Introducing PerMet 1.0: A Metaphor-Annotated Corpus for Persian

Mohammad Saeid Miri

Allameh Tabataba'i University, Tehran, Iran
ms_miri@outlook.com

Abstract

Metaphor plays a central role in human language and thought, and corpus-linguistic approaches enable its systematic investigation. Such research requires large, representative collections of metaphor-annotated linguistic data from diverse contexts. Despite the increasing availability of metaphor corpora in various languages, Persian remains underrepresented, with few publicly available resources and no large-scale register-diverse metaphor corpus. This paper introduces PerMet 1.0, a metaphor-annotated corpus for Persian. The corpus consists of approximately 120,000 tokens (about 99,000 lexical units) drawn from five registers: academic, news, fiction, social media, and spoken discourse. Five independent annotators labeled the corpus using Metaphor Identification Procedure Vrije Universiteit (MIPVU), with adaptations for Persian. Inter-annotator agreement showed a high level of consistency (κ = 0.952), confirming the reliability of the annotation. Preliminary analysis shows that 13.1% of the lexical units are related to metaphor, with the academic register showing the highest proportion, followed by news, social media, spoken, and fiction. PerMet 1.0 offers a foundational resource for research on metaphor in Persian, cross-linguistic comparative studies, and the development and fine-tuning of machine learning or large language models for automatic metaphor identification.

Keywords: metaphor corpus, metaphor identification, Persian metaphor corpus, metaphor identification procedure, semantic annotation

1. Introduction

Metaphor, the process of understanding one domain in terms of another, is a crucial aspect of how people talk, think, and communicate. It shapes how we speak and understand concepts, enabling language users to comprehend complex or unfamiliar concepts by relating them with more concrete ones (Lakoff and Johnson, 1980). It seems unlikely for a text to lack metaphorical words. For example, in "He's a heavy smoker" (Lakoff et al., 1991), the word "heavy" conceptualizes the quantity of smoking in terms of weight, illustrating the conceptual metaphor AMOUNT IS A PHYSICAL PROPERTY.

Metaphor has been discussed from a variety of philosophical and linguistic perspectives (Black, 1955; Bowdle and Gentner, 2005; Lakoff and Johnson, 1980; Ortony, 1993). In addition to its theoretical development within linguistics, metaphor has received increasing attention in corpus linguistics over the past two decades. This attention reflects a growing interest in studying metaphor through concordance lines and real linguistic data, rather than relying on made-up examples (Krennmayr and Steen, 2017). To this end, researchers required a valid and reliable method for determining whether a word or lexical item is used metaphorically. Metaphor identification protocols, most notably Metaphor Identification Procedure (MIP, Pragglejaz Group, 2007) and its enhanced version Metaphor Identification Procedure Vrije Universiteit (MIPVU, Steen et al., 2010), were developed in response to this need. These protocols provide the groundwork for building metaphor-annotated corpora, such as the VU Amsterdam Metaphor Corpus (VUA, Krennmayr and Steen, 2017).

The role of metaphor in everyday communication, on the one hand, and recent advances in Natural Language Processing (NLP) and Artificial Intelligence (AI), on the other, have intensified the need for language models to understand metaphor and make it a crucial topic of inquiry (Ge et al., 2023; Ptiček and Dobša, 2023). Language models often struggle with understanding and interpreting figurative expressions, especially metaphors (Ge et al., 2023) Providing high-quality, metaphor-annotated data can enhance the ability of language models to process figurative language and perform key NLP tasks (e.g., machine translation, sentiment analysis) more accurately. It is, therefore, necessary to build a large and representative metaphor corpus to facilitate metaphor analyses in real contexts and enable the development and fine-tuning of language models to better handle non-literal language.

Since the introduction of MIP and later MIPVU, several metaphor corpora have been published, most notably VUA (Krennmayr and Steen, 2017). Although there is growing interest in figurative language in Persian, particularly metaphor, and several general-purpose Persian corpora are available, only a few metaphor-annotated datasets exist, and those are not representative and register-diverse. As a result, there remains a significant lack of metaphor-annotated corpora for the Persian language. This paper addresses this gap by introducing the first version of a large, representative, register-diverse corpus of Persian metaphors, named PerMet 1.0. It contains 120,343 tokens (99,079 lexical units) from five

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4835–4845
11–16 May 2026.

distinct registers: academic texts, news, social media, fiction, and spoken. The corpus was manually annotated by five annotators following the MIPVU (Steen et al., 2010) and includes detailed labels for metaphor-related words (MRWs). PerMet 1.0 provides high-quality data for metaphor research and the development of language models.

The paper is structured as follows. Section 2 presents a brief overview of the literature and evolution of ways to identify metaphors, along with a survey of published metaphor corpora. Section 3 describes the methodology and details the development of PerMet 1.0. Section 4 shows the results of inter-annotator agreement tests. Section 5 presents a quantitative analysis of the corpus data. Finally, Section 6 concludes the paper by summarizing the main findings, discussing the limitations, and suggesting future directions.

2. Metaphor Identification and Corpus Compilation

Conceptual Metaphor Theory (CMT, Lakoff and Johnson, 1980) established a basis for the study of metaphor in language use and inspired growing interest among linguists. This concern comprised not only the cataloging of metaphorical expressions but also the providing of empirical support for CMT's theoretical assertions through the analysis of large corpora. Early efforts to operationalize CMT and address its methodological challenges appeared in the late 1990s, most notably in the volume *Researching and Applying Metapho*r (Cameron and Low, 1999). At that time, metaphor research generally lacked explicit identification procedures, and researchers relied largely on intuition. This method worked well for canonical examples, but it was insufficient for corpus-based studies because it was difficult to determine if many examples were (non-)metaphorical without clear criteria (Stefanowitsch, 2007).

Proposals by Steen (1999, 2002) and Crisp et al. (2002) eventually culminated in the Metaphor Identification Procedure (MIP; Pragglejaz Group, 2007), the first fully elaborated method for identifying metaphor. MIP defined a lexical unit as metaphorical when its contextual meaning contrasts with a more basic meaning—defined as more concrete, precise, bodily related, and historically older—and that contrast can be understood through comparison (Pragglejaz Group, 2007). For instance, in "Both left and right agree taxes must rise" (The Guardian, emphasis added), the basic meaning of *rise* is "to move upwards," while its contextual meaning is "to increase", according to The Cambridge Dictionary (n. d.). There is a contrast between the two meanings, which can be understood by comparison: we can understand the increase in quantity in terms of physical movement.

