Journal of Language Research
ZABANPAZHUHI
Quarterly Scientific Journal of Language Research, Vol. 16, No.52, Autumn 2024
Alzahra University, http://jlr.alzahra.ac.ir
Article Type: Research
pp.219-246

Building a specialized comparable corpus: PARSA

Elham Alayiaboozar[1]
Aliasghar Hojjatpanah[2]

Received: 2023/09/17
Accepted: 2023/12/02

1. Introduction

Based on the language used in their constituent texts, corpora are categorized as monolingual, bilingual, or multilingual. A comparable corpus is a bilingual or multilingual corpus that includes similar texts in the same subject areas. In other words, a comparable corpus is a collection of documents in two different languages that cover similar topics. Comparable corpora can be composed of general texts, providing various possibilities for discourse analysis, pragmatics, analysis of text genres, and sociolinguistics. Examples of such corpora could include collections of encyclopedia entries, or literary texts from a certain period of time. However, the most common types of comparable corpora, which attract many audiences are those related to specialized fields and containing a high density of vocabulary and technical terms. Such a corpus is called a specialized comparable corpus. In this study, a specialized comparable corpus was built from the Persian and English abstracts of theses and dissertations registered in IranDoc. The corpus is named PARSA.

2. Materials and methods

The process of building PARSA involved several steps: sampling, preprocessing (normalization and tokenization), POS tagging, and checking Persian tags. In the sampling process, 295000 Persian abstracts and their English versions were selected, covering texts from different fields of science (humanities, social sciences, art, engineering, etc.). Preprocessing consisted of two main processes: *normalization* and *tokenization*. Before annotation, it was essential to ensure the homogeneity of the text units, a process called *normalization.* Normalization converts the text into a machine-readable format. In this process, some parts were automatically normalized using the "HAZM" tool. For Unicode homogeneity, especially for characters with different Unicode representations in Persian and Arabic (<ی>, <ک>, and *Ezafe*

[1] Assistant Professor, Iranian Research Institute for Information Science and Technology (IranDoc) (corresponding author); alayi@irandoc.ac.ir
[2] Iranian Research Institute for Information Science and Technology (IranDoc); hojjatpanah@irandoc.ac.ir

construction over <ۀ> /je/), TSQL was used in the SQL Server database to replace Arabic Unicode characters with their Persian counterparts.

Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either word, characters, or subwords. For space homogeneity in the *Tokenization* process, the morphological information (regarding Persian inflectional and derivational affixes) in Sadeghi (1991-1993), Keshani (1992), Lazar (2010) and Ghatre (2007) was used. The output of using "HAZM" tool and Persian morphological information was checked; lots of words were not separated and couldn`t be read. So, the output file was checked manually to correct the wrong spell and separate words manually or using search and replace tools. An example:

1) مساحتپهنههایمتفاوتخطراتموردمطالعهمشخصگردید

After checking:

2) مساحت پهنه‌های متفاوت خطرات مورد مطالعه مشخص گردید

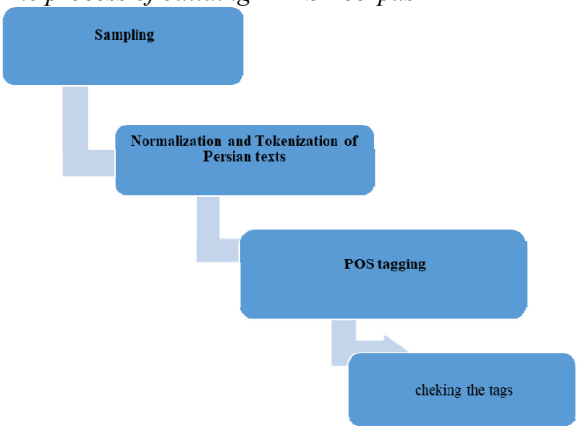
For POS tagging of Persian texts, HAZM tool was used and the tags were controlled. For English texts, NLTK was used.

3. Conclusion

In this study, a specialized comparable corpus (PARSA) was built, which includes Persian and English abstracts of theses and dissertations registered in IranDoc. The process of building this corpus is schematically represented in Figure (1).

Figure 1

The process of building PARSA corpus



Keywords: specialized corpus, comparable corpus, normalization, tokenization, tagging