International Journal of Information Science and Management
Vol. 20, No. 4, 2022, 231-243
https://dorl.net/dor/20.1001.1.20088302.2022.20.4.14.3

Original Research

Steps for Creating Two Persian Specialized Corpora

Elham Alayiaboozar
Assistant Prof., Iranian Research Institute for Information Science and Technology (IranDoc). Tehran, Iran.
Corresponding Author: alayi@irandoc.ac.ir
ORCID iD: http://orcid.org/0000-0002-1584-8085

Ali Asghar Hojjatpanah
MS. In Computer Software, Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran.
hojjatpanah@irandoc.ac.ir
ORCID iD: https://orcid.org/0000-0003-4851-7183

Received: 29 January 2022
Accepted: 07 March 2022

Abstract

Currently, most linguistic studies benefit from valid linguistic data available at corpora. Compiling corpora is a common practice in linguistic research. The present study introduces two specialized corpora in Persian; a specialized corpus is used to study a particular type of language or language variety. For building such corpora, first, a set of texts were compiled based on pre-established criteria used in the sampling process (including the mode of the texts, type of the texts, domain of the texts, language/ language varieties of the texts and the date of the texts). The corpora are specialized because they include technical terms in information processing and management, librarianship, linguistics, computational linguistics, thesaurus building, managing, policy-making, natural language processing, information technology, information retrieval, ontology and other related interdisciplinary domains. After compiling data and Metadata, the texts were preprocessed (normalized and tokenized) and annotated (automated POS tagging); finally, the tags were manually checked. Each corpus includes more than four million words. Since not many specialized corpora are built in Persian, such corpora could be considered valuable resources for researchers interested in studying linguistic variations in Persian interdisciplinary texts.

Keywords: Persian Corpus, Specialized Corpus, Building Corpora, Text Preprocessing, Corpus Annotation.

Introduction

We live in an information age, surrounded by a world full of information. In such an age, systematic data collection is considered a fundamental task. One of the systematic data collection methods is building corpora. The word "corpus", is derived from the Latin word meaning "body", may also be used to refer to any text in written or spoken form. In modern linguistics, corpus, plural corpora, is defined as a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language (Crystal, 1991). A corpus is defined in terms of both its form and its purpose. Linguists have always used the word corpus to describe a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study. More recently, the word has been reserved for collections of texts (or parts of

text) that are stored and accessed electronically. Because computers can hold and process large amounts of information, electronic corpora are usually larger than the small, paper-based collections previously used to study aspects of language (Hunston, 2002). McEnery and Wilson (2001) also define corpus as big electronic texts that are annotated and classified. Corpus linguistics is a field of computational linguistics that deals with principles and the method of using corpus in linguistic studies. Corpus linguistics studies aspects of language based on naturally occurring examples of language in context. Since the 1960s, scholars have tried to collect many valid linguistic texts via computers; Brown corpus is a pioneer in that regard, including Standard English texts. In the early 1980s, there were few electronic corpora, but in the 1990s, the number of projects for building corpora increased; nowadays, information technology and computer science, as well as availability of huge source of texts, have made it possible for scholars to make different kinds of corpora. Some linguists interested in theoretical subjects do not believe in the validity of corpus and are skeptical that it can be representative of natural language. They think that many sentences in natural language are absent in corpora, and sometimes corpora contain wrong or impolite sentences, which should not be included in corpora; nevertheless, it should be kept in mind that there is no corpus which haves all aspects of language including phonology, morphology, syntax, semantics and the like; however, it should be pointed out that any corpus which is studied, even a very small one, include facts about language, which we have not faced anywhere; so, any corpus has something to learn.

Different fields of linguistic research benefit from corpora, including natural language processing, speech understanding, artificial intelligence, text-to-speech conversion and automatic speech recognition, dictionary and encyclopedia compiling, building linguistic databases, collocations in different languages, machine translation, language learning, dialectology, semantics, discourse analysis, sociolinguistics, forensic linguistics, studying literary genre and so on so forth. Corpora are of different kinds. Atkins, Clear & Ostler (1992) have studied corpora from various aspects and classified them based on the content, language, date and other criteria. Corpora can be classified as monolingual (it is the most frequent type of corpus. It contains texts in one language only), parallel (it consists of the exact text translated into one or more languages. The texts are aligned. Such corpus allows searches in one or both languages to look up or compare translations), comparable (it is a corpus consisting of texts from the same domain in more languages. In contrast to a parallel corpus, the texts are not translations of each other and belong to the same domain with the same Metadata. An example of a comparable corpus is a corpus made from Wikipedia, diachronic (it is a corpus containing texts from different periods and is used to study the development or change in language), synchronic (texts in a synchronic corpus come from the same point of time. It is a snapshot of language in one moment), static (a static corpus is a corpus whose development is complete. The content of such corpus does not change), learner (it contains a collection of texts produced by learners of a language used to study errors and mistakes made by learners of languages), specialized (a specialized corpus contains texts limited to one or more subject areas, domains, topics etc. Such corpus is used to study how the specialized language is used. They may contain a sample of this type of text or, if the dataset is finite and of a manageable size (for example, all of Jane Austen`s novels) be complete), monitor (a monitor corpus is used to monitor the language change. It is a corpus which is regularly (or even continuously) updated, new texts are added as they are produced), general (it consists of general texts, i.e., texts do not belong to a single text type, subject field, or register. An example of a general corpus is the British National