Since its introduction, MIP has been widely used for metaphor research, including corpus-based metaphor analysis (e.g., Deignan, 2005; Semino et al., 2017), critical metaphor analysis (e.g., Charteris-Black, 2004), and automatic metaphor identification (e.g., Choi et al., 2021). Although MIP is explicit, straightforward, and based on minimal assumptions, its drawbacks became evident, especially in terms of its validity and reliability, when applied to a large corpus. Steen et al. (2010) developed the Metaphor Identification Procedure Vrije Universiteit (MIPVU) to resolve these issues; it is an enhanced and more comprehensive version of MIP. Its basic procedure can be summarized as follows (Steen et al., 2010, pp. 25–26):

1. Find metaphor-related words (MRWs) by examining the text on a word-by-word basis.
2. When a word is used indirectly and that use may potentially be explained by some form of cross-domain mapping from a more basic meaning of that word, mark the word as metaphorically used (MRW).
3. When a word is used directly and its use may potentially be explained by some form of cross-domain mapping to a more basic referent or topic in the text, mark the word as direct metaphor (MRW, direct).
4. When words are used for the purpose of lexico-grammatical substitution, such as third person personal pronouns, or when ellipsis occurs where words may be seen as missing, as in some forms of co-ordination, and when a direct or indirect meaning is conveyed by those substitutions or ellipses that may potentially be explained by some form of cross-domain mapping from a more basic meaning, referent, or topic, insert a code for implicit metaphor (MRW, implicit).
5. When a word functions as a signal that a cross-domain mapping may be at play, mark it as a metaphor flag (MFlag).
6. When a word is a new-formation coined, examine the distinct words that are its independent parts according to steps 2 through 5.

Returning to the aforementioned example, "Both left and right agree taxes must rise", the basic and contextual meanings of *rise* are as described above. According to MIPVU, the two senses are sufficiently distinct, as they have two separately numbered sense descriptions. In addition, these two meanings display a form of similarity, which can be explained through a cross-domain mapping from a more basic, concrete meaning (physical movement) to an abstract one (increasing in quantity). Therefore, in this context, *rise* is labeled as an indirect metaphor-related-word (MRW).

MIPVU established the basis not only for systematically analyzing metaphor in discourse (Steen et al., 2010), but also for developing

metaphor identification procedures in other modalities (e.g., VISMIP, Šorm and Steen, 2018; FILMIP, Bort-Mir, 2019), for adapting MIPVU to other languages (e.g., Nacey, Dorst, et al., 2019), and for compiling large metaphor-annotated corpora. The VU Amsterdam Metaphor Corpus (Steen et al., 2010) was the first published metaphor-annotated corpus based on MIPVU. Building on the success of MIPVU and VUA, a growing body of research has focused on language-specific adaptations of MIPVU, and metaphor corpora have since been developed in several languages. Badryzlova et al. (2013) applied MIPVU to annotate metaphors in Russian. Lu and Wang (2017) introduced a 30,000-token corpus for Mandarin Chinese in three registers (academic, news, and fiction), the PSU Chinese Metaphor Corpus (CMC). Antloga (2020) compiled KOMET 1.0, a 200,000-token metaphor corpus for Slovenian from news, fiction, and online texts. Ptiček (2025) introduced an 87,000-token corpus for Croatian, drawn from news, culture, and literary science texts.

The development of procedures such as MIP and MIPVU has transformed metaphor research by enabling the compilation of metaphor-annotated corpora in multiple languages and allowing scholars to investigate metaphor usage across diverse registers. Nevertheless, Persian remains underrepresented in this line of research, with few published metaphor corpora available. Tsvetkov et al. (2014) analyzed data from multiple languages, including 364 sentences in Persian; however, the Persian data is inaccessible. Assi et al. (2022) applied a modified version of MIPVU to teacher–student discourse among primary school children, and Miri (2024) evaluated an adapted version of MIPVU for the development of a Persian metaphor corpus. As of now, there is only one publicly available metaphor corpus for Persian, the Persian MIPVU Corpus (PMC, Bakhtiyari and Iravani, 2025). The PMC consists of 30,000 tokens obtained from news texts and manually annotated based on MIPVU. Consequently, accessible resources are limited, and a large, register-diverse, and representative metaphor corpus does not currently exist for Persian. PerMet 1.0 is intended to serve as a representative resource for metaphor research in Persian.

3. Method

This section outlines the method for designing PerMet 1.0. It includes the data collection, annotation process, lexical resources used for applying MIPVU to Persian texts, and strategies for addressing language-specific issues.

3.1 Corpus Data

PerMet 1.0 is a 120,343-token (99,079 lexical units) corpus from five distinct registers: academic, news, fiction, social media, and spoken. Corpus data were collected through four methods: (1) random sampling from publicly accessible corpora, (2) web scraping (using the BeautifulSoup[1] Python library), (3) Optical Character Recognition (OCR) of scanned texts (using Google Vertex AI Platform[2]), and (4) transcription of audio materials (using Avanegar[3] speech-to-text tool). The academic subcorpus consists of a random sampling from the Ferdowsi Annotated Academic Language Corpus (Kamyabi Gol et al., 2018) and web scraping from published articles in Persian scientific journals in four disciplines: engineering, medical sciences, natural sciences, and humanities. For the fiction subcorpus, 32 random samples were taken from four Persian novels (Maroufi, 1989, 2003; Pirzad, 2001; Vafi, 2005) and digitized using OCR. The news subcorpus includes data gathered using web scraping from 60 articles published in four major Persian news websites (IRNA, Shargh Daily, IRIB News, and Hamshahri Online) in five domains: politics, economy, society, culture, and sports. For the social media subcorpus, 1,350 unique tweet entries (2,264 tweets in total) were collected from the Large Scale Colloquial Persian Language (Abdi Khojasteh et al., 2020). The spoken subcorpus took 36 speeches from the Hambam Corpus (Haig and Rasekh-Mahand, 2025). It includes diaries, media interviews, and lectures from native Persian speakers (20 female and 16 male). Table 1 shows further details for each register.

Register	No. of items	No. of LUs	No. of tokens
Social Media	1,350	19,718	24,465
Academic	97	19,088	23,735
Fiction	89	19,228	24,341
News	60	21,635	25,119
Spoken	36	19,410	22,683
Total	**1,632**	**99,079**	**120,343**

Table 1: Corpus data information (LU: Lexical Unit)

The collected texts required normalization due to their diverse sources. Subsequently, they underwent preprocessing prior to annotation. This process involved sentence segmentation, tokenization, lemmatization, and Part-of-Speech (POS) tagging, all done with the Stanza library (Qi et al., 2020), which fully supports Persian. Minor corrections in tokenization and POS tagging were applied to the text, particularly for social media and spoken data.

