https://doi.org/10.31820/f.37.2.4

Kristina Štrkalj Despot, Ana Ostroški Anić, Polona Gantar, Mija Bon, Matej Klemen, Marko Robnik Šikonja, Simon Krek, Benedikt Perak, Jaka Čibej

CroSloMet: A STRUCTURED METAPHOR DATASET FOR CROATIAN AND SLOVENE

dr. sc. Kristina Štrkalj Despot, Institute of Croatian Language, Zagreb
kdespot@ihjj.hr orcid.org/0000-0001-9004-5103
dr. sc. Ana Ostroški Anić, Institute of Croatian Language, Zagreb
aostrosk@ihjj.hr orcid.org/0000-0001-9999-0750
dr. sc. Polona Gantar, University of Ljubljana
apolonija.gantar@guest.arnes.si orcid.org/0000-0001-5822-6414
Mija Bon, University of Ljubljana
mija.bon@ff.uni-lj.si orcid.org/0000-0002-3377-1640
Matej Klemen, University of Ljubljana
matej.klemen@fri.uni-lj.si orcid.org/0000-0002-7852-2357
dr. sc. Marko Robnik Šikonja, University of Ljubljana
marko.robniksikonja@fri.uni-lj.si orcid.org/0000-0002-1232-3320
dr. sc. Simon Krek, University of Ljubljana
simon.krek@ff.uni-lj.si orcid.org/0000-0001-8965-6863
dr. sc. Benedikt Perak, University of Rijeka
bperak@uniri.hr orcid.org/0000-0003-4177-5307
dr. sc. Jaka Čibej, University of Ljubljana
jaka.cibej@ff.uni-lj.si orcid.org/0000-0002-3037-6848

izvorni znanstveni rad
UDC 811.163.42’373.612.2
811.163.6’373.612.2
rukopis primljen: 17. rujna 2025; prihvaćen za tisak: 11. studenoga 2025.

Recent advancements in large language models (LLMs) have opened new avenues for processing figurative language, yet their performance in metaphor interpretation continues to fall short of human-level understanding. One limitation lies in the inadequacy of existing metaphor datasets, which often lack explicit connections to conceptual metaphors and are predominantly monolingual. In this paper, we present CroSloMet, a novel dataset of over 1,120 metaphorical and 1,120 literal sentences in Croatian and Slovene, grounded in the MetaNet.HR framework. Each example is annotated with the corresponding conceptual metaphor, linguistic multi-word expression (MWE), canonical forms, and literal usage, enabling both metaphor identification and explanation tasks. We present preliminary evaluations of the dataset through two experiments: metaphor classification using CroSloEngual BERT, achieving 88.5% accuracy, and metaphor explanation generation with LLama 3-8B, where strict exact-match evaluation yielded low scores despite semantically valid outputs. To address this, we propose a multi-level validation framework combining manual annotation, natural language inference, semantic similarity, and LLM-based judgment. Our findings highlight the importance of capturing generality and specificity in metaphor mappings and call for more nuanced evaluation methods. CroSloMet provides a resource for advancing metaphor understanding in LLMs and contributes to cross-linguistic and cognitively informed metaphor research.

Keywords: *metaphors; metaphor dataset; metaphor explanation; metaphor understanding; large language models*

1. Introduction

Computational approaches to metaphor processing have evolved significantly, yet challenges persist due to the variability of metaphorical expressions across languages and contexts, and different theoretical approaches and understandings of what metaphor is. Recent advancements in natural language processing (NLP) have demonstrated the potential of transformer-based models in figurative language processing. However, LLMs often struggle with contextual nuances and domain-specific metaphorical expressions. While LLMs show improvement over chance in metaphor identification, their performance lags behind human capabilities (Tong et al. 2024; Kim et al. 2023; He et al. 2023; Despot, Ostroški Anić, and Veale 2023; Liu et al. 2022).

Our study presents a dataset designed to improve metaphor identification and metaphor name generation, with initial experimental results indicating promising performance improvements over previous bench-

marks. The limitations of existing datasets have prompted the creation of this new figurative language dataset, which was built on the basis of the MetaNet.HR database (Despot et al. 2019). The CroSloMet dataset aims to enhance metaphor classification and recognition through improved linguistic structure and annotation detail. Our approach is based on the cognitive theory of metaphor, i.e., conceptual metaphor theory (Lakoff and Johnson 1999). Conceptual metaphor theory (CMT) underscores the central role of metaphors in everyday language use and positions them as a fundamental component of our conceptual system crucial for understanding abstract concepts. The MetaNet.HR project (Despot et al. 2019) has sought to systematically document conceptual metaphorical mappings within the Croatian language, capturing source-target relations (metaphorical mappings) between different image schemas and semantic frames. By leveraging this structured resource, we have constructed a dataset that aims to provide a linguistic foundation for metaphor identification.

Existing datasets for metaphor research, most notably the VU Amsterdam Corpus (Steen et al. 2010, more on this in section 2), have provided valuable resources for computational experiments. However, these datasets exhibit certain limitations that we discuss in the next chapter. Our dataset builds upon previous research by introducing a structured bilingual dataset, explicitly linking conceptual metaphors (CMs) with linguistic realizations and including detailed annotations of metaphorical multi-word expressions (MWEs). The dataset structure facilitates machine learning applications, including fine-tuning LLMs for metaphor detection and interpretation.