Corpus); other kinds of corpora may include: whole text, samples, open, closed, single, central, cluster, nuclear and perimeter.

Many general corpora are built in different languages. Although there are many specialized corpora in English (including Air Traffic Control, ATC), British Academic Spoken English (BASE), British Academic Written English (BAWE), Business Letters Corpus, Translational English Corpus (TEC), Characterizing Individual Speakers (CHAINS), Corpus of Spoken Professional American-English (CSPA), Corpus of Written British Creole (CWBC), and many others (Weisser, 2016), there are not so many specialized corpora in many other languages including Persian; therefore, this study aimed at building two specialized corpora from interdisciplinary sources, one from the published articles of the Iranian journal of information processing and management (called *Pazhuheshname*) and the other one from digital books published at *Iranian Research Institute for Information Science and Technology (IranDoc)* (called *PEKA*); both are considered as valuable references for scholars whose main concern is working on valid interdisciplinary sources of language. Considering the language, both are monolingual (Persian).

Although prior identification of target users is a prerequisite in corpus building, it does not mean there is no overlap among the target users regarding the utilization of a corpus. Nobody imagined that famous corpora, such as Brown Corpus, would ever be used as precious resources in many domains of language research, including English language teaching or cultural studies. At the time of creating the two specialized corpora, *Pazhuheshname* and *PEKA*, the general assumption was that the proposed corpora could be used in various linguistics works. The main application of these corpora has been visualized in natural language processing, information retrieval, language technology, machine translation, database compilation, sociolinguistics, syntax analysis, lexicology, grammar writing, and other related domains.

Some speech and text specialized monolingual and bilingual corpora (Persian-English) have been generated in Persian, including Persian Emotional Speech (Keshtiari, Kuhlmann, Eslami & Klann-Delius, 2015), ICT English-Persian comparable textual corpus (Dashtbani, Mansoorizade & Mohammad, 2015), Bilingual Plagiarism Detection Corpus (Asghari, Khoshnavataher, Fatemi & Faili, 2015), ShEMO: a Large-scale Validated Database for Persian Speech Emotion Detection (Mohamad Nezami, Jamshid Lou & Karami, 2019), and many others; however, for studying technical terms in different fields of science as well as jargon studies, there is a need for extracting such terms and their behavior in linguistic contexts from specialized corpora, whose content is texts from special fields of science. Since *Pazhuheshname* and *PEKA* are building out of texts in interdisciplinary fields with associated technical terms and jargon, they are considered valuable resources in linguistic studies such as studying the language of science and the like.

The architecture of building these two specialized corpora is presented in a block diagram (Figure 1).

