

# Introduction to the Corpus of Law, Academic, and News (CLAN)

*Ariana N. Mohammadi*

---

## ABSTRACT

The Corpus of Law, Academic, and News (CLAN) is a collection of Persian text samples representing the pattern and structure of formal written Persian. The corpus is subdivided into three sub-corpora, including legal documents, academic abstracts, and news articles. The CLAN consists of a total of 400 freestanding XML text files. Each file includes an internal DTD declaration and is encoded as UTF-8. The corpus is annotated and can be used in a variety of linguistic research studies such as text analysis and genre analysis.


**Keywords:** Persian corpora; corpus of written Persian; corpus of law, corpus of academic persian, corpus of news

---

## 1. Description

The Persian Corpus of Law, Academic, and News (CLAN) contains formal written texts in Persian that are divided into three sections: legal texts, academic abstracts, and newspaper articles. The corpus represents the pattern and structure of formal written Persian in published materials. The CLAN consists of 400 XML text files that are encoded as UTF-8 and include an internal DTD declaration. All text files are marked with a header and a body. The header markups vary in each sub-corpus based on genre, and the body of each file is tagged with a title (or headline) and body paragraphs.

---

**CONTACT**  [ariana-mohammadi@lcca.ca](mailto:ariana-mohammadi@lcca.ca)

The corpus is divided into three genres, and each genre is sub-divided into proportionally balanced representation of different text types. The legal section contains texts that are extracted from official legal text publications, including the civil penal code, the criminal penal code, and the constitution of the Islamic Republic of Iran. The academic sub-corpus also contains published academic abstracts in different disciplinary areas including Art and Humanities, Social Sciences, and Natural Sciences. Further, the sub-corpus of news articles is extracted from the archive of ten Iranian news outlets within the past ten years (2010-2020).

The typical length of text segments in the corpus ranges from 400 to 5,000 word-tokens, and no single text file contributes more than 5.5% of the data to the corpus. Biber (1993), through statistical analyses, concludes that in English common linguistic features are distributed fairly consistently across text segments of 1,000 words. That is, text samples of 1,000 words or more can reliably represent common linguistic features of English in a corpus. However, this cut-off point is unknown in Persian. Therefore, text samples of diverse lengths have been included in the composition of the CLAN. Further, in order to preserve the integrity of the text samples, the original length of segments has been maintained and text files are not fragmented (Mohammadi, 2018).

The number of text samples in a corpus sub-division that represents a particular genre partially depends on the degree of internal variations within that genre (Biber, 1993). For instance, academic abstracts have a higher degree of internal variations in comparison to legal texts. On the other hand, academic abstracts are naturally shorter than legal documents. As a result, the number of text samples that are included in the legal sub-corpus is considerably fewer than the number of text samples that are included in the academic sub-corpus. The number of text samples in legal sub-section is 48, while the academic sub-section contains 274 text samples. Further, the news sub-section is composed of 78 text files.

## 2. Orthographic Standardization

Orthographic standardization in Persian may be somewhat challenging due to Persian's rich morphology and orthographic inconsistencies (Rasooli, Kholi, & Habash, 2013). One of the main considerations in standardization of Persian texts is the use of semi-space (Rasooli et al., 2013). Rasooli (2013) estimates that about 8% of tokens in Persian dependency treebank include semi-space. In Persian texts, semi-space can be used to mark both inter- and intra- word boundaries. For example, the progressive prefix می 'mi' can be affixed to the verb with no space (e.g., می‌رود), with semi-space (e.g., می‌رود), or with full space (e.g., می رود). On the other hand, compound nouns may be written using semi-space (e.g., رئیس‌جمهور) or full space (e.g., رئیس جمهور).

Although semi-space is common and expected in Persian texts, in order to improve the uniformity of spacing in the corpus, all semi-space characters have been removed throughout the corpus. Another important consideration in standardizing Persian texts is the use of optional white space (Megerdooian, 2004). Persian words ending with /a/, /d/, /r/, /z/ may precede another distinct word with the white space (e.g., مرد قوی) or without the white space (e.g., مردقوی). To achieve orthographic consistency across the corpus, the optional white space has been invariably used between distinct word sequences.