2. Related Work and Theoretical Background

Our approach to metaphor is based on a conceptual metaphor theory (CMT) (Lakoff and Johnson 1980; 1999; Lakoff 1993), that sees metaphors as a fundamental component of our conceptual system. This perspective posits that metaphors are integral to the encoding, storage, representation, and retrieval of concepts, with the activation of metaphorical structures being an inherent part of conceptual thought. According to CMT, abstract concepts are predominantly understood through metaphorical mappings to more concrete experiences. As a result, our understanding of abstract concepts like *love*, *friendship*, and *morality* is predominantly shaped by metaphorical thinking. We often describe and interpret these ideas through

more concrete experiences, such as *warmth*, *closeness*, or *cleanliness*. Extensive experimental research supports the existence of a cognitive connection between metaphorical source and target domains, particularly in cases of primary metaphors (for an overview, see Dancygier and Sweetser 2014: 36–38). We define a metaphor as a frame-to-frame mapping, a (one-way) mapping that projects conceptual structure from the source domain (e.g., *warmth*) onto another domain, referred to as the target domain (e.g., *affection*), and consequently, there is a set of correspondences or mappings between the constituent elements of the source domain and those of the target domain (Despot 2024). Conceptual metaphors underline both conventional and novel, creative metaphorical expressions.

Metaphor processing has been a challenge in Natural Language Processing (NLP), with researchers exploring various computational approaches to metaphor identification, interpretation, and generation. Metaphors present unique difficulties for LLMs because they require a deeper, more abstract level of reasoning beyond literal word meanings that spans different cognitive domains, which necessitates several advanced reasoning capabilities. It requires analogical reasoning (Holyoak and Thagard 1995) and categorization (Lakoff and Johnson 1999) as well as contextual and world knowledge. One key issue is common sense reasoning. Many metaphors depend on implicit world knowledge that is not explicitly stated in the text. For instance, understanding the metaphor *"She has a heart of stone"* requires knowing that stone is hard and unyielding, and metaphorically extending that property to emotions. While modern LLMs are trained on vast amounts of textual data, they still struggle with implicit reasoning because they lack true experiential understanding of the world. Unlike humans, who develop intuition from sensory and social experiences, LLMs rely on probabilistic associations between words. Additionally, metaphors are often highly context-dependent, meaning the same phrase can carry different meanings in different situations. For example, *"He's a rock"* could imply strength and reliability in one scenario, but emotional unresponsiveness in another. Even those LLMs with sophisticated contextual embeddings (Devlin et al. 2019) struggle with subtle shifts in meaning because they rely on statistical correlations rather than genuine understanding. Cultural knowledge further complicates metaphor comprehension. Many metaphors are deeply rooted in cultural traditions, making them difficult for LLMs to interpret correctly without specific training data (Yang et al. 2025, Chen and Wang 2025).

Early work on computational metaphor processing relied on rule-based methods and linguistic theories (for an overview, see Shutova 2011). More recent advancements have leveraged deep learning models and large-scale annotated corpora to improve metaphor understanding in LLMs (for an overview, see Tong et al. 2024). Recent research has explored various strategies to enhance metaphor processing in LLMs. One key approach is expanding training datasets to include a richer set of metaphorical expressions. For example, the Metaphor Understanding Challenge Dataset (MUNCH) (Tong et al. 2024) has been introduced to provide a standardized benchmark for evaluating LLM metaphor comprehension. MUNCH is designed to evaluate the metaphor understanding capabilities of LLMs and provides over 10 000 paraphrases for sentences containing metaphor use, as well as 1500 instances containing inapt paraphrases. All apt and inapt paraphrases were manually annotated. Metaphorical sentences cover natural metaphor use across four genres (academic, news, fiction, and conversation), and they exhibit different levels of novelty. Experiments with LLaMA and GPT-3.5 demonstrated that MUNCH presents a challenging task for LLMs (Tong et al. 2024).

Similarly, analogy-based learning techniques (Turney and Littman 2005) have been explored to improve LLMs' ability to draw conceptual parallels between metaphorical source and target domains. By explicitly training models on structured analogy datasets, researchers aim to improve metaphor generalization. Another promising direction is contextual embedding techniques, as representing words and phrases as vectors that capture their contextual meaning can help LLMs disambiguate metaphorical language and understand its intended meaning (Devlin et al. 2019).

One of the most widely used resources for computational metaphor research is the VU Amsterdam Metaphor Corpus (Steen et al. 2010), which provides a manually annotated corpus of metaphorical and literal language instances based on the British National Corpus (BNC). The VU Metaphor Corpus has been instrumental in training and evaluating NLP models for metaphor detection. However, the corpus is based entirely on English texts, which restricts its applicability to models intended for cross-linguistic figurative language processing. Additionally, unlike MetaNet.HR, which explicitly links metaphors to conceptual domains (e.g., TIME IS MONEY, EMOTIONS ARE FORCES), the VU Corpus mainly annotates metaphorical words without systematically mapping them to underlying conceptual metaphors. Another common issue with this and other similar existing metaphor

datasets is the fragmented nature of annotations—multi-word metaphors are often labeled as separate single-word metaphors, which may hinder more accurate and holistic modeling of metaphorical expressions.

Several studies have attempted to improve upon these limitations by integrating additional metaphor resources. Liu et al. (2022) tested various LLMs on metaphor interpretation tasks and found that performance on the VU Corpus was lower compared to datasets with richer conceptual structure. Similarly, He et al. (2023) introduced a dataset for simile interpretation and found that metaphor detection models trained on VU data struggled with more nuanced figurative language tasks. Other corpora, such as the LCC metaphor datasets (Mohler et al. 2016), and metaphor-emotion dataset (Mohammad et al. 2016), provide metaphor annotations at the word or sentence level. While widely used for automated metaphor identification (see Tong et al. 2021), they lack information on metaphor interpretation. In contrast, some datasets focus on interpretation as a paraphrasing task (Shutova 2011; Bizzoni and Lappin 2018; Joseph et al. 2023), but they are often small (200–1000 instances) and do not explicitly model the reasoning process behind metaphor interpretation, which remains an open challenge (for an overview, see Tong et al. 2024). Klemen and Robnik Šikonja (2023) conducted experiments using multiple transformer-based large language models on four variants of two publicly available Slovene corpora: KOMET (Antloga 2020) and G-KOMET (Antloga and Donaj 2022). Their study included monolingual, multilingual, and cross-lingual settings, incorporating the VU Amsterdam Metaphor Corpus as an additional source of metaphorical knowledge. Model performance was evaluated quantitatively using the word-level F1 score. Other figurative language resources exist, such as MetaNet, MetaNet.HR, Cordoba Metonymy Database, etc. (for an overview see Bolognesi, Brdar, and Despot 2019), but were not extensively used for NLP purposes. For a good overview of other non-English resources and work on metaphor detection, see Klemen and Robnik Šikonja 2023.