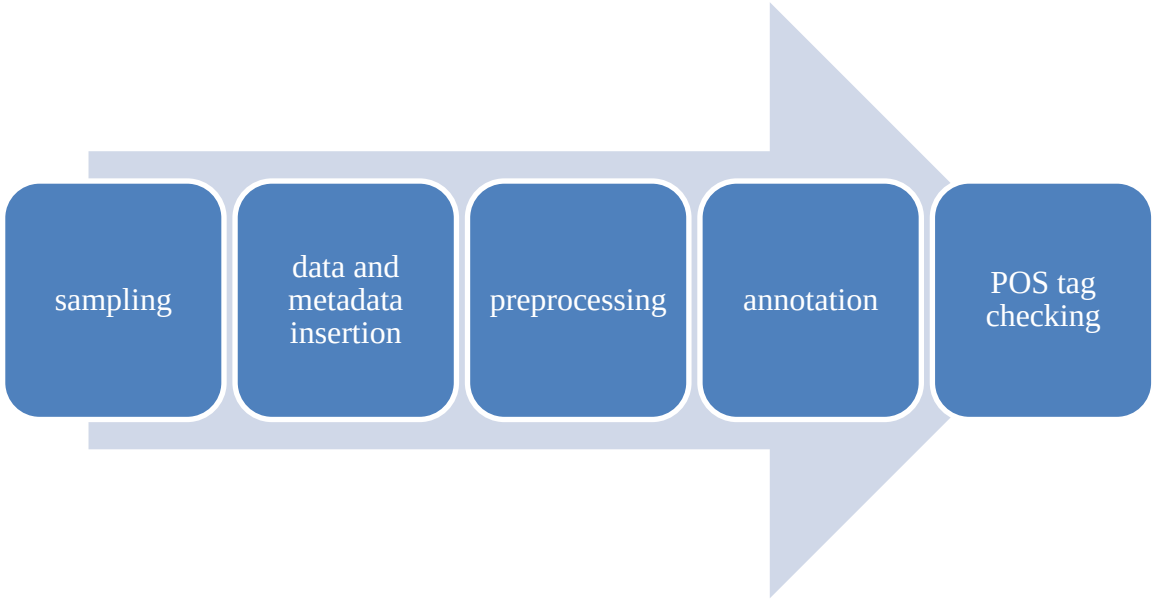

Figure 1: block diagram of corpus building

In the following parts of the article, after a review of related literature, the method of compiling data for each corpora, the process of preprocessing, and annotation are presented.

Literature Review

Given the characteristics of building a corpus, Waynne (2005) has reviewed different renowned scholars' viewpoints about steps to be followed in building a corpus. Brezina, Hawtin & McEnery (2020) explain about Written British National Corpus 2014; this corpus is built at the University of Lancaster; it contains 100 million present-day British English words, and actually, this corpus has been constructed as a comparable counterpart of the original British National Corpus, which was compiled in the early 1990s.

Regarding the high efficiency of specialized corpora in many linguistic analyses, which involve working on actual data from the language of science, some scholars have been interested in building specialized corpora, among whom are Davies (2021), Torrida (2016) and Beloso (2015). Davies (2021) explains about Coronavirus corpus, which at March 2021 contains 900 million words; this corpus is considered a sub-corpus of News on the Web (NOW) corpus, which includes 12 billion words. Torrida (2016) presents the steps to be followed in building a specialized corpus; the steps could include sampling, deleting non-content morphemes (functional morphemes), text analysis via AntConc, making a list of frequency of words, POS tagging, adding a definition for words as well as their collocations and the like. Beloso (2015) introduces the Corpus of Architecture Discourse in Contemporary English (CADCE), which contains 500,000 words from written language collected from different sources; this corpus is monolingual, not annotated, and it contains published texts in North American, British, Irish, Canadian and Australian English (2007-2008).

Other scholars have built comparable or parallel corpora: Lapshinova-Koltunski (2013) present a comparable translation corpus created to investigate translation variation phenomena in terms of contrasts between languages, text types and translation methods (machine vs. computer-aided vs. human). There are also some Persian-English comparable and parallel corpora, including those introduced by Karimi, Ansari and Sadeghi (2017), Dashtbani et al.

(2015), Kashefi (2018), Baradaran Hashemi, Shakery & Faili (2010), Mohammadi and Riahi (2016) and Mohammadi (2012). Other Persian corpora may include *Persian written corpus* (Bijankhan, Sheykhzadegan, Bahrani & Ghayoomi, 2011), *Persian Database* (Assi, 2005), *PersPred*, A Syntactic and Semantic Database for Persian Complex Predicates (Samvelian & Faghiri, 2013), *Persian Syntactic Dependency Treebank* (Rasooli, Kouhestani & Moloodi, 2013), and Hamshahri (Aleahmad, Amiri, Darrudi, Rahgozar & Oroumchian, 2009).

Material and Methods

Sampling

In building any kind of corpus, whether is written or speech some general criteria should be taken into account, including sampling, representativeness, balance, size, the type of corpus (general vs. specialized), and homogeneity (Figure 1).