[1] https://pypi.org/project/beautifulsoup4/
[2] https://cloud.google.com/vertex-ai
[3] https://ivira.ai/speech-to-text/

3.2 Annotation Process

The corpus was manually annotated following MIPVU (Steen et al., 2010) protocols. Four Persian native speaker postgraduates with backgrounds in linguistics were recruited in the annotation process, alongside the author. Except for the author, other annotators had no prior experience in metaphor annotation. They independently analyzed texts word by word to determine whether a lexical unit should be labeled as a Metaphor-Related Word (MRW). They received an initial training package consisting of 3 hours of video tutorials on what the MIPVU is, how to apply it to Persian data, and how to use dictionaries. After training, they completed a pilot task of annotating a 141-token text. Subsequently, they received a half-hour video for clarifying the annotation process and resolving their deviations. Each annotator labeled one register, except the author, who supervised other annotators and labeled fiction and parts of the academic and spoken subcorpora.

The annotation process was conducted in five phases. In the first phase, each annotator independently labeled parts of the assigned texts and sent their results to the author. In the second phase, the author looked over the results, noted any deviations, and sent the texts back to them for changes. In the third phase, annotators resolved the deviations and labeled the remainder of the data. In the fourth phase, after a comprehensive review of the data, multiple rounds of group discussion were conducted via a social media platform, and two online meetings were held to resolve remaining disagreements and reach a consensus on the majority of conflicting issues. Finally, the annotators revised the data based on the discussions.

3.3 Lexical Resources

We looked up the basic and contextual meanings of lexical units (LUs) in contemporary, corpus-based dictionaries to determine whether they are used metaphorically (Steen et al., 2010). The primary resources used were the Advanced Learner's Persian Dictionary (ALPD, Assi, 2019) and Sokhan Comprehensive Dictionary (SCD, Anvari, 2002). Where necessary, particularly in cases of annotator disagreement or missing entries, we used FarsNet (Shamsfard et al., 2010) as a supplementary reference. In very few cases, where none of them had an entry for a word, we used the online version of the Dehkhoda Dictionary[4].

3.4 Annotation Categories

PerMet was formatted in CoNLL-U Plus, which is an extended version of CoNLL-U that allows the addition of task-specific annotations. PerMet kept six essential columns from the original CoNLL-U format: ID (token index), FORM (word form or punctuation symbol), LEMMA (lemma form), UPOS (universal part-of-speech tag), XPOS (optional language-specific part-of-speech tag), and MISC (other annotations). Several metaphor-related columns were added, some of which are based on the VUA tagset (Krennmayr and Steen, 2017), other versions of MIPVU (e.g., Nacey, Krennmayr, et al., 2019), and PARSEME 1.3 corpus (Savary et al., 2023). Figure 1 provides detailed information about the annotation format. Each column is described as follows:

- **LU**: Identifies whether a token is considered as a lexical unit (*LU*).
- **LU:TYPE**: Specifies the type of a lexical unit: W (any word), MWE (multi-word expressions), CN (compound nouns), PRT (parts of an MWE), or I (ignored, e.g., punctuation marks, numerals)
- **MWE:TYPE**: Specifies the type of multi-word expression: LVC.full (for compound verbs), VID (for verbal idioms), and UNCAT (for other MWEs).
- **MRW**: Identifies whether a lexical unit is metaphorical (MRW).
- **MRW:TYPE**: Specifies the type of relation to metaphor: IND (indirect), DIR (direct), IMPL (implicit), or SIG (metaphor flag).
- **BL:TYPE**: Marks lexical units whose metaphorical status is unclear:
    - *WIDLII* ('When In Doubt, Leave It In'): For unresolved cases even after group discussion.
    - *DFMA* (Discarded From Metaphor Analysis): For cases where it is impossible to determine the contextual meaning.
- **MFLAG**: Specifies the linguistic level of the metaphor flag: LEX (lexical, a word), MORPH (morphological, part of a word), PHRASE (phrasal, a phrase), or PSUS (pseudo-sentence).

The following examples (1-3) demonstrate how the mentioned tagset was applied in PerMet 1.0. The most common example of MRWs is indirect metaphor. According to MIPVU, a lexical unit is labeled as an indirect metaphor if the contextual meaning exhibits some form of cross-domain mapping from its (more) basic meaning. Example (1), taken from the news subcorpus, issues a warning about desertification in Iran.

(1) *biyābān-zāyi* *dar* *kešvar* ***be***
desertification in country to
noqte=ye *hošdār*
point=EZ warning
reside ***ast***
reach-PST-PTCP be-PRS.3SG
'Desertification in the country has reached the warning point.'

[4] https://dehkhoda.ut.ac.ir/fa/dictionary

Three lexical units are used metaphorically in this context, marked in bold: ***be*** ('to'), ***noqte=ye*** ('point'), and ***reside ast*** ('has reached'). The contextual meaning of the preposition ***be***, 'an indication of the direction of movement or the purpose of a process' (ALPD, sense 3) can be understood in contrast with the basic meaning, which is 'an indication of the destination or end of a route' (ALPD, sense 1). The basic meaning refers to a concrete, physical location, while the contextual meaning is an abstract concept; thus, this lexical unit was labeled as an indirect metaphor. ***Noqte=ye*** is another indirect metaphor that means 'a position or stage in the course of something' (SCD, sense 8), which contrasts with

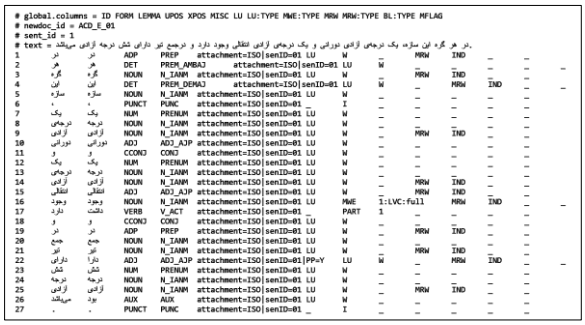

Figure 1: Example of metaphor annotation in PerMet 1.0

the basic meaning, 'place or location' (SCD, sense 1). The meanings are sufficiently distinct, and a cross-domain mapping from LOCATION to STATE is evident. The contextual meaning of *residan* ('to reach') is 'to be in a certain stage or situation' (SCD, sense 17), contrasting the basic meaning 'to arrive at one's desired destination or place' (SCD, sense 1) and a cross-domain mapping from LOCATION to STATE can be seen. Accordingly, these three lexical units were coded as indirect metaphors, together evoking the conceptual metaphor CHANGE OF STATE IS CHANGE OF LOCATION (Lakoff et al., 1991), as if desertification is a moving object that has reached a new location, the "warning point".