Alongside metaphor identification, metaphor generation has long been a focus in artificial intelligence and natural language processing research. Early approaches involved using probabilistic word associations to generate basic "A is like B" metaphors, given a target concept and its attributes (Abe et al. 2006), or explored transforming literal expressions into metaphorical ones using word embeddings and sequence-to-sequence models (Stowe, Ribeiro, and Gurevych 2020). Thesaurus Rex (Veale and Hao 2007), a web-

based tool, takes two input words (e.g., *soccer* and *basketball*) and returns shared conceptual categories such as *sport* or *game* to support metaphor generation. Thesaurus Rex leverages web data to generate a diverse range of alternative perspectives on familiar concepts. It selects from this pool and applies vertical reasoning through WordNet to produce precise similarity judgments. Additionally, Rex identifies the most informative perspective to clearly explain each comparison or, in generative mode, to suggest creative analogies. For example, to convey the potential harmfulness of coffee, it may propose comparisons to alcohol, tobacco, or pesticide—substances commonly classified as toxic substances on the web. A web service called MetaphorMagnet (Veale 2019) uses web data as a knowledge resource for metaphor to generate and understand deliberate metaphors on demand allowing other applications to exhibit a measure of their own figurative creativity. Similarly, Kim et al. (2023) introduced Metaphorian, an LLM-powered metaphor generation tool, a system designed to assist science writers in creating scientific metaphors by supporting the processes of searching, expanding, and iteratively refining metaphors. It employs a workflow based on large language models, guided by heuristic strategies identified in a study involving six professional writers. This approach demonstrated that LLMs can generate coherent metaphors when given structured prompts, but still struggle with metaphor interpretation tasks.

Our work builds upon these existing studies by introducing a structured bilingual dataset that addresses several critical gaps in previous research, such as the tendency of the datasets to focus exclusively on the English language and a lack of explicit links between linguistic expressions and the underlying conceptual metaphors they instantiate. In contrast, our dataset incorporates aligned Croatian and Slovene data, and each metaphorical expression in our dataset is explicitly mapped to a conceptual metaphor, drawing from the MetaNet.HR framework.

3. CroSloMet Dataset Construction

The CroSloMet dataset presented in this study was developed to enhance computational metaphor processing by incorporating detailed linguistic, but also conceptual information. It is based on the MetaNet.HR repository, a structured database of Croatian conceptual and linguistic metaphors, systematically mapping source-target relations among semantic frames and image schemas (Despot et al. 2019). This repository covers

all main metaphor families, including domains such as Event Structure, Emotions, Mind, Morality, Time, Economics, Governance, and Well-Being. The method of populating the database involved using computational tools like Word Sketches and Thesaurus from Sketch Engine (Kilgarriff et al. 2014) for compiling the lists of target words frequently occurring in metaphorical contexts and manual metaphor annotation using the Metaphor Identification Procedure (MIP) (Pragglejaz 2007). Collected examples are systematically classified according to their metaphor type, level, source and target domains, and linguistic realization. For each of the target words all the Word Sketches and a random concordance sample (300 lines per target word) are analyzed. Then, the metaphorical expressions and collocations are annotated using the MIP procedure (Pragglejaz 2007). At least two annotators annotated the same concordance sample, and points of disagreement are discussed by the entire group. A mutual decision is reached in cases of disagreement (for other details on the method see Despot et al. 2019).

For the dataset compilation, we first extracted all conceptual metaphors and their examples from the original repository. For each CM, we then collected additional example sentences from the HrWaC corpus in order to obtain a minimum of three examples per CM. This procedure resulted in a new dataset containing more than 1,120 annotated sentences featuring metaphorical multi-word expressions (MMWEs). Two annotators independently searched the HrWaC corpus for additional metaphorical examples. All retrieved examples were subsequently reviewed by at least two additional authors. Any points of disagreement were discussed collectively by the entire research team. Following the reasoning of Basile et al. (2021), we did not measure inter-annotator agreement (IAA), as the collaborative resolution procedure and dataset type make formal IAA measures unnecessary. Using the same procedure, we then collected an equivalent number of literal instances of the same MWEs from the HrWaC corpus, resulting in a balanced set of metaphorical and literal examples.

The dataset has a broad coverage since the MetaNet.HR project included all the most important metaphor families as mentioned above. The dataset primarily focuses on conventional metaphors, as these are considerably more frequent in natural language use and therefore more readily retrievable from large corpora. However, it also contains a smaller number of creative or novel metaphorical expressions that emerged naturally in the source data. The dataset also includes both general (e.g., ACTION IS MOTION) and specific conceptual metaphors (LOVE RELATIONSHIP IS A BOAT TRIP), as this is how

the MetaNet.Hr (that served as the starting point) was structured. This distinction reflects the hierarchical nature of conceptual metaphors: more general metaphors (e.g., ACTION IS MOTION) often underpin a wide variety of more specific instantiations (e.g., COMPETITION IS A RACE, LOVE RELATIONSHIP IS A BOAT TRIP). However, at this stage of the project, we did not annotate examples for their degree of novelty or their degree of generality/specificity. Distinguishing between conventional and novel metaphorical uses requires clear operational criteria, yet existing definitions vary and often rely on graded, context-dependent judgments. Introducing novelty annotation or specificity annotation without a well-defined framework would risk inconsistency and reduce the reliability of the dataset. For this reason, we deferred systematic specificity and novelty annotation to future work, when more rigorous guidelines can be established and tested. This version of CroSloMet dataset is therefore structured to facilitate only metaphor detection (whether an MWE is metaphorical or not) and conceptual metaphor recognition (i.e., the generation of the name of the conceptual metaphor that a certain MWE reflects).[1]