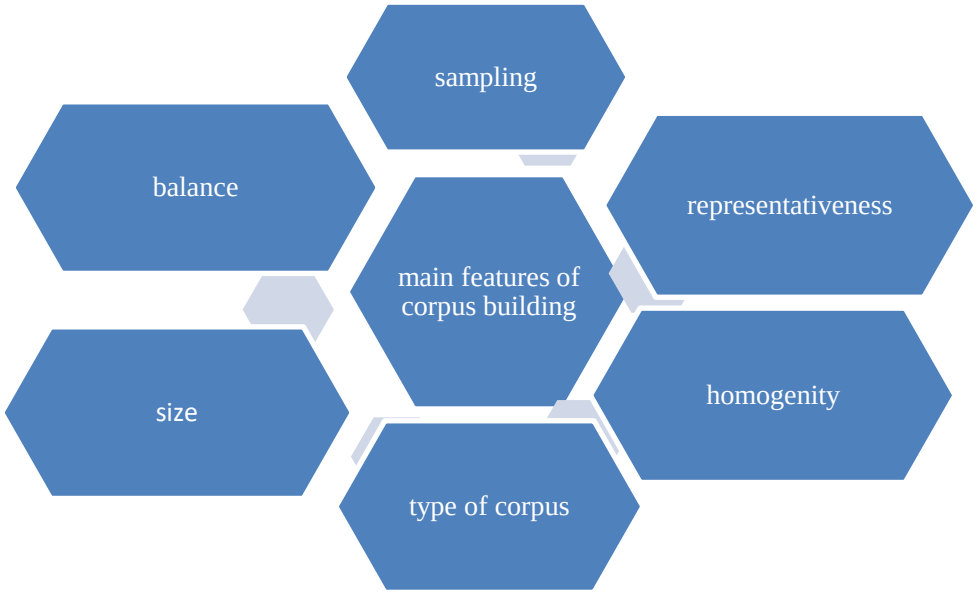

Figure 1: main features of corpus building

Sampling is defined as the process of choosing texts of different genres based on the aim of corpus building. Any selection/sampling must be made on some criteria, and the first major step in corpus building is the determination of the criteria on which the texts that form the corpus will be selected. Standard criteria include 1. The mode of the texts; whether the language originates in speech or writing, or perhaps nowadays in electronic mode; 2. The type of the texts; for example, if written, whether a book, a journal, a notice or a letter; 3. The domain of the texts; for example, whether academic or popular; 4. The language/ language varieties of the corpus; 5. The location of the texts; for example (the English of) the UK or Australia and 6. The date of the texts (Sinclair, 2004).

Representativeness refers to the process of choosing texts to be representative of the language of the corpus. Balance in corpus building refers to linguistic variety or official/unofficial variety used in texts. Size of the corpus is the number of the words. The type of the corpus refers to whether the corpus is general or specialized. The last but not the least point to be considered in corpus building is homogeneity. Most of the aforementioned criteria depend on the first item, i.e., sampling.

As mentioned in the the introduction, two specialized corpora were built in this study (*Pazhuheshname* and *PEKA*). Regarding most of the criteria mentioned in sampling, the mode of the texts is written electronic, the type of the texts is articles and books, the domain of the texts is academic, the language/language variety of the texts is written and formal, and the location of the texts is Iran.

Regarding the sampling frame of corpora, the *Pazhuheshname* corpus contains 1089 texts of articles (1972-2020), and the *PEKA* corpus contains 281 texts of books (1969-2021). The subjects covered in these two specialized corpora include information processing and management, librarianship, linguistics, computational linguistics, thesaurus building, management, policy-making, natural language processing, information technology, information retrieval, ontology and other related interdisciplinary domains. The process of sampling for corpora is as follows:

For *Pazhuheshname*, no specific approach was followed in gathering articles; out of 1089 articles, 600 WORD formats of them were available, and the rest were in PDF format; all 600 articles were considered to be used in the corpus.

Generally, corpora are built to enable scholars to study linguistic items (words, phrases, sentences …) in contexts. For *PEKA*, it should be mentioned that since some of IranDoc digital books contain statistical information, bibliography, thesauri and the like, these books were excluded in the sampling process due to a lack of enough texts to be used in linguistics studies. Another point that was taken into account in this step was that many books were published before the 1990s; since the WORD format of most books was not available, using OCR (Optical Character Recognition) for converting old PDF files to WORD format would not be that accurate for Persian, so, many of the old files, regardless of their subject matter, were excluded as well. Finally, 68 files were chosen to be used in the corpus: 31 files in WORD format and 37 files in PDF format. OCR was used for converting PDF files to WORD format, whose accuracy is more than 90 percent. After PDF conversion to WORD format, the files were manually checked to ensure the accuracy of the conversion process.

It is worth mentioning that at the time of building *Pazhuheshname*, OCR was not used due to some financial and practical considerations.

Besides data gathering, Metadata pertaining to each text file, article and book, was also inserted in corpora. The type of Metadata used in this study is a descriptive, which is the descriptive information about a resource (here, the articles and books used in corpora). It includes elements such as the title of the books and articles, author(s), and the subjects of the articles and books used in corpora. It is worth mentioning that all the titles were checked to make sure not including one document twice; moreover, all the tables, graphs and tables were deleted in the sampling process.

Preprocessing

Text preprocessing is an approach for cleaning and preparing text data for use in a specific context. While working to build a corpus, you might face difficulties, and before processing the corpus, these problems should be removed at a step named *Preprocessing* (Ghayoomi, Momtazi & Bijankhan 2013). Preprocessing may include two main processes: *Normalization* and *Tokenization*. Before annotation, it is essential to check the homogeneity of the text units; this process is called *Normalization*. Normalization process turns the text into a machine-readable one. Normalization for Persian texts may include Unicode homogeneity, correcting wrong

spelling, linking different ways of writing of one word and so on. One of the main problems in Persian text processing is the existence of different character encodings in text documents; for example, the word «کتاب» (book), might have different encodings in different documents that, cause the text processing algorithms to consider them as different words. In the present study for Unicode homogeneity of some characters whose Unicode is different in Persian and Arabic (<ی>, <ک> and *Ezafe construction*[1] over <ۀ>/je/), TSQL was followed in the SQL Server database, and Arabic Unicode of mentioned letters were replaced by the Persian ones (Table 1).