(2) *az* *birun* ***mesl=e*** *yek* ***qāzi,***
from outside like=EZ a judge
jeddī, *va* *az* *darun* ***mesl=e***
serious and from inside like=EZ
yek ***dalqak,*** *lowde*
a clown ludicrous
'From the outside, serious like a judge, and from the inside, ludicrous like a clown.'

Example (2), taken from the fiction subcorpus, illustrates a direct metaphor, in which a lexical unit is used directly through linguistic cues signaling a cross-domain mapping. In this fragment, one character describes another person's trait using direct metaphors (marked in bold), *qāzi* ('judge') and *dalqak* ('clown'), with *mesl=e* ('like') functioning as the metaphor flag. The character overtly used words that signify a topic shift from someone's traits to a judge and clown, which can be understood through a cross-domain mapping. Thus, *qāzi* and *dalqak* were annotated as direct metaphors, and *mesl=e* received a SIG tag. Also, it should be noted that the words ***az*** ('from'), ***birun*** ('outside'), and ***darun*** ('inside') were annotated as indirect metaphors, as they exhibited a contrast based on a cross-domain mapping from their basic spatial meanings.

The third category of MRWs is implicit metaphor, when a substitution or ellipsis occurs, and the substituted or omitted element conveys a meaning that can be understood by some form of cross-domain mapping (Steen et al., 2010). Example (3), drawn from the academic subcorpus, concerns mutation testing (or mutation analysis), a software testing method in which small changes ('mutations') are introduced into a program's code to evaluate its quality.

(3) *in* *kār* *bā* *estexrāj=e*
this task with extraction=EZ
vižegi-hā=ye *moxtalef* *az*
feature-PL=EZ various from
jaheš-hā *va* *estefāde*
mutant-PL and use
az ***ānhā*** *barāye* *āmuzeš=e*
from 3PL for training=EZ
model-hā=ye *yādgiri=ye* *māšin*
model-PL=EZ learning=EZ machine
anjām *mi-šav-ad*
do IPFV-become.3SG
'This task is done by extracting various features from the mutants and using them to train machine learning models.'

The third-person plural pronoun (*ānhā*) substituted *jaheš-hā* ('mutants'), which was labeled as an indirect metaphor; both are marked in bold. Since the dictionaries have no entry for 'mutant' and the authors of sentence (3) used *jaheš-hā* ('mutations') to translate 'mutant', annotators used the closest lexical entry, *jaheš* ('mutation'). The contextual meaning of *jaheš*, which is not listed in dictionaries, is 'a small change in a program's code'. This contrasts with the basic meaning: '[Biology] a sudden change in chromosomes or genes of a living organism that causes changes in the structural features or growth processes of that organism and its next generation' (ALPD, sense 3). Comparing the two meanings reveals a cross-domain mapping from GENE to CODE, conceptualizing software mutants as biological organisms. Accordingly, *jaheš-hā* was labeled as an indirect metaphor, and *ānhā* as an implicit metaphor.

3.5 Language-Specific Issues

Although MIPVU provides explicit protocols to demarcate different types of linguistic units, since it was primarily developed for English, any application in other languages requires several adjustments. There are several linguistic features in Persian that need methodological adjustments to align with MIPVU.

Compound Verbs (CVs): One of the language-specific issues is Persian Compound Verbs or Complex Predicates, which are highly productive in Persian. Since the number of simplex verbs is limited, most Persian verbs are made through combinations of a verb and non-verbal elements (Samvelian and Faghiri, 2013). These may be noun-verb, such as *yād gereftan* 'to learn' (lit. learning-take), adjective-verb, such as *xošhāl kardan* 'to make happy' (lit. happy-do), particle-verb, such as *bar-dāštan* 'to pick up' (lit. PARTICLE-have), or prepositional phrase-verb, such as *be dast āvardan* 'to obtain' (lit. to-hand-bring) constructions (Dabir-Moghaddam, 1997; Samvelian and Faghiri, 2013). Except for some particle-verb constructions, Persian CVs typically consist of more than one token, but demonstrate one action, process, or event (Mirzaei, 2015). PerMet 1.0 adopted the approach of treating multi-token verbs as single lexical units, following MIPVU's protocols, while also making further adjustments suited to Persian CVs.

Specifically, each CV, whether single- or multi-token, received the tag 'MWE', stands for Multi-Word Expression in 'LU:TYPE' column, and the first element received its specific tag in 'MWE:TYPE' column, which is based on PARSEME 1.3 tagset for Persian (Savary et al., 2023). For example, the first token of the compound verb *be dast mi-āvar-ad* (PREP hand IPFV-bring-3SG, 'he/she obtains') received a 'MWE' tag and the other elements were tagged with 'PRT'. Only the first element, *dast*, has 'LU' and '1:LVC.full' tag. Other parts received '1' tag in 'MWE:TYPE' column. This protocol supported both separable and non-separable CVs, infinitives, and past participles. Annotators used dictionaries to identify CVs; if an element-verb combination was found as a main or sub-entry, it was a CV that should be treated as a single lexical unit.

Polywords: Another language-specific issue is polywords, like *banā bar in* 'therefore' or *afzun bar in* 'in addition', which are fixed multi-word expressions with discourse functions. According to MIPVU, such expressions were treated as single lexical units because they designate a single referent in discourse (Steen et al., 2010). PerMet 1.0 followed the MIPVU protocols and treated polywords as single lexical units. However, unlike English, identifying polywords in this study was not an effortless job. Dictionaries include only a limited number of highly frequent polywords and often do not cover their variants. For example, *az taraf=e digar* ('on the other hand') is listed in the ALPD, but the variant *az taraf-i digar* is not included. To address this, PerMet 1.0 treated polywords as single lexical units, assigned the tag 'MWE' to the first element and 'PRT' to the rest, and tagged the first element with 'UNCAT' in 'MWE:TYPE' column. It also added the "ListedForm=" tag in the MISC column to link the unlisted variants to the recorded form in dictionaries. The same approach was applied to compound prepositions (e.g., *bar pāye=ye* 'based on') and idiomatic expressions (e.g., *āberu=ye kasi rā bordan* 'to disgrace someone').

Compound Nouns: PerMet 1.0 followed the MIPVU protocols in treating compound nouns as single lexical units, similar to the decision for compound verbs. Compound nouns were tagged with the 'CN' (for the first element) and 'PRT' labels and treated as single lexical units, receiving one LU tag.