The key components of the dataset include:

a) Conceptual metaphor names in English and in Croatian.

b) Annotated Croatian example: Sentences containing metaphorical MWEs are given and metaphorical MWE is marked with asterisks. E.g., *Sve će napraviti kako bi *došao do cilja*. 'He will do anything to *reach his destination/goal*.'* The annotated expression is the linguistic realization of the conceptual metaphor listed under a). At least three examples per conceptual metaphor are provided. Every sentence contains only one annotated metaphorical expression, that is the linguistic example of the conceptual metaphor listed in the first column. We have not taken into consideration possible other metaphorical expressions in a sentence that are not the reflection of the metaphor name given in the first column.

c) Annotated Slovene example: *Vse bo naredil, da bi *prišel do cilja**.

d) The structure of the MWE: MWEs are categorized as either S (if the source domain is the only one explicitly lexically expressed) or ST (if both source and target domains are lexically expressed). E.g., *On je *dno dna*. 'He is at the *bottom*'* (S), where only the target word bottom is

[1] The dataset can be accessed here: https://zenodo.org/records/17751591.

expressed, and we cannot be certain what the target is without the context, and *On je *moralno dno dna*. He is at the *moral bottom** (ST), where both the source word *bottom,* and the target word *moral* are explicitly expressed.

e) The canonical form of Croatian metaphorical MWE: Standardized versions of metaphorical expressions are provided to facilitate automated corpus searches and further metaphor analysis (e.g., *doći do cilja* 'to reach a goal'). Column E lists the metaphorical MWEs in their canonical forms, i.e., as they appear in a dictionary. This does not apply to all examples, e.g. for MWEs in which the form in the plural is predominantly used (*trošiti novčane resurse* 'to spend financial resources'), or in examples of personification (*društvo pati*, 'society suffers'; *vatra proždire,* 'fire eats'), see some examples in Table 1.

f) The canonical form of Slovene metaphorical MWE: e.g., *priti do cilja* 'to reach a goal'

g) Annotators' comments that were not taken into consideration during the experiments

h) A literal example in Croatian from the web corpus hrWaC 2.1 (Ljubešić & Klubička, 2016): Examples of sentences in which the metaphorical expressions are used in their literal or primary meaning. This means that the base unit of the MWE (usually the word expressing the source of the metaphor) is used in another context in its literal meaning, enabling models to distinguish between metaphorical and literal contexts.

i) Literal MWE in Croatian: Due to their morphological richness, Croatian and other Slavic languages have developed an interesting mechanism of metaphorical specialization: e.g., we have the verb *motriti* 'to look, to see', and the verb *razmotriti*, which became specialized for the metaphorical meaning of 'considering, reflecting upon', reflecting the conceptual metaphor KNOWING IS SEEING. In these cases, it is impossible to find a literal example of *razmotriti* that would still mean 'to see, to look.' Thus, in the literal expression, we resorted to its basic form *motriti*. We didn't want to delete these examples as they are frequent and easy to detect – each appearance of the word *razmotriti* reflects the metaphor KNOWING IS SEEING.

j) A literal example in Slovene

k) A literal equivalent for MWE in Slovene.

Table 1. *Examples of metaphorical MWEs (MMWEs) from the dataset.*

Cro MMWE	Translation to English
ući u proces	enter the process
izaći iz procesa	exit the process
doći do cilja	reach the goal
slijediti korake	follow the steps
doći do dogovora	reach an agreement
doći do rješenja	come to a solution
put do prosvjetljenja	path to enlightenment
put do finala	path to the final
stići do izjednačenja	reach a tie
simptom siromaštva	symptom of poverty
simptom problema	symptom of the problem
graditi od nule	build from scratch
graditi odnos	build a relationship
graditi demokraciju	build democracy
birokratska hobotnica	bureaucratic octopus
kraci birokratske hobotnice	tentacles of the bureaucratic octopus
birokratska hobotnica	bureaucratic octopus
labirint birokracije	labyrinth of bureaucracy
labirint zakona	labyrinth of laws
neman birokracije	monster of bureaucracy
birokracija proždire	bureaucracy devours
žrvanj birokracije	grindstone of bureaucracy
korupcijska hobotnica	corruption octopus
glava hobotnice	the head of the octopus
krak hobotnice	tentacle of the octopus
boriti se s osjećajem	fight with a feeling
napad emocija	attack of emotions
borba s emocijama	struggle with emotions
boriti se s emocijama	fight with emotions
pobijediti emocije	overcome emotions
izljev emocija	outpouring of emotions
kanalizirati emocije	channel emotions
bujica emocija	flood of emotions
uzburkane emocije	stirred emotions
emocije preplave	emotions overwhelm

Cro MMWE	Translation to English
melodije prolaze	melodies pass
melodija se kreće	melody moves
muzika dolazi	music comes
kanal za priljev gotovine	cash inflow channel
tijek gotovine	cash flow
priljev gotovine	cash inflow
imati ideju	have an idea
ukrasti ideju	steal an idea
moja ideja	my idea
vlastita ideja	own idea
svježa ideja	fresh idea
slatka ideja	sweet idea
graditi karijeru	build a career
temelj karijere	the foundation of a career

