Information Structure and Thematic Structure of Persian Language
A Corpus-based Analysis

Khadijeh Taheri[1]
Habib Gowhari*[2]

Abstract
Other things being equal, grammatical subject is coded as the theme of the clause. In marked cases, however, other non-subject arguments and prepositional phrases may appear as the theme due to some processes, including pre-posing. Accordingly, the goals of the present study are three-fold. First, it aims at investigating the information status of grammatical subject as the unmarked theme of the clause. Secondly, it intends to find the marked themes in the non-canonical constructions, including inversion, topicalization, passivization, and scrambling. Preposing as a cover terms includes all these processes. Accordingly, this study is set out to examine information status, syntactic category, and grammatical function of preposed constituents (marked themes). The degree of matching between grammatical subject and theme is also investigated. This study is a corpus-based one in which the data are collected from two different corpora: a movie (Separation by Farhadi) and a football report. The theory of information structure by Prince (1992) revised in Birner (2013) is adopted as theoretical model of the study. Based on this theory, information is classified in terms of discourse and hearer. Accordingly, information status of any constituent can be one of these for cases: discourse old-hearer old, discourse old- hearer new, discourse new- hearer old, and discourse new-hearer new. This classification is applied to the collected themes (both marked and unmarked) of both corpora. Halliday (2004) is another theoretical basis of this study. Different types of themes are taken from Halliday(2004) to be analyzed in terms of information status. The focus in this study is on topical themes (not interpersonal or textual themes). The target themes were, then, coded based on the theoretical model and they were analyzed by SPSS (version 23). In terms of syntactic markedness, the results indicated that in football report, there were more numbers of marked

---

[1] PhD Candidate of General Linguistics, Department of Linguistics, Ilam Branch, Islamic Azad University, Ilam, Iran; khtaheri@ilam-iau.ac.ir
[2] Assistant Professor, Department of Linguistics, Ilam Branch, Islamic Azad University, Ilam, Iran, (corresponding author); gowhari@ilam-iau.ac.ir

syntactic structures comparing with the film corpus. As for the match between subject and theme, the results indicated a higher amount of match between these two concepts in film corpus comparing with football report corpus. The lower match between subject and theme in football report can be accounted for by the higher number of movements and dislocations in the syntactic category of this type of language. As stated, the main goal of this study was to investigate the information status of both marked and unmarked themes. As for unmarked theme (subject), the results showed that in film corpus such themes are mainly old (more than 90%) both in terms of discourse and hearer. In contrast, the themes were found to be mainly discourse-old but hearer new (more than 65%) in the football report corpus. This is due to the different nature of these two corpora. In daily conversations (film), we have topic continuity, while in football the sentences are expressed to refer to different individuals (no topic continuity). As stated, preposed (marked themes) are also investigated. As for information status, there is almost the same pattern as unmarked themes. Most of the marked themes in film corpus are old both in terms of discourse and hearer. In football report corpus, however, the marked themes (preposed themes) are mainly discourse- old and hearer new. This is somehow problematic, since the discourse old entities cannot be hearer-new. Prince (1992) resolved this problem by raising the issue of schemata and set relations. She believed that many referents are implied in the discourse and hence counted as being discourse-old. She called them inferable. In football report corpus, there were many such inferable marked themes. Another unexpected finding was the high amount of new preposed themes in film corpus which seems to contradict the previous literature (Birner, (1996), for example). Careful examination of literature does not ban the occurrence of new pre-posed items in theme position, although themes naturally have identifiable referents. 'recency of new-ness/ old-ness proposed by Birner (1996) can account for the emerged pattern in the present study. The results of this study are important both theoretically and practically. Many findings showed that the constraints on preposing are not observed in the two investigated corpora. It implies that Persian language, due to its flexible word order, does not follow the rules of other languages. It seems there is a connection between language type and the kind of the constraints in that language. This can be a suggestion for further study to find the appropriate constraints governing information status of themes in Persian. Another implication of this study is the fact that in the same language, we do not have the same patterns as far as the information status of themes is concerned. As it was found, the information status of themes in film corpus was dramatically different from football report corpus. Therefore, one cannot generalize the results obtained in a study to the whole language.