Clitics: A clitic is "a form which resembles a word, but which cannot stand on its own as a normal utterance being phonologically dependent upon a neighbouring word (its host) in a construction" (Crystal, 2008, p. 80). In PerMet 1.0, clitics (such as pronominal enclitics or enclitic forms of the 'to be' verb) were treated as separate tokens but single lexical units. Following the PerDT corpus (Rasooli et al., 2013) conventions, the 'attachment=' tag was used to indicate their attachment status (ISO = isolated, NXT = attached to the next token, PRV = attached to the previous token).

```
# global.columns = ID FORM LEMMA UPOS XPOS MISC LU LU:TYPE MWE:TYPE MRW MRW:TYPE BL:TYPE MFLAG
# newdoc_id = ACD_S_12
# sent_id = 116
# text = افزون بر این، مطالعه‌ی ترمودینامیکی نشان داد، فرایند جذب داروی دوکسوروبیسین با جاذب خودبه‌خودی، گرمازا و با کاهش آنتروپی همراه است.
1   افزون   افزون   ADJ    ADJ_AJP  attachment=ISO|senID=116  LU  MWE   1:UNCAT  _    _    _  _
2   بر      بر      ADP    PREP     attachment=ISO|senID=116  _   PART  1        _    _    _  _
3   این     این     PRON   PR_DEMON attachment=ISO|senID=116  _   PART  1        _    _    _  _
4   ،       ،       PUNCT  PUNC     attachment=ISO|senID=116  _   I     _        _    _    _  _
5   مطالعه‌ی مطالعه  NOUN   N_IANM   attachment=ISO|senID=116  LU  W     _        _    _    _  _
6   ترمودینامیکی ترمودینامیکی ADJ ADJ_AJP attachment=ISO|senID=116 LU W _     _    _    _  _
7   نشان    نشان    NOUN   N_IANM   attachment=ISO|senID=116  LU  MWE   2:LVC:full MRW IND _  _
8   داد     داد     VERB   V_ACT    attachment=ISO|senID=116  LU  PART  2        _    _    _  _
9   ،       ،       PUNCT  PUNC     attachment=ISO|senID=116  _   I     _        _    _    _  _
10  فرایند  فرایند  NOUN   N_IANM   attachment=ISO|senID=116  LU  W     _        _    _    _  _
11  جذب     جذب     NOUN   N_IANM   attachment=ISO|senID=116  LU  W     _        _    _    _  _
12  داروی   دارو    NOUN   N_IANM   attachment=ISO|senID=116  LU  W     _        _    _    _  _
13  دوکسوروبیسین دوکسوروبیسین PROPN N_IANM attachment=ISO|senID=116 LU W _      _    _    _  _
14  با      با      ADP    PREP     attachment=ISO|senID=116  LU  W     _        MRW  IND  _  _
15  جاذب    جاذب    NOUN   N_IANM   attachment=ISO|senID=116  LU  W     _        _    _    _  _
16  خودبه‌خودی خودبه‌خودی ADJ ADJ_AJP attachment=ISO|senID=116  LU  W     _        _    _    _  _
17  ،       ،       PUNCT  PUNC     attachment=ISO|senID=116  _   I     _        _    _    _  _
18  گرمازا  گرمازا  ADJ    ADJ_AJP  attachment=ISO|senID=116  LU  W     _        _    _    _  _
19  و       و       CCONJ  CONJ     attachment=ISO|senID=116  LU  W     _        _    _    _  _
20  با      با      ADP    PREP     attachment=ISO|senID=116  LU  W     _        MRW  IND  _  _
21  کاهش    کاهش    NOUN   N_IANM   attachment=ISO|senID=116  LU  W     _        _    _    _  _
22  آنتروپی آنتروپی NOUN   N_IANM   attachment=ISO|senID=116  LU  W     _        _    _    _  _
23  همراه   همراه   ADJ    ADJ_AJP  attachment=ISO|senID=116  LU  W     _        MRW  IND  _  _
24  است     است     AUX    AUX      attachment=ISO|senID=116  LU  W     _        _    _    _  _
25  .       .       PUNCT  PUNC     attachment=ISO|senID=116  _   I     _        _    _    _  _
```

Figure 2: Example of annotating compound verbs and polywords in PerMet 1.0

To help understand the aforementioned guidelines, Figure 2 shows the annotation of a sample sentence containing the compound verb *nešān dād* (show give-PST.3SG, '[it] showed'), spanning tokens 7 and 8, and the polyword *afzun bar in* ('in addition'), comprising tokens 1 through 3. Both the CV and the polyword are assigned the 'LU' tag only on their initial components (*nešān* and *afzun*, respectively), while subsequent tokens are labeled with an underscore (_) in the 'LU' column. Furthermore, the 'LU:TYPE' column distinguishes these components by assigning the 'MWE' tag to the initial token and the 'PART' tag to all succeeding elements. Detailed information regarding the specific type of CV or polyword is included in the 'MWE:TYPE' column. Notably, the 'MRW' tag is restricted to the initial component of the CV (*nešān*), the only token with the 'LU' tag.

4. Inter-Annotator Agreement

Two rounds of inter-annotator agreement tests were conducted to ensure the reliability of PerMet 1.0. In these tests, the focus is on whether a lexical unit is an MRW (marked as 1) or not (marked as 0), since the other MRW types (direct and implicit) are very infrequent. In both tests,

reliability was calculated using Fleiss' kappa coefficient. The first test was conducted at the beginning of the project, after the training phase (before annotating the corpus). Four annotators labeled a short news article containing 125 lexical units. The resulting kappa was 0.594, which indicates a *moderate agreement* (Landis and Koch, 1977). The second test was conducted at the final phase, after iterative discussions. A random sample of 5,068 lexical units (i.e., 5% of the corpus) was selected. The sample consisted of approximately 1,000 lexical units from each register. Three annotators independently labeled the data using the same binary classification scheme. The results, summarized in Table 2, show very high Fleiss' kappa values, ranging from 0.934 to 0.964, with an overall value of 0.952, which indicates *almost perfect agreement* (Landis and Koch, 1977).

Test	Register	LUs	Fleiss' κ
1	News	125	0.594
2	News	1,010	0.964
	Academic	1,012	0.961
	Spoken	1,016	0.947
	Social Media	1,015	0.946
	Fiction	1,015	0.934
	Overall	5,068	0.952

Table 2: Results of inter-annotator agreement tests (LU: Lexical Unit)

The main reason for inconsistency in the first test was deviation from the procedure. In some cases, annotators violated protocols for demarcating lexical units, such as incorrectly identifying a CV as a simple verb and assigning MRW tags to one (or all) of them. At times, they treated any semantic difference as a reason for metaphoricity or made errors in determining the more basic meaning. Comparison of the two tests shows a substantial improvement in inter-annotator agreement during the development of the corpus. The improvement was achieved through multiple rounds of review and revision throughout the annotation, multiple group discussions after finishing it, and resolving the conflicting issues.