To evaluate the cross-linguistic applicability of the CroSloMet dataset, a translation initiative was carried out to create a Slovene counterpart to the original Croatian entries. We automatically translated the Croatian sentences using ChatGPT4, and manually reviewed the translations for the Slovene part of the dataset. We performed the same procedure when selecting sentences with literal use of the metaphorical expression. If, in the Slovene part of the dataset, the translation of the Croatian sentence was not appropriate or the wording of the metaphorical or literal expression was incorrect, we replaced the Slovene translation with a selected sentence from the Gigafida 2.0 Corpus of Written Standard Slovene (Krek et al. 2020). While many conceptual metaphors are shared between the two languages, structural and semantic differences have required careful adaptation. In cases where multiple Slovene equivalents existed for a Croatian metaphorical expression, we selected the variant that was more frequent and widely used, provided it remained semantically consistent with the original metaphorical frame. For example, within the conceptual metaphor CENSORSHIP IS THE PHYSICAL RESTRAINING OF MOUTH, the Croatian expression *staviti flaster na usta* 'to put a band-aid over (someone's) mouth' was matched with Slovene equivalents such as *zapreti usta* 'to shut (someone's) mouth', *zalepiti usta* 'to tape/seal (someone's) mouth', or *zavezati jezik* 'to tie (someone's) tongue'. In other instances, variations occurred at the syntactic level, e.g., the Croatian *pritisnut dugom* 'burdened

with debt' corresponds to the Slovene *dolgovi pritiskajo* 'debts are weighing down'. Morphological differences were also observed, such as the Croatian *žrvanj birokracije* 'the millstone of bureaucracy', which is typically rendered in Slovene in the plural form as *birokratski mlini* 'the millstones of bureaucracy'. These adaptations ensure that the Slovene data remains faithful to the metaphorical intent while reflecting natural language use.

4. Preliminary Validation

To evaluate the dataset's effectiveness, we conducted a sanity-check experiment using the CroSloEngual BERT model (~100M parameters). The model was trained to classify sentences as either containing a metaphorical or a literal expression. The results showed a high accuracy of 88.5%, indicating the dataset in general contains clear distinctions between metaphorical and literal meanings (compared to what the annotators had determined), making it a valuable resource for future computational metaphor research. While the results are not directly comparable due to different experimental settings, the high accuracy shows great potential in contrast to existing work done on the Slovene KOMET corpus (Klemen and Robnik Šikonja 2023), where the maximum achieved F1 score on the word-level metaphor detection task was 60.7%.

The second experiment aimed to assess the ability of an LLM to generate explanations (conceptual metaphor name in "target frame is source frame" form) for metaphorical expressions. While LLMs have shown promise in metaphor detection, generating accurate and human-like explanations remains a significant challenge. This experiment was conducted using a fine-tuned LLama 3-8B model, evaluated on a balanced validation set of 220 examples, with an equal number of metaphorical and literal sentences. The model was fine-tuned to generate conceptual metaphor names for metaphorical expressions, i.e., the system is expected not only to recognize the expression containing a metaphor, but it is supposed to recognize which conceptual metaphor it relates to (a feature not tested in previous work with LLMs and figurative language). We took the column "English metaphor name" as the expected explanation, and for the literal expressions we took "Literal" as the expected explanation (as there should not be explanations for a non-metaphorical text). The model was trained using a training subset of the dataset, leveraging the structured format of conceptual metaphors and their annotated linguistic realizations. The model demonstrated

strong performance at not generating explanations for the literal examples: 70.43% recall and 77.66% precision. This high recall confirms that the dataset effectively separates metaphorical and literal instances, allowing the model to make reliable distinctions.

However, quantifying the accuracy of metaphor explanations (conceptual metaphor name generation) proved challenging due to the inherent subjectivity in metaphor interpretation. Using a strict exact-match metric – where only verbatim matches with the human-assigned metaphor label were considered correct – the model achieved only 5% accuracy. However, this metric is overly rigid, as many generated explanations were semantically valid but varied in specificity or wording. The 5% accuracy recorded under exact-match evaluation does not reflect the semantic validity of many explanations. The model often captured the correct conceptual metaphor but expressed it differently (e.g., *Poverty is a crime against humanity* instead of *Poverty is a crime*), see Table 2.

Table 2. *Examples of semantically valid system predictions, currently treated as inaccurate.*

Annotator's Label	The System's Prediction
Poverty is a crime	Poverty is a crime against humanity
Memorizing is putting objects into the container	Memory is storage
Analysis of social problems is a diagnosis of affliction	Poverty is a disease
Love is an object that is found	Love is an object that is searched for
Profit is a point on a scale	Profit is a vertical movement up a vertical axis
Opportunities are objects	Opportunities are grasped
The economic crisis is a disease	Economic problems are diseases
A word is a limited commodity	Words are resources
Improving economic status is an upward motion	Success is up
Debt is a physical impediment to motion	Money is a burden
Poverty is a plant	Poverty is a tree

While these outputs were not exact matches, they correctly captured the conceptual metaphor. This suggests that strict accuracy metrics underestimate the model's true capability. The model displayed tendencies that require further investigation: a) tendency to be more general or more specific (some explanations were more general or more specific than the human-labeled conceptual metaphor); b) structural mismatches (some generated explanations had correct conceptual mappings but used different wording or syntactic structures); c) some explanations were simply incorrect. Our next steps involve carefully manually analyzing and validating all these cases (see next chapter), where we will treat only cases under c) as incorrect.

One of the key theoretical challenges reflected in the construction and annotation of the CroSloMet dataset and in metaphor research and annotation in general is the issue of generality and specificity in metaphor analysis. As Dancygier and Sweetser (2014) emphasize, complex target domains – such as *debate* – are rarely structured by a single, simple metaphorical mapping. Instead, such domains often activate multiple overlapping conceptual metaphors at varying levels of abstraction. For instance, the metaphor ARGUMENT IS WAR includes specific projections such as IDEAS ARE WEAPONS, but these depend on more general ontological metaphors like IDEAS ARE OBJECTS, which reify abstract entities and allow us to speak of "giving," "losing," or "stealing" ideas. This layered structure raises important questions about how to best represent metaphorical mappings in a corpus. Should a metaphor be annotated as COMPETITION IS A RACE, or more generally as ACTION IS MOTION, given that the latter underpins the former? This issue becomes especially salient when metaphors involve nested or overlapping conceptual structures, such as COMMUNICATION IS OBJECT TRANSFER or COMMUNICATION IS A CHANNEL, which may be embedded in more specific war-related metaphors (e.g., in the expression "firing back" in a debate) (Despot 2024). In our database, this complexity is reflected in the careful selection of metaphorical labels, which must balance linguistic specificity with cognitive plausibility, but on the other hand, we did want to have very general metaphors represented in our database as well. For example, expressions like *udarac ispod pojasa* ("a low blow") clearly evoke the frame of physical combat rather than war in a geopolitical sense, whereas *pucati iz svih raspoloživih oružja* ("to fire all available weapons") aligns more naturally with the broader frame of warfare. Consequently, our annotation process involves constant negotiation between coarse-grained and fine-grained metaphor labels, acknowledging that metaphor systems

are hierarchical and interdependent. There is no single "correct" level of metaphor annotation; instead, context and discourse usage guide our decisions about which conceptual frame best captures the metaphorical meaning of an expression.