Table 1
Some of the Arabic and Persian Unicode

Arabic letter Unicode	U+1610 for <ی>	U+1603 for <ک>	U+1577 for <ۀ>
Persian letter Unicode	U+1740 for <ی>	U+1705 for <ک>	U+1607 for <ه>

Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either word, characters, or subwords. English text processing differs from Persian; letters/graphemes are written separately, and no letter/grapheme is attached to the other one in the Standard English writing system; though, in the Persian writing system, most affixes are connected to the stem. In the Persian writing system, some morphemes in multi-token words can be entirely separated by white space, linked via ZWNJ[2] (so-called half-space), or attached (using no space). If affixes are connected to the stem using white space, rather than ZWNJ (when needed), in text processing, it will lead to considering a single word as two or more words by machine (depending on the number of affixes attached to the stem); for example, <نرفته‌ام> /narafteʔam/ (I have not gone) is a single word consisting of three morphemes:

*<ن> (negative prefix attached to the verb) + <رفت> /raft/ (went) + <‌ام> /ʔam/ (first person singular inflectional suffix) = <نرفته‌ام>, <نرفته ام>

‌<نرفته‌ام>‌ (ZWNJ between stem and suffix)

-<نرفته ام> (white space between stem and suffix)

In <نرفته‌ام>, if white space is used between the stem and the first person singular suffix, in machine processing, this word is considered as two words (<نرفته> and <ام>), which would lead to allocating two tags (e.g., two POS tags) to a single word (نرفته‌ام).

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. Take the following examples:

In Persian the suffix <ها-> /hɒ/ is a plural morpheme attached to a stem like <کتاب> /ketɒb/ (book), which leads to three forms: *<کتاب> + <ها-> = <کتابها>, <کتاب ها>, <کتاب‌ها>; moreover,

1. Ezāfeh is an unstressed vowel /e/ considered as a morpheme, which is pronounced, but not written in Persian. Its function is making links between some constituents building a noun phrase in which this phenomenon is named "Ezāfeh construction" (Ghayoomi et al: 2013).
2. Zero-Width Non-Joiner(ZWNJ)

some derivational suffixes including <-گیری> /giri/ (take), <-نویس/نویسی> /nevis, nevisi/ (writing), <سنج/سنجی> /sandʒ, sandʒi/ (measure) might attach to stems via white space or ZWNJ, which leads to the following forms:

*<آبمیوه> /ʔɒbmive/ (juice)+ <-گیری> /giri/ (take) = <آبمیوه گیری> or <آبمیوه‌گیری> (juicer)

*<برنامه> /barnɒme/ (program)+ <-نویسی> /nevisi/ (writing) = <برنامه نویسی> or <برنامه‌نویسی> (programming).

Another problem regarding Persian writing system is that as some letters never attach to the next letter (including <و>/v,o,u/, <ژ>/ʒ/, <ز> and <ذ> /z/, <ا>/ɒ/, <ر>/r/ and <د>/d/), if white space is not used properly, it would lead to wrong analysis; for example, <وباهاوآمد> /va bɒ ʔu ʔɒmad/ (and he/she came with him/her) is considered as a single unit/word, rather than four unites/words).

For space homogeneity, employing computer programming derivational and inflectional affixes (including <ها->, <می->, <-گیر/گیری>, <-سنج/سنجی>, <-نویس/نویسی>, <-تر/ترین>, <-ای>, <-کار/کاری>, <-گر/گری>, <-دان/دانی>, <مند> etc) were detected in the texts and if they were not attached to the stem (white space used), the program attaches them to their stem via ZWNJ.

Not only computer programming but also an available Persian tokenizer tool (HAZM tool) was used for space homogeneity in the *Tokenization* process.