The result is consistent with findings from previous studies in other languages and in Persian. Earlier research has reported a range from 0.70 to 0.97 (mostly after discussion) and 0.7 0.760 to 0.964 for Persian (Bakhtiyari and Iravani, 2025; Miri, 2024). The kappa value of 0.952 for PerMet demonstrates the effectiveness of the identification procedure, training, and discussions.

5. Corpus Analysis

This section presents a quantitative overview of PerMet 1.0, focusing on the distribution of metaphor-related words (MRWs) across different registers and expanding on the distribution of MRWs across major word classes. Then, the results will be compared with findings from other studies.

5.1 MRWs across registers

Table 3 displays the overall number and proportion of MRWs and non-MRWs in each register. Out of 99,122 lexical units, only 13.1% were identified as MRWs. The academic register has the most MRWs (19.7%), followed by news (16.8%), social media (11.8%), and spoken (9.7%). Counter-intuitively, the fiction (7.4%) register displays the lowest proportion.

	LUs	Non-MRWs	MRWs	%MRWs
Acad.	19,088	15,323	3,765	19.7
News	21,635	18,001	3,634	16.8
Soc. M.	19,718	17,398	2,320	11.8
Spoken	19,410	17,528	1,882	9.7
Fiction	19,228	17,801	1,427	7.4
Total	99,079	86,051	13,028	13.1

Table 3: Distribution of metaphor-related words across registers

To contextualize the distribution of metaphors in PerMet 1.0, Table 4 shows the percentage of metaphor-related words (MRWs) across various registers with other metaphor corpora, including the VU Amsterdam Corpus for English (Krennmayr and Steen, 2017), the PSU Chinese Metaphor Corpus (Lu and Wang, 2017), and the Persian Metaphor Corpus (Bakhtiyari and Iravani, 2025).

PerMet 1.0 has 13.1% MRWs, which is about the same as VUA (13.6%) and slightly higher than PMC (11.9%) and CMC (11.2%). In all corpora, except PMC, academic and news texts exhibit the greatest MRWs across all registers. There is also a fairly high percentage of MRWs (11.8%) in the social media register. However, fiction and spoken registers have lower MRWs (9.7% and 7.4%, respectively), which contrasts with the VUA results. These discrepancies are likely due to limitations in data collection and metaphor corpus compilation for Persian, which will be discussed in Section 6.

	PerMet 1.0	VUA	CMC	PMC
Acad.	19.7	18.5	16.3	-
News	16.8	16.4	9.5	11.9
Soc. M.	11.8	-	7.9	-
Spoken	9.7	7.7	-	-
Fiction	7.4	11.9	-	-
Total	13.1	13.6	11.2	11.9

Table 4: Comparison of metaphor percentage across PerMet 1.0 and VUA, CMC, and PMC.

5.2 Metaphor Types across Registers

Table 5 illustrates the distribution of different MRW types, metaphor flags, and borderline cases across registers. As expected, indirect metaphors

show the majority (95.2%) of all MRWs, consistent with findings from other languages (e.g., Steen et al., 2010; Nacey, Dorst, et al., 2019). Direct and implicit metaphors are almost rare, with less than 5% of the total.

	IND	DIR	IMPL	MF	WIDLII	DFMA
Acad.	3,749	10	6	0	24	0
News	3,573	49	12	5	26	0
Soc.	2,114	206	0	23	14	38
Spoken	1,826	48	8	14	7	1
Fiction	1,141	283	3	55	13	0
Total	12,403	596	29	97	84	39

Table 5: Distribution of metaphor types across registers (IND: Indirect, DIR: Direct, IMPL: Implicit, MF: Metaphor Flag)

The distribution varies across registers: the fiction register shows the greatest diversity of metaphor types (indirect, 79.9%; direct, 19.8%; implicit, 0.2%; 55 metaphor flags*), but the academic register is dominated by indirect metaphors (99.5%). Additionally, the social media subcorpus shows a great variety, including the most DFMA tags. This variability is due to the nature of social media texts, particularly tweets, which are often short and have a limited context, which makes it difficult to determine whether a lexical unit is metaphorical.

5.3 MRWs across Major Word Classes

The distribution of metaphor-related words (MRWs) across major lexical categories is detailed in Table 6. Results indicate that metaphor-related words not only appear in expected categories, such as nouns, verbs, and adjectives, but also in those typically regarded as literal, such as prepositions and determiners. Determiners, mostly demonstrative adjectives, and pronouns (e.g., *in* 'this', *ān* 'that', *hamin* 'this', *hamān* 'that'), and prepositions (e.g., *az* 'from', *dar* 'in', *bā* 'with', *be* 'to') show high metaphorical proportions, 61.3% and 42.5% respectively. Nouns and adjectives also contribute significantly, accounting for a large proportion of MRWs. In contrast, verbs, adverbs, and pronouns display lower proportions. Determiners and prepositions are generally among high proportions in metaphor corpora (e.g., Steen et al., 2010; 13-35% of prepositions and 19-31% of determiners), but PerMet 1.0 displays a significant language-specific difference.

Several reasons may explain these differences. First, there is no universal protocol for demarcating lexical units, and each researcher must adapt MIPVU to the target language. Such adaptations can lead to variation across corpora. For instance, compound verbs, and verbal idioms were demarcated as single lexical units, and only the first token received the LU tag, which is typically a non-verbal element. This methodological decision lowered the proportion of verbs while raising that of prepositions, adjectives, and nouns. Second, some differences occur because of the annotation decisions. Similar to English, Persian demonstratives can function as MRWs when there is a cross-domain mapping from 'the one that someone is looking at' to 'referring to an abstract concept in discourse'. However, these forms may also serve as definitive markers, which may be considered as non-MRWs and make it hard to distinguish different functions. Addressing this issue has been deferred to future versions of PerMet.