Future work could take advantage of the hierarchical organization of conceptual metaphors in the MetaNet.HR database (Despot et al. 2019) to refine the evaluation of model-generated metaphor explanations. Instead of relying solely on exact matches between predicted and gold-standard conceptual metaphors, we could implement a graded evaluation scheme that accounts for metaphor hierarchy. MetaNet.HR already encodes such inheritance relationships among metaphors, allowing us to calculate conceptual distance or similarity between predicted and reference metaphors even if its formulation is more abstract or differently phrased. Such an approach would help move beyond binary evaluation and better reflect the layered nature of metaphor understanding. Including both levels is not arbitrary, but a theoretically motivated choice grounded in Conceptual Metaphor Theory, which emphasizes that complex conceptual domains are structured by overlapping mappings at different levels of abstraction. Representing both general and specific metaphors in the dataset would ensure that models can be evaluated not only on their ability to detect the presence of figurative language, but also on the depth of conceptual understanding they achieve. For instance, correctly classifying only the general metaphor ACTION IS MOTION would demonstrate recognition of metaphorical structure but limited sensitivity to finer conceptual distinctions. Conversely, identifying a specific metaphor such as LOVE RELATIONSHIP IS A BOAT TRIP without connecting it to the broader ACTION IS MOTION frame would suggest reliance on surface lexical cues. By explicitly encoding generality and specificity, the dataset would support experiments that probe these different levels of abstraction and would avoid reducing metaphor processing to a binary task of literal versus figurative interpretation.

5. Conclusions and Next Steps

This study presents a novel CroSloMet dataset for figurative language understanding and evaluates the performance of Large Language Models in metaphor detection and explanation. By leveraging the structured annotation framework of the MetaNet.HR database, we have developed a bilingual dataset that enhances computational metaphor processing.

Our preliminary experiments have demonstrated that the CroSloEngual BERT model trained on the new CroSloMet dataset demonstrated 88.5% accuracy in distinguishing metaphorical from literal expressions. This result was significantly higher than previous tests on existing Slovene datasets (Klemen and Robnik Šikonja 2023), suggesting that our dataset provides clearer examples of metaphorical meaning distinctions. The structured nature of our dataset (explicit annotation of metaphor types and linguistic expressions) and its broad coverage played a crucial role in improving model performance.

However, in generating metaphor explanations (recognizing which conceptual metaphor is reflected in the linguistic expression), strict evaluation metrics underestimated model performance, as many generated outputs were semantically valid but expressed differently than the predefined human labels. The LLama 3-8B model struggled with exact-matching explanations, achieving only 5% strict accuracy, but qualitative analysis revealed that its outputs were often conceptually correct.

Given the difficulty of evaluating metaphor explanations (conceptual metaphor name generation), our next steps will involve validating these metaphor explanations both manually and automatically (utilizing techniques such as natural language inference, paraphrasing, semantic similarity, and additional LLMs). As part of our future work, we plan to develop a comprehensive framework for validating metaphor explanations, combining both manual and automated methods. Manually, evaluators will follow clear guidelines to assess whether a generated explanation is valid for the original metaphorical text, whether it captures the same conceptual mapping as the gold-standard explanation, and whether it is more specific or more general. On the automated side, we will implement a multi-pronged evaluation strategy. This includes using natural language inference (NLI) to test entailment relations between generated and true explanations, identifying cases where one explanation is a specification or generalization of the other. Paraphrase detection and semantic similarity measures (e.g., cosine similarity between text embeddings) will complement the NLI-based approach. Furthermore, we aim to incorporate large language models (LLMs) to assess explanation sensibility in context, by asking whether the generated explanation is valid given the original metaphor and a reference explanation. In parallel, we are exploring the training of a custom classifier for this task, which would require a labeled dataset containing valid and invalid explanations. This multifaceted evaluation pipeline will allow us to capture the nuances of metaphor inter-

pretation and move beyond overly rigid, exact-match metrics toward a more semantically grounded assessment. In future work, we will also explore ways to incorporate the hierarchical structure of conceptual metaphors in MetaNet.HR into our evaluation framework, enabling more nuanced assessments that go beyond exact matches by accounting for conceptual distance and metaphor generality.

These initial experiments with CroSloMet highlight both the potential and limitations of LLMs in explaining metaphors. While metaphor identification has improved significantly with structured datasets like ours, conceptual metaphor recognition (conceptual metaphor name generation) remains a challenge due to variability in conceptual interpretation and the lack of standardized evaluation metrics. While current models struggle with strict evaluation criteria, they demonstrate a strong ability to recognize figurative language and produce semantically relevant explanations. Future work will focus on refining evaluation methods and improving model interpretability by focusing on incorporating paraphrase detection, natural language inference (NLI), and human evaluation to improve assessment methods. Additionally, the inclusion of Slovene data has revealed cross-linguistic variations in metaphor realization, so additional languages may be incorporated in the future to study metaphor universality and variation.

Overall, this research contributes to the field of NLP by providing a structured, high-quality dataset for figurative language processing and identifying key areas where LLMs require further refinement. Our dataset emphasizes the MWE level (unlike the common word-level approach typical of VU Amsterdam Corpus) and conceptual grounding, which likely contributed to the model's higher accuracy. By addressing the current limitations in connecting correct conceptual metaphors with metaphorical MWEs, we move closer to developing systems capable of interpreting figurative language with the depth and nuance of human cognition. To achieve this, we need diverse evaluation strategies and datasets tailored to different layers of metaphor processing.

