Metaphor Annotation Criteria: A Step towards Building a Metaphor Corpus

Mohammad Saeid Miri[1]

INTRODUCTION

Due to the development of smart devices, the ability of computers to understand human language has become a key issue in technology. By learning and analyzing machine-readable, annotated linguistic data, computers are able to comprehend human language (corpus). Corpora plays a crucial role in helping computers comprehend human language. Metaphor is one of the most complicated linguistic data that computers cannot comprehend. Despite the prevalence of metaphor in everyday language use and the importance of identifying it, no corpus has been published for Persian yet. Compiling a corpus of Persian metaphors is the initial step in learning metaphors for computers.

To compile a corpus of Persian metaphors, it is necessary to meet two main criteria. Deciding the best definition is the first prerequisite. The best definition is both comprehensive and applicable; comprehensive in the sense that it covers a significant proportion of metaphorical instances, and applicable in the sense that it could be used to build a corpus. The second requirement is the construction of a method for metaphor identification. Without a straightforward data annotation method, it is impossible to identify metaphors.

MATERIALS AND METHODS

Various definitions and theories of metaphor exist in the academic literature (Black, 1993; Fauconnier & Turner, 2002; Gibbs, 1999; Glucksberg & Keysar, 1990; Lakoff & Johnson, 1980; Ortony, 1993). In addition to theoretical endeavors, the literature on operationalizing metaphor identification is expanding (Cameron, 1999, 2003; Deignan, 1999; Low, 1999; Steen, 1999). The Pragglejaz Group (2007) introduced the first serious method for identifying 'linguistic' (not 'conceptual') metaphor: the Metaphor Identification Procedure (MIP). Although MIP is an explicit, step-by-step procedure, feedback from numerous studies suggest that disagreement among experts, average reliability, and the exclusion of other metaphor examples (inadequate validity) prompted Steen et al. (2010b) to introduce MIPVU, a revised method for metaphor identification. The following is the general guideline (Steen et al., 2010b: 23-24):

---

[1] Ph.D. Student in General Linguistics, Allameh Tabataba'i University, Tehran, Iran; ms_miri@outlook.com

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 from 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 coordination, 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 by the author, examine the distinct words which are its independent parts according to steps 2 through 5.

Given that the main objective of this paper is to operationalize metaphor identification, MIPVU is the best, most 'comprehensive' and 'applicable' method. MIPVU's guidelines are provided in Steen et al (2010b). This paper assesses the procedure to determine whether it will be accepted in Persian due to its language-specific characteristics.

RESULTS AND DISCUSSION
While the MIPVU is explicit, detailed, and (in some cases) adaptable, it is not without flaws. Steen et al. (2010b) welcome new versions of MIPVU (in multiple languages) in order to identify its shortcomings. This section will examine one of the difficulties associated with identifying metaphors: the demarcation of lexical units.
An essential element of the MIPVU is the unit of analysis. Steen et al. (2010b: 27) called a word a lexical unit "for theoretical reasons." Even in the English version of the MIPVU, there are exceptions (such as phrasal verbs) for which Steen et al. (2010b) provided guidelines. The most prevalent issue in the other variants of MIPVU is lexical unit demarcation (Herrmann et al., 2019; Nacey et al., 2019b; Pasma, 2019). In Persian, polywords, compound verbs, and compound nouns pose the greatest difficulty in demarcating lexical units. This paper suggests that it may be possible to define three labels: 'cv' (for compound verbs), 'p' (for polywords), and a numbered 'extra element'[1] with a numerical attribute. We can solve the demarcation issue by labeling the 'extra element' with 'cv' (or 'p') and assigning it a number.

CONCLUSION
This paper aimed to introduce a method for identifying Persian metaphors based on MIPVU. The most significant finding of this study is the validation of the proposed method for identifying Persian metaphors. This procedure is straightforward, takes into account language-specific properties, and allows the researcher to make case-specific decisions. The proposed method was able to pass all of the reliability tests

---

[1] The main idea of metaphor corpus annotation belongs to Nacey, Greve, and Johansson-Falck (2019). However, this paper modifies the tagset because of Persian language characteristics.

($\kappa = 0.964$) and is an effective method for identifying Persian metaphors. Statistical analysis and reliability results will be discussed in another paper.

The ability to quantify the study of metaphors in Persian is yet another accomplishment of this paper. With its various constraints on the demarcation of lexical units and the analysis of their basic and contextual meanings, the proposed method enables the researcher to provide a quantitative, detailed, measurable, and trustworthy analysis.