Annotation

Corpus annotation, sometimes called ‘tagging’, can be broadly conceptualized as the process of enriching a corpus by adding linguistic and other information inserted by humans or machines (or a combination of them) in service of a theoretical or practical goal (Hovy & Lavid, 2010). A typical and familiar case of corpus annotation is grammatical tagging (also called word-class tagging, part-of-speech tagging and POS tagging). In this case, a label or tag is associated with a word to indicate its grammatical class (Garside, Leech & McEnery, 2013). Annotation adds additional value to the corpus. Some scholars, like Sinclair (2004), prefer not to engage themselves in annotation due to being skeptical that it would be of any help in linguistic analyses. They believe that raw corpus/ unannotated corpus is a pure one, which is more useful in linguistic analyses; however, most researchers believe that a corpus is useful if one can extract knowledge or information from it. Raw corpus contains no direct information in grammar or other related fields, which can hinder many of the applications to which a corpus can be put. Given the example of <left> in English: in opposite to <right>, <left> could be an adjective (my left hand), and adverb (turn left) or a noun (on your left); moreover, <left> could be past or past participle of the verb <leave> (I left early). So, various meanings and uses of this word cannot be detected from its orthographic form; however, if a corpus is grammatically well annotated, a label indicating the word class will accompany each occurrence of <left> (Garside, Leech & McEnery). Many kinds of annotations/tagging include Part-Of-Speech (POS), phonetic, syntactic, semantic, pragmatic, discourse, stylistic and lexical.

Manual annotation is time-consuming and costly; the automated approach is not free of mistakes, but a combination of them seems more reliable because first, some developed tools designed for tagging automatically annotate the corpus; then, the tags are checked by a human. If a well-designed tagger is used for this purpose, it will take less time to check the annotation accuracy.

Among different kinds of annotations, POS tagging is the most famous; POS tagging, also called grammatical tagging or morpho-syntactic annotation, means allocating part of speech

tags (NOUN, ADJECTIVE, ADVERB, VERB and the like) to each word in the corpus. POS tagging has been one of the first widely used corpus annotations, and is today by far the most common type. It is the most basic type of annotation, which is the basis for further linguistic analyses, like parsing and semantic annotation. A POS-tagged corpus is useful for a wide applications ranging from homograph disambiguation to more complicated uses such as computing the occurrences of word classes in a corpus, machine translation and the like. Due to the high application of POS tagging for many linguistic analyses, it was decided to add POS tags to the words in corpora (*Pazhuheshname* and *PEKA*). For POS tagging, the HAZM tool was used. HAZM is a Python tool for Persian text processing, whose accuracy is reported as 97.1 percent for POS tagging (github.com/sobhe/ham). The tag list used in HAZM is the one introduced in Bijankhan et al. (2011)- table 2. It is worth mentioning that to show Ezafe construction in Persian texts, if Ezafe follows a word, HAZM adds <e> to indicate Ezafe construction in Persian. For example, in phrase <قراردادهای رسمی> /gharardad hɒje rasmi/ (official contracts), the tags are as follows:

* <قراردادهای> /Ne/ (Noun+ Ezafe construction) and <رسمی> /Adj/

Table 2
Tag list for Persian texts (Bijankhan et al., 2011)

POS tag	Tag Name
N	Noun
PREP	Preposition
PUNC	Punctuation
AJ	Adjective
V	Verb
CON	Conjunction
NUM	Number
PRO	Pronoun
DET	Determiner
ADV	Adverb
POSTP	Postposition
RES	Residual
CL	Classifier
INT	Interjection

As expected the frequency of some tags, including nouns (N), adjectives (AJ), verbs (V), conjunctions (CONJ), Prepositions (P) and punctuations (PUNC) seems to be high; for example, in *Pazhuheshname* the frequency of mentioned tags are as follows: N=945798, Ne (noun+ Ezafe construction)= 900911, AJ= 308823, AJe (adjective + Ezafe construction)= 107820, V= 347951, CONJ= 294695, P=430372, and PUNC=664338.

After automated POS tagging, the accuracy of tags was checked manually. A semi-supervised process is used in annotation, because first the data was annotated automatically using HAZM; then the tags were checked manually; so, a hybrid of manual and automated methods was used for POS tagging.

Manual checking was done following some steps: first, sample data was chosen to check the accuracy of POS tags; then the tags in samples were reviewed, and the wrong tags were corrected; in addition to correcting the incorrect tags, error pattern analysis was done and it was

understood that most errors were originated from preprocessing, i.e., Tokenization step. This problem was identified in building the first corpus (since the corpora were built one by one), so the morphological information regarding Persian affixes was used to solve the blank space problem: using ZWNJ instead, wherever needed.

Although checking the accuracy of the tags was done manually; still there exist some POS tag errors: consider the following data:

The word <زیادی> (more), which is an adverb in context, is wrongly tagged as an adjective (AJ). <بهینه‌ترین> (the best), which is an adjective, is improperly tagged as a noun (N). <به> (to), which is a preposition, is improperly tagged as an adverb (ADV) and verb /V/. It should be taken to account that in some cases, wrong tags are due to using white space in a word; for example, <بموقع/به‌موقع> (timely), which is an adjective or adverb, is written as <به موقع>, that leads to allocating two tags to this single word (<به> /preposition/ (P), which is wrongly tagged as an adverb and <موقع>, which is tagged as a noun (N)). Extracted annotated data from these two corpora show that most tags are correct, and few wrong tags exist, which could be neglected (fewer than 5 or 6 incorrect tags in every 300 words in the corpus).

