2024. Cross-lingual metaphor detection for low-resource languages. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 22–34.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Alina Karakanta, Mayra Nas, and Aletta G Dorst. 2025. Metaphors in literary machine translation: Close but no cigar? In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 276–286.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, et al. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In

Proceedings of the Tenth Conference on Machine Translation, pages 355–413.

Zoltán Kövecses. 2010b. Metaphor and culture. *Acta Universitatis Sapientiae, Philologica*, 2(2):197–220.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*, volume 1. University of Chicago press Chicago.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the second workshop on figurative language processing*, pages 18–29.

Chee Wee Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 vua metaphor detection shared task. In *Proceedings of the workshop on figurative language processing*, pages 56–66.

Zhengjian Li and Lang Chen. 2025a. Mind vs. machine: Comparative analysis of metaphor-related word translation by human and ai systems. *Training, Language and Culture*, 9(1):10–27.

Zhengjian Li and Lang Chen. 2025b. Mind vs. machine: Comparative analysis of metaphorrelated word translation by human and ai systems. *Training, Language and Culture*, 9(1):10–27.

Jiahui Liang, Aletta G Dorst, Jelena Prokic, and Stephan Raaijmakers. 2025. Using gpt-4 for conventional metaphor detection in english news texts. *Computational Linguistics in the Netherlands Journal*, 14:307–341.

Jiahui Liang and Lifeng Han. 2026. MetaHOPE: A metaphor-oriented evaluation framework for analysing mt and llm translation errors. In *Proceedings of the 9th International Conference on Natural Language and Speech Processing (ICNLSP 2026)*. Accepted for publication. Preprint available at arXiv:2607.00848.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall, London.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Aske Plaat, Max Van Duijn, Niki Van Stein, Mike Preuss, Peter Van der Putten, and Kees Joost Batenburg. 2025. Agentic large language models, a survey. *Journal of Artificial Intelligence Research*, 84.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sebastian Reimann and Tatjana Scheffler. 2025. Using large language models to perform mipvu-inspired automatic metaphor detection. In *The 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, page 10.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and large language models: When surface features matter more than deep understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17462–17477.

Christina Schäffner. 2004. Metaphor and translation: some implications of a cognitive approach. *Journal of pragmatics*, 36(7):1253–1269.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Elena Semino. 2008. *Metaphor in discourse*. Cambridge University Press Cambridge.

Ekaterina Shutova. 2010. Models of metaphor in nlp. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 688–697.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Gerard Steen. 2017. Deliberate metaphor theory: Basic assumptions, main tenets, urgent issues. *Intercultural Pragmatics*, 14(1):1–24.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010a. Vu amsterdam metaphor corpus.

Gerard J Steen, Aletta G Dorst, Tina Krennmayr, Anna A Kaal, and J Berenike Herrmann. 2010b. A method for linguistic metaphor identification.

Raymond van den Broeck. 1981. The limits of translatability exemplified by metaphor translation. *Poetics Today*, 2(4):73–87.

Shun Wang, Ge Zhang, Han Wu, Tyler Loakman, Wenhao Huang, and Chenghua Lin. 2024. Mmte: Corpus and metrics for evaluating machine translation quality of metaphorical language. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 11343–11358.

Tamara Wit, Lifeng Han, Carly Heipon, David Lindevelt, Anne Stiggelbout, and Suzan Verberne. 2026. Measuring the practice of shared-decision making (option12): An investigation into open-sourced smaller llms (os-sllms) for better privacy and sustainability. *Preprint*, arXiv:2607.06127.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101.

A Initial Annotation Guidelines

Two files: "Annotation Guideline for The Categorization of Potentially Universal Metaphors and Culture" and "Annotation Guidelines for Metaphor Translation Strategies and Quality Evaluation" both available at `https://github.com/Jiahui84/MetaHOPE`