Lexical Category	LUs	MRWs	%MRWs
Prepositions	11171	4744	42.5
Nouns	37181	3916	10.5
Determiners	2480	1521	61.3
Adjectives	9843	1373	13.9
Verbs	13811	725	5.2
Pronouns	5656	459	8.1
Adverbs	3317	141	4.3
Others*	15664	149	1.0

Table 6: Distribution of metaphor-related words (MRWs) by lexical category
* proper nouns, numerals, conjunctions, interjections, and the object marker *rā*

6. Conclusions

This study introduced PerMet 1.0, a large-scale, register-diverse Persian metaphor corpus, based on MIPVU protocols. The data was collected through various ways and labeled by five independent annotators in multiple phases. The annotation process was elaborated by conducting multiple tests, rounds of review, and group discussions. The results of inter-annotator reliability ($\kappa = 0.952$) showed an almost perfect agreement between the annotators, indicating that MIPVU can be applied to Persian data with some modifications, confirming the findings of Assi et al. (2022), Miri (2024), and Bakhtiyari and Iravani (2025).

The analyses demonstrated that 13.1% of all lexical units are related to metaphor. The distribution varied across registers, metaphor types, and lexical categories. The academic texts had the most MRWs (19.7%), while the fiction texts had the least (7.4%). In terms of metaphor type, indirect metaphors were the most frequent type by far, followed by direct and implicit metaphors. Across the lexical categories, prepositions, nouns, and determiners exhibited the highest frequency of MRWs, with prepositions and determiners showing a high proportion in relation to metaphor.

This study, however, encountered specific limitations. The fiction subcorpus was not well represented, because most publishers did not

grant permission to use their books. Only two publishers allowed partial use of their books. This problem may explain the significant difference in distribution of MRWs. Another limitation concerns the lack of freely available, representative spoken data. The HamBam corpus mostly contains diaries, interviews, and lectures, but few real conversations. Thus, the difference in the proportion of MRWs might be due to this limitation. Finally, a frequently noted limitation in other studies (e.g., Lu and Wang, 2017; Nacey, Dorst, et al., 2019) pertained to difficulties associated with dictionaries. Annotators sometimes had to put aside the corpus-based dictionary (ALPD) and look up in the comprehensive and historical one (SCD). Moreover, many times the sense description was vague or conflated (e.g., for prepositions).

As the first version of PerMet, this corpus is, metaphorically speaking, only at the beginning of its JOURNEY! Future work should expand PerMet to include more data, particularly spoken and fiction, to make it more balanced. Next versions could also enrich the corpus with additional morphological, syntactic, and semantic features to enable more in-depth metaphor analysis in Persian. Given that the results align with findings from other metaphor corpora in different languages, PerMet holds a potential for cross-linguistic comparative analyses. Moreover, it offers a foundation to develop and fine-tune machine learning and large language models for automatic metaphor identification in Persian.

7. Ethical Considerations

This study followed general ethical principles for corpus linguistics, particularly in data collection. Data for the academic, social media, and spoken subcorpora were obtained from freely available corpora, as described in Section 3.1. The fiction subcorpus data was collected with the explicit permission from the respective publishers. Moreover, no personal information about speakers has been disclosed in the spoken subcorpus.

8. Corpus Availability

The PerMet 1.0 corpus is available at https://github.com/ms-miri/PerMet and is released under the Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0) license, allowing researchers to use and share the corpus for non-commercial purposes with appropriate attribution.

9. Acknowledgments

I would like to thank Samansa Mahoozi, Mahdi Khorram, Marjan Akbari, and Atefe Sadat Paliz for their valuable work in annotating the corpus. I am thankful to Dr. Azadeh Mirzaei for her insightful comments and guidance and to Abolfazl Alamdar for his helpful feedback. I also would like to thank the publishers Markaz and Qoqnoos for granting permission to use their books in this study. This work also benefited from the insightful comments provided by three anonymous reviewers.

10. Bibliographical References

Abdi Khojasteh, H., Ansari, E., & Bohlouli, M. (2020). LSCP: Enhanced Large Scale Colloquial Persian Language Understanding. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6323–6327). European Language Resources Association. https://aclanthology.org/2020.lrec-1.776/

Antloga, Š. (2020). *Metaphor corpus KOMET 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1293

Anvari, H. (2002). *Sokhan Comprehensive Dictionary* (Vols. 1–8). Sokhan.

Assi, S. M. (2019). *Advanced Learner's Persian Dictionary*. Samt.

Assi, S. M., Farzane Bakhtiari, Arsalan Golfam, & Shahin Nematzade. (2022). The application of MIPVU procedure for identifying metaphors in Persian: Teacher-student discourse in 1st and 2nd grade of primary schools. *ZABANPAZHUHI (Journal of Language Research)*, *14*(42), 173–201. https://doi.org/10.22051/jlr.2021.34991.1992

Badryzlova, Y., Shekhtman, N., Isaeva, Y., & Kerimov, R. (2013). Annotating a Russian corpus of conceptual metaphor: A bottom-up approach. *Proceedings of the First Workshop on Metaphor in NLP*, 77–86.

Bakhtiyari, F., & Iravani, A. (2025). Towards the First Persian Metaphor Annotated Corpus. In M. Vasheghani Farahani & Z. Ghane (Eds.), *New Frontiers in Corpus Based Studies of Persian: Challenges, Innovations and Applications* (pp. 3–24). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98989-6_1

Black, M. (1955). Metaphor. *Proceedings of the Aristotelian Society, New Series*, *5*, 273–294.

Bort-Mir, L. (2019). *Developing, applying and testing FILMIP: The filmic metaphor identification procedure*. https://doi.org/10.13140/RG.2.2.18345.03688

Bowdle, B. F., & Gentner, D. (2005). The Career of Metaphor. *Psychological Review*, *112*(1), 193–216. https://doi.org/10.1037/0033-295X.112.1.193

Cameron, L. J., & Low, G. (Eds.). (1999). *Researching and Applying Metaphor*.

Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9781139524704

Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Palgrave Macmillan UK.

Choi, M., Lee, S., Choi, E.-K., Park, H., Lee, J., Lee, D., & Lee, J. (2021). *MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories*. https://doi.org/10.18653/V1/2021.NAACL-MAIN.141

Crisp, P., Heywood, J., & Steen, G. (2002). Metaphor identification and analysis, classification and quantification. *Language and Literature: International Journal of Stylistics*, *11*(1), 55–69. https://doi.org/10.1177/096394700201100105

Crystal, D. (2008). *A dictionary of linguistics and phonetics* (6. ed). Blackwell.

Dabir-Moghaddam, M. (1997). Compound verbs in Persian. *Studies in the Linguistic Sciences*, *27*(2), 25–59.

Deignan, A. (2005). *Metaphor and corpus linguistics*. J. Benjamins Pub.