Discussion

This study built two specialized corpora (*Pazhuheshname* and *PEKA*). Both are monolingual (Persian) and annotated (POS tagging). Regarding pre-established criteria pertaining to sampling, the mode of the texts is written electronically, the texts are articles and books, the domain of the texts is academic, the language variety of the texts is written and formal, and the location is Iran. The sampling frame of corpora is as follows: *Pazhuheshname* contains 1089 texts, and *PEKA* contains 281 texts. Both are specialized corpora and have texts in interdisciplinary domains. In the sampling process, out of 1089 articles in the Iranian journal of information processing and management, 600 WORD formats were available, all of which were used in *Pazhuheshname*. It is worth mentioning that in the sampling process in the second

corpus, some texts were excluded due to their non-linguistics content as well as the date of their publication, which was before the 1990s; so, 68 files were used in *PEKA*. Besides data gathering, descriptive Metadata on each text file was also inserted in corpora, i.e., the title of the books and articles, author(s), and the subjects of the articles and books. In preprocessing step, the Normalization and tokenization of texts were done before annotation. For POS tagging,, an available tool for Persian texts was used (automated annotation); then, the tags were manually checked.

Although Some monolingual and bilingual corpora have been generated in Persian, including Persian Emotional Speech, Bilingual Plagiarism Detection Corpus, ShEMO: a Large-scale Validated Database for Persian Speech Emotion Detection and many others; studying technical terms in different fields of science as well as jargon studies requires having access to specialized corpora whose content is texts from special fields of science to extract such terms and their linguistic behavior in linguistic contexts. Since Pazhuheshname and PEKA are built out of texts in interdisciplinary fields with associated technical terms and jargon, they are considered valuable resources in linguistic studies such as studying the language of science and the like.

As mentioned, Some speech and text specialized monolingual and bilingual corpora (Persian-English) have been generated in Persian, including Persian Emotional Speech (Keshtiari et al., 2015), ICT English-Persian comparable textual corpus (Dashtbani et al., 2015), Bilingual Plagiarism Detection Corpus (Asghari et al., 2015), ShEMO: a Large-scale Validated Database for Persian Speech Emotion Detection (Mohamad Nezami et al., 2019); most of them are not available to see the linguistic content for comparing with the content of Pazhuheshname and PEKA. In the process of building Pazhuheshname and PEKA we faced a lot of challenges in preprocessing step regarding the changes in words spell after OCR, which we had to correct a long list misspelling prior to POS tagging, but we have no idea whether other scholars who have built similar corpora have faced such challenges or not. The present study is considered as unique in a sense that no one have ever built corpora out of the materials used in present study. Pazhuheshname and PEKA are built out of texts in interdisciplinary fields with associated technical terms and jargons, they are considered as valuable resources in linguistic studies such as studying language of science and the like. Another point regarding the present study is that for POS tagging a hybrid method was used: a combination of automated POS tagging (using HAZM tool) and manual checking of the accuracy of automated tagging.

Conclusion

Finally, it is worth mentioning that due to the low availability of highly specialized corpora to be used in linguistic studies, there is an obvious need to collect and share specialized texts in Persian for building more specialized corpora in Persian, which are considered the basis of scientific language studies. As an extension to this project, it is suggested to add other annotations to corpora, including semantic, discourse, syntactic and other linguistic annotations, which make corpora a valuable resource for all linguistic analysis.

Many machine translation (MT) projects benefit from bilingual corpora; since not enough specialized bilingual corpora (so-called comparable and parallel corpora) in Persian-English exists, it is suggested that building bilingual specialized corpora (Persian-English) in different fields of study could be considered as a priority.

References

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M. & Oroumchian, F. (2009). *Hamshahri: A Standard Persian Text Collection. Knowledge-Based Systems*. 22(5), 382–387. https://doi.org/10.1016/j.knosys.2009.05.002

Asghari, H., Khoshnavataher, K., Fatemi, O. & Faili, H. (2015). Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus. In *The 13th Evaluation Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN15).* Retrieved from http://ceur-ws.org/Vol-1391/148-CR.pdf

Assi, M. (2005). Persian database in Internet *Journal of researchers*, 2. Retrieved from https://hawzah.net/fa/Magazine/View/6280/6282/69701 [in Persian]

Atkins, S., Clear, J. & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*. 7 (1), 1-16. Retrieved from http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf

Baradaran Hashemi, H., Shakery, A. & Faili, H. (2010(. Creating a Persian-English comparable corpus. In *Proceedings of Conference on Multilingual and Multimodal Information Access Evaluation (CLEF)*, (pp. 27-39). Padua, Italy.