6. Limitations

Despite the promising results and the richness of the CroSloMet dataset, several limitations should be acknowledged. First, while our dataset includes both Croatian and Slovene examples, its scope remains limited to two closely related South Slavic languages, and future work should extend the framework

to more typologically diverse languages. Second, although our dataset contains both general and highly specific conceptual metaphors, as well as both conventional and novel linguistic metaphors, we did not annotate the examples according to these distinctions. Determining the appropriate level of specificity for such annotations remains a complex and partially subjective task, and applying these labels reliably across examples can be challenging. Introducing these categories without a clearly operationalized framework could therefore reduce annotation consistency rather than enhance it. Nonetheless, systematically incorporating these distinctions could be valuable for future work, and our findings highlight the need to further refine and formalize criteria for metaphor-type annotation in subsequent iterations of the dataset. Third, while our evaluation of metaphor explanation generation offers useful insights, it currently relies on a small validation set and overly strict metrics, which do not fully capture the semantic adequacy of generated explanations. More nuanced and scalable validation procedures are required. Lastly, the current implementation does not systematically include metaphorically rich genres such as poetry, which limits the representativeness of the dataset for creative metaphor use. Addressing these limitations will be a key focus in the continued development of CroSloMet.

Acknowledgments

This work was created as part of the bilateral Croatian-Slovene project *Automatic identification of semantic relations in figurative context in Croatian and Slovene* (2023-2025), the *Metaphor and Metonymy in Language and Thought* project funded by the European Union – NextGenerationEU (2024-2027), and the research programmes *No. P6-0411 (Language Resources and Technologies for Slovene)* and *No. P6-0215 (Slovene Language – Basic, Contrastive, and Applied Studies)* funded by the Slovenian Research and Innovation Agency.

References

Keiga Abe, Kayo Sakamoto, and Masanori Nakagawa (2006) "A computational model of the metaphor generation process." In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 28. https://escholarship.org/uc/item/5d96219g.

Špela Antloga (2020) "Korpus metafor KOMET 1.0." In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pp. 167–170.

Špela Antloga, and Gregor Donaj (2022) "Corpus of metaphorical expressions in spoken Slovene language G-KOMET 1.0." *Slovenian language resource repository CLARIN.SI*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma (2021) "We Need to Consider Disagreement in Evaluation." In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, Association for Computational Linguistics, pp. 15–21. https://doi.org/10.18653/v1/2021.bppf-1.3

Yuri Bizzoni, and Shalom Lappin (2018) "Predicting Human Metaphor Paraphrase Judgments with Deep Neural Networks." In *Proceedings of the Workshop on Figurative Language Processing*, New Orleans, Louisiana, pp. 45–55.

Marianna Bolognesi, Mario Brdar, and Kristina Š. Despot (Eds.) (2019) *Metaphor and Metonymy in the Digital Age. Theory and methods for building repositories of figurative language*, John Benjamins, Amsterdam.

Shubin Chen, and Zhen Wang (2025) "Challenges and Solutions for Large Language Models in Metaphor Translation of Political Texts." *Open Journal of Applied Sciences*, 15, 2360–2375. https://doi.org/10.4236/ojapps.2025.158158.

Barbara Dancygier, and Eve Sweetser (2014) *Figurative language*. Cambridge University Press, Cambridge.

Kristina Š. Despot (2024) *Metafora: Dekodiranje jezika imaginacije*. Institut za hrvatski jezik, Zagreb.

Kristina Š. Despot, Ana Ostroški Anić, and Tony Veale (2023) "Somewhere along your pedigree, a bitch got over the wall! A proposal of implicitly offensive language typology." *Lodz Papers in Pragmatics*, 19, 2, pp. 385–414. https://doi.org/10.1515/lpp2023-0019.

Kristina Š. Despot, Mirjana Tonković, Mario Brdar, Mario Essert, Benedikt Perak, Ana Ostroški Anić, Bruno Nahod, and Ivan Pandžić (2019) "MetaNet.HR: Croatian Metaphor Repository." In *Metaphor and Metonymy in the Digital Age*, edited by Marianna Bolognesi, Mario Brdar, and Kristina Š. Despot, John Benjamins Publishing Company, Amsterdam, pp. 123–146. https://doi.org/10.1075/milcc.8.06des.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019) "BERT: Pre-training of deep bidirectional transformers for language

understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao (2023) "Can Pre-trained Language Models Interpret Similes as Smart as Humans?" *ArXiv, abs/2203.08452*. https://doi.org/10.48550/arXiv.2203.08452.

Keith J. Holyoak, and Paul Thagard (1995) *Mental leaps: Analogy in creative thought*. MIT Press, Boston.

Rohan Joseph, Timothy Liu, Aik Beng Ng, Simon See, and Sunny Rai (2023) "NewsMet: A 'do it all' Dataset of Contemporary Metaphors in News Headlines." In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics, pp. 10090–10104.

Adam Kilgarriff, Vít Baisa, Jan Busta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, P. Rychlý, and Vít Suchomel (2014) "The Sketch Engine: ten years on." *Lexicography*, 1, pp. 7–36, https://doi.org/10.1007/S40607-014-0009-9.

Jeong Chul Kim, Sang Guk Suh, Lydia B. Chilton, and Haijun Xia (2023) "Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing." In *Designing Interactive Systems Conference (DIS '23)*. ACM, Pittsburgh, PA, New York, NY, https://doi.org/10.1145/3563657.3595996.

Matej Klemen, and Marko Robnik-Šikonja (2023) "Neural Metaphor Detection for Slovene." In *Selected papers from the CLARIN Annual Conference 2022*, edited by Tomaž Erjavec, and Maria Eskevich, Linköping Electronic Conference Proceedings 198, pp. 77–89, https://doi.org/10.3384/ecp198008.