Ge, M., Mao, R., & Cambria, E. (2023). A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, *56*, 1829–1895. https://doi.org/10.1007/s10462-023-10564-7

Haig, G., & Rasekh-Mahand, M. (2025). *HamBam—The Hamedan-Bamberg Corpus of Contemporary Spoken Persian* (Version 3.0) [Dataset]. Otto-Friedrich-Universität Bamberg. https://doi.org/10.48564/UNIBAFD-V80BG-H0243

Kamyabi Gol, A., Akhlaghi Baghujeri, E., Asgarian, E., & Habibi, H. (2018). Extracting information from language corpus: Introducing the corpus of scientific articles of Ferdowsi University of Mashhad. *Library and Information Sciences*, *21*(2), 3–25. https://doi.org/10.30481/lis.2018.61800

Krennmayr, T., & Steen, G. (2017). VU Amsterdam Metaphor Corpus. In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 1053–1071). Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2_39

Lakoff, G., Espenson, J., & Schwartz, A. (1991). *Master Metaphor List* (Tech. No. 2). University of California at Berkeley. https://web.archive.org/web/20180417173025/araw.mede.uic.edu/~alansz/metaphor/METAPHORLIST.pdf.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lu, X., & Wang, B. P.-Y. (2017). Towards a metaphor-annotated corpus of Mandarin Chinese. *Language Resources and Evaluation*, *51*(3), 663–694. https://doi.org/10.1007/s10579-017-9392-9

Maroufi, A. (1989). *Samfoniy-e Mordegān [Symphony of the Dead]*. Qoqnoos.

Maroufi, A. (2003). *Sāl-e Balvā [Year of Turmoil]*. Qoqnoos.

Miri, M. S. (2024). Metaphor Identification in Persian: Annotation, Data Analysis, and Reliability Assessment for Compiling a Metaphor Corpus for Persian. *Language and Linguistics*, *19*(38), 181–208. https://doi.org/10.30465/lsi.2024.47498.1730

Mirzaei, A. (2015). Every Verb, An Eventual Reality: Persian Verb System. *Dastoor*, *11*, 57–92.

Nacey, S., Dorst, A. G., Krennmayr, T., & Reijnierse, W. G. (Eds.). (2019). *Metaphor Identification in Multiple Languages: MIPVU around the world* (Vol. 22). John Benjamins Publishing Company. http://www.jbe-platform.com/content/books/9789027261755

Nacey, S., Krennmayr, T., Dorst, A. G., & Reijnierse, W. G. (2019). What the MIPVU protocol doesn't tell you (even though it mostly does). In S. Nacey, A. G. Dorst, T. Krennmayr, & W. G. Reijnierse (Eds.), *Converging Evidence in Language and Communication Research* (Vol. 22, pp. 42–67). John Benjamins Publishing Company. https://doi.org/10.1075/celcr.22.03nac

Ortony, A. (1993). Metaphor, language, and thought. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed., pp. 1–16). Cambridge University Press; Cambridge Core. https://www.cambridge.org/core/books/metaphor-and-thought/metaphor-language-and-thought/9B241278803ED88BC07EB1920E1319DB

Pirzad, Z. (2001). *Čerāq-hā rā Man Xāmuš Mikonam [I Will Turn Off the Lights]*. Markaz.

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

Ptiček, M. (2025). *Contemporary approaches to natural language processing—Metaphor identification in Croatian language* [Doctoral Thesis, University of Zagreb]. https://urn.nsk.hr/urn:nbn:hr:211:648471

Ptiček, M., & Dobša, J. (2023). Methods of Annotating and Identifying Metaphors in the

Field of Natural Language Processing. *Future Internet*, *15*, 201. https://doi.org/10.3390/fi15060201

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Rasooli, M. S., Kouhestani, M., & Moloodi, A. (2013). Development of a Persian Syntactic Dependency Treebank. In L. Vanderwende, H. Daumé III, & K. Kirchhoff (Eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 306–314). Association for Computational Linguistics. https://aclanthology.org/N13-1031/

Samvelian, P., & Faghiri, P. (2013). Re-thinking Compositionality in Persian Complex Predicates. *Annual Meeting of the Berkeley Linguistics Society*, *39*(1), 212. https://doi.org/10.3765/bls.v39i1.3882

Savary, A., Ben Khelil, C., Ramisch, C., Giouli, V., Barbu Mititelu, V., Hadj Mohamed, N., Krstev, C., Liebeskind, C., Xu, H., Stymne, S., Güngör, T., Pickard, T., Guillaume, B., Bejček, E., Bhatia, A., Candito, M., Gantar, P., Iñurrieta, U., Gatt, A., … Walsh, A. (2023). PARSEME corpus release 1.3. In A. Bhatia, K. Evang, M. Garcia, V. Giouli, L. Han, & S. Taslimipoor (Eds.), *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)* (pp. 24–35). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.mwe-1.6

Semino, E., Demjén, Z., Hardie, A., Payne, S., & Rayson, P. (2017). *Metaphor, Cancer and the End of Life: A Corpus-Based Study* (1st ed.). Routledge. https://doi.org/10.4324/9781315629834

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., & Bagherbeigi, S. (2010, January 1). Semi Automatic Development Of FarsNet: The Persian Wordnet. *Proceedings of 5th Global WordNet Conference (GWA2010)*.

Šorm, E., & Steen, G. J. (2018). Chapter 3. VISMIP: Towards a method for visual metaphor identification. In G. J. Steen (Ed.), *Visual metaphor: Structure and process* (pp. 47–88). John Benjamins Publishing Company. https://doi.org/10.1075/celcr.18.03sor

Steen, G. J. (1999). From linguistic to conceptual metaphor in five steps. In R. W. Gibbs & G. J. Steen (Eds.), *Current Issues in Linguistic Theory* (Vol. 175, p. 57). John Benjamins Publishing Company. https://doi.org/10.1075/cilt.175.05ste

Steen, G. J. (2002). Towards a procedure for metaphor identification. *Language and Literature: International Journal of Stylistics*, *11*(1), 17–33. https://doi.org/10.1177/096394700201100103

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., & Krennmayr, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins Pub. Co.

Stefanowitsch, A. (2007). Corpus-Based Approaches to Metaphor and Metonymy. In A. Stefanowitsch & S. Th. Gries (Eds.), *Corpus-Based Approaches to Metaphor and Metonymy* (pp. 1–16). De Gruyter Mouton. https://doi.org/10.1515/9783110199895.1

The Cambridge Dictionary. (n. d.). RISE. In *The Cambridge Dictionary*. The Cambridge University Press. https://dictionary.cambridge.org/dictionary/english/rise

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C. (2014). Metaphor Detection with Cross-Lingual Model Transfer. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 248–258. https://doi.org/10.3115/v1/P14-1024

Vafi, F. (2005). *Royāy-e Tabbat [Dream of Tibet]*. Markaz.