Beloso, B. S. (2015). Designing, describing and compiling a corpus of English for architecture. *Procedia-social and behavioral sciences*, 198,.459-464.

Bijankhan, M., J. Sheykhzadegan, M. Bahrani & Ghayoomi, M. (2011). Lesson from building a Persian written corpus: Peykare. *Language resources and evolution* 45 (2), 143-164. https://doi.org/10.1007/s10579-010-9132-x

Brezina, V., Hawtin, A. & McEnery, T. (2020). The written British national corpus 2014-design and comparability. *Journal of text and talk*, 41 (5-6). 595-615. https://doi.org/10.1515/text-2020-0052

Crystal, D. (1991). *A dictionary of linguistics and phonetics*. 3rd edition, Blackwell.

Dashtbani, S., Mansoorizade, M. & Mohammad, N. (2015). ICT English-Persian comparable textual corpus. *Iranian journal of comparative linguistic research*. 4(8), 121-141. [in Persian]

Davies, M. (2021). The Coronavirus corpus: Design, construction and use. *International journal of corpus linguistics*, 26(4), 583-598. https://doi.org/10.1075/ijcl.21044.dav

Garside, R., Leech, G. & McEnery, T. (2013). *Corpus annotation: Linguistic information from computer text corpora.* Routledge.

Ghatre, F. (2007). Inflectional features in contemporary Persian. *Dastoor*. 3, 52-81. [in Persian]

Ghayoomi, M., Momtazi, S. & Bijankhan, M. (2013). A study of corpus development for Persian. *International journal on Asian language processing*, 20 (1), 17-33. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=850B3BDAE5AC3A2DFE7B7E9DDAE6AA9F?doi=10.1.1.388.4367&rep=rep1&type=pdf

Hovy, E. & Lavid, J. (2010). Towards a science of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*. 22(1), 13-36.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

Karimi, H., Ansari, E. & Sadeghi, B. (2017). Extracting an English-Persian parallel corpus from comparable corpora. *Arxiv: 1711.00681v3 [cs.CL]*.

Kashefi, O. (2018). MIZAN: A large Persian-English parallel corpus. *arXiv preprint arXiv:1801.02107*

Keshani, KH. (1992). *Suffix derivation in contermporary Persian*. Markaz-e Nashr-e Daneshgahi. Tehran. [in Persian]

Keshtiari, N., Kuhlmann, M., Eslami, M. & Klann-Delius, G. (2015). Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD). *Behavior Research Methods*, 47(1), 275-294. https://doi.org/10.3758/s13428-014-0467-x

Lapshinova-Koltunski, E. (2013, August). VARTRA: A comparable corpus for analysis of translation variation. In *Proceedings of the sixth workshop on building and using comparable corpora* (pp. 77-86).

Lazar, G. (2010). *Contemporary Persian grammar*. Bahreini, Mahasti (translation). Hermes publication. [in Persian]

Mohammadi, R. (2012). *Building a Persian-English comparable corpus and extracting parallel sentences*. Master thesis. University of Alzahra. [in Persian]

McEnery, T. & A. Wilson. 2001. *Corpus linguistics: An introduction*. Edinburgh University Press.

Mohammadi, S. R. & Riahi, N. (2016(. Presenting an Optimal Method for Constructing an English-Persian Comparable Corpus. International Journal of Intelligent Information Systems. 5 (3): 42-47.

Mohamad Nezami, M.O., Jamshid Lou, P. & Karami, M. (2019) . ShEMO: A large-scale validated database for Persian speech emotion detection. *Language Resources & Evaluation* 53(1),1-16. https://doi.org/10.1007/s10579-018-9427-x

Rasooli, M., Kouhestani, M. & Moloodi, A. (2013). Development of a Persian syntactic dependency treebank. *In The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, (pp. 306-314). Atlanta, USA.

Sadeghi, A. (1991-1993?). Word formation methods in Persian. *Danesh publication*. [in Persian]

Samvelian, P. & Faghiri, P. (2013(. Introducing PersPred, A syntactic and semantic database for persian complex predicates. In *Proceedings of the 9th Workshop on Multiword Expressions*, (pp. 11-20). *Atlanta, Georgia, USA.* Association for Computational Linguistics.

Sinclair, J. (2004). Corpus and Text - Basic Principles. In Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. ahds.literature, languages and linguistics: University of Oxford, UK.

Toriida, M. C. (2016). Steps for creating specialized corpus and developing an annotated frequence-based vocabulary list. *TESL Canada journal/ revue TESL du Canada*, 34 (1), 87-105. https://doi.org/10.18806/tesl.v34i1.1257

Weisser, M. (2016). Specialised Corpora. Retrieved from http://martinweisser.org/corpora_site/spec_corpora.html