Simon Krek, Špela Arhar Hodlt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc (2020) "Gigafida 2.0: the reference corpus of written standard Slovene." In *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Marseille, France*, edited by Nicoletta Calzolari, ELRA – European Language Resources Association, Paris, pp. 3340–3345, http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

George Lakoff (1993) "The contemporary theory of metaphor." In *Metaphor and Thought* (2nd ed.), edited by Andrew Ortony, Cambridge University Press, Cambridge, pp. 202–251.

George Lakoff, and Mark Johnson (1980) *Metaphors We Live by*. University of Chicago Press, Chicago.

George Lakoff, and Mark Johnson (1999) *Philosophy in the flesh*. Basic Books, New York, NY.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig (2022) "Testing the Ability of Language Models to Interpret Figurative Language." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Seattle, pp. 4437–4452, https://doi.org/10.18653/v1/2022.naacl-main.330.

Nikola Ljubešić, and Filip Klubička (2016) "Croatian web corpus hrWaC 2.1." *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042, http://hdl.handle.net/11356/1064.

Saif Mohammad, Ekaterina Shutova, and Peter Turney (2016) "Metaphor as a medium for emotion: An empirical study." In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, Berlin, Association for Computational Linguistics, pp. 23–33.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson (2016) "Introducing the LCC metaphor datasets." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, European Language Resources Association (ELRA), pp. 4221–4227.

Pragglejaz Group (2007) "MIP: A Method for Identifying Metaphorically Used Words in Discourse." *Metaphor and Symbol*, 22, 1, pp. 1–39, https://doi.org/10.1080/10926480709336752.

Ekaterina Shutova (2011) "Computational approaches to figurative language." *Language and Linguistics Compass*, 5,4, pp. 246–264.

Gerard J. Steen, Aletta G. Dorst, Berenike J. Herrmann, Anna A. Kaal, Tina Krennmayr, and Tryntje Pasma (2010) *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins Publishing Company, Amsterdam.

Kevin Stowe, Leonardo Ribeiro, and Iryna Gurevych (2020) Metaphoric paraphrase generation. *arXiv preprint arXiv:2002.12854 (2020)*, https://doi.org/10.48550/arXiv.2002.12854.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova (2024) Metaphor Understanding Challenge Dataset for LLMs. *arXiv:2403.11810v1 [cs.CL] 18 Mar 2024*, https://doi.org/10.48550/arXiv.2403.11810.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis (2021) "Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp. 4673–4686, https://doi.org/10.18653/v1/2021.naacl-main.372.

Peter Turney, and Michael L. Littman (2005) "Corpus-based Learning of Analogies and Semantic Relations." *Machine Learning*, 60, pp. 251–278, https://doi.org/10.1007/s10994-005-0913-1.

Tony Veale, and Yanfen Hao (2007). "Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language." *AAAI*, pp. 1471–1476, https://dl.acm.org/doi/abs/10.5555/1619797.1619881.

Tony Veale (2019). "Metaphor in the age of mechanical production (Or: Turning potential metaphors into deliberate metaphors)." In M. Bolognesi, M. Brdar, & K. Š. Despot (Eds.), *Metaphor and metonymy in the digital age: Theory and methods for building repositories of figurative language* (pp. 75–98). John Benjamins. https://doi.org/10.1075/milcc.8.04vea.

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Zhang, Yiliao Song, Hongfei Lin, and Feng Xia (2025). "Cultural Bias Matters: A Cross-Cultural Benchmark Dataset and Sentiment-Enriched Model for Understanding Multimodal Metaphors". In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pp. 26301–26317, https://doi.org/10.48550/arXiv.2506.06987.

SAŽETAK

Kristina Štrkalj Despot, Ana Ostroški Anić, Polona Gantar, Mija Bon, Matej Klemen, Marko Robnik Šikonja, Simon Krek, Benedikt Perak, Jaka Čibej

CROSLOMET: STRUKTURIRANI METAFORIČKI SKUP PODATAKA ZA HRVATSKI I SLOVENSKI JEZIK

Ubrzan razvoj velikih jezičnih modela otvorio je nove mogućnosti za obradu figurativnoga jezika, no njihovo tumačenje značenja metafora i metaforičkih izraza i dalje zaostaje za razinom ljudskoga razumijevanja. Jedno od ograničenja jezičnih modela proizlazi iz nedostatnosti postojećih skupova podataka o metaforama, koji često nemaju jasno izražene veze s konceptualnim metaforama te su uglavnom jednojezični. U ovom radu predstavljamo CroSloMet, novi skup podataka s više od 1120 metaforičkih i 1120 doslovnih rečenica na hrvatskom i slovenskom jeziku, utemeljen na bazi metafora MetaNet.HR. Svaki je primjer označen pripadajućom konceptualnom metaforom, višerječnim jezičnim izrazom, kanonskim oblicima i doslovnom upotrebom, što omogućuje provedbu zadataka određivanja i objašnjavanja metafora. U radu su prikazane preliminarne evaluacije skupa podataka kroz dva eksperimenta: klasifikaciju metafora s pomoću modela CroSloEngual BERT-a, gdje je postignuta točnost od 88,5 %, te generiranje objašnjenja metafora s pomoću modela LLama 3-8B, pri čemu je stroga evaluacija točnoga podudaranja dala niske rezultate unatoč semantički valjanim rezultatima. Kako bismo to prevladali, predlažemo višerazinsku metodologiju validacije koja kombinira ručno označavanje, zaključivanje prirodnim jezikom, semantičku sličnost i prosudbu temeljenu na velikom jezičnom modelu. Naši rezultati naglašavaju važnost obuhvaćanja razina općenitosti i specifičnosti u metaforičkom preslikavanju te pokazuju na potrebu za nijansiranijim metodama evaluacije. CroSloMet je resurs za unaprjeđenje razumijevanja metafora u velikim jezičnim modelima i doprinosi međujezičnom i kognitivno utemeljenom istraživanju metafora.

Ključne riječi: *metafore; metaforički skup podataka; objašnjavanje metafora; razumijevanje metafora; veliki jezični modeli*