Further, to preserve uniformity within and between the text files throughout the corpus the following conventions have been used:

- a) The suffix ها /-ha/ 'plural' and its inflections such as های /-haie/ or هایِ /-hain/ are treated as separate items. A white space is used between /-ha/ and the noun it attaches to.
- b) The suffixes تر /-tar/ 'comparative' and ترین /-tarin/ 'superlative' are treated as part of the adjective to which they are attached. The suffixes /-tar/ and /-tarin/ are only treated as separate items where the adjacent adjective ends with /consonant + h/.
- c) The suffixes ام /-am/, ای /-i/, ست /-ast/, ه /-eh/, ایم /-im/, اید /-id/, اند /-an(d)/ (denoting copula) and ام /-am/, ات /-at/, اش /-ash/, مان /-mân/, تان /-tân/, شان /-shân/ (denoting possession) are typically attached to their adjacent word. The suffixes are only treated as separate items when the adjacent noun or adjective ends with /consonant + h/.
- d) The progressive morpheme می /mi-/ and its negative form نمی /nemi-/ are always attached to the adjacent verb.
- e) The prefixes نا /nâ-/ and بی /bi-/ are attached to their adjacent word when they are used as adjectival negative affixes. When /bi-/ is used as a binary adjectival negative prefix, as in بیخرد /bi-xerad/ 'unwise' versus باخرد /bâ-xerad/ 'wise', it is treated as part of the adjective to which it is attached. The exception to this rule is when the adjective starts with the vowels /a/, /e/, /o/, /â/. Also, when /bi/ denotes absence, as in آرش بی کار زندگی می‌کنه /Ârash bi kâr zendegi mikoné/ 'Arash lives without a job', /bi/ is treated as a separate item.
- f) Likewise, با /bâ-/ as a binary adjectival prefix is attached to its adjacent word. However, the preposition /bâ/ is always treated as a separate lexicon.

The standardization conventions of the Corpus of Law, Academic, and News are similar to the conventions used in the Corpus of Conversational Persian Transcripts (Mohammadi, 2019a). This allows the user to combine the two corpora for a more comprehensive representation of the Persian language.

### 3. Corpus Composition

The CLAN is composed of 400 text samples from three different genres: legal texts, academic abstracts, and newspaper articles. The general composition of the corpus is illustrated in Table (1).

Table 1. Composition of CLAN

File ID	Channel	Mode	Genre	Text type	Tokens
CLAN001- CLAN274	written	print (online archive)	academic	abstracts	85,765
CLAN275- CLAN322	written	print (online archive)	legal documents	- constitutional text - penal codes	88,170
CLAN323- CLAN400	written	print (online archive)	news	front-page articles	101,055

#### 3.1. Academic Sub-Corpus

The academic sub-corpus consists of 274 text files including a diverse and proportionally balanced representation of published academic abstracts in different disciplinary areas (i.e., Arts and Humanities, Social Sciences, Natural Sciences). The abstracts are retrieved from Iran’s Scientific Information Database (SID), a comprehensive and up-to-date data bank of Iranian scientific journals. The SID platform is available in both English and Persian. The research abstracts are extracted from the Persian section of the SID data bank.

The research abstracts are obtained from published papers in three major areas of Humanities, Social Sciences, and Natural Sciences. The abstracts are selected from different subject areas such as art, history, nutrition, medicine, veterinary, geology, psychology, physics, architecture, computer sciences, etc. The files in this section are annotated with a header that includes the genre, text type, field, topic, and journal specifications (e.g., name, volume, number, and web address). The body of the text is also tagged with a title and one to five paragraphs <p>.

### 3.2. Legal Sub-Corpus

The legal section of the corpus contains 78 text files. Text samples included in the legal sub-corpus are extracted from published legal materials including the constitution of the Islamic Republic of Iran as well as the civil and criminal penal codes of Iran. Documents in this section represent the standard written language of institutional type and show high degrees of content factuality. Legal documents in Persian show the highest level of formality among all spoken and written genres (Mohammadi, 2018). In the legal sub-corpus, each file is annotated with a header that indicates the genre, text type (e.g., constitution, civil, criminal), section, and chapter. The body of the text is also tagged with a title, one or two subtitles, and multiple paragraphs <p>.

### 3.3. News Sub-Corpus

The news sub-corpus is composed of the front-page news articles from various newspapers. The front-page articles typically cover the most important national or international news of the day in political, economic, social, or cultural domains. The news articles were extracted from 10 different newspapers including Keyhân, Ebtekâr, Resâlat, Irân, Etemâd, Farhixtegân, Ârmân-e-meli, Vatan-e-emruz, Doyây-e-eqtesâd, and Jâm-e-Jam. The majority of the newspapers cover general topics. However, some of the papers are more specialized, and their articles are primarily associated with a special domain (e.g., economy or education).

To extract the news articles, 78 issues of the newspapers in the past 10 years were selected from the newspaper archives in the public domain. The articles extracted from each newspaper equally contribute to the news sub-section, and the number of articles from each newspaper is evenly distributed. The main news article on the front page of each issue was fully extracted. However, the images and image captions were excluded. The text files are marked with a header that indicates the genre, the name of the newspaper, the date of publication, and the newspaper's copyright notice. The body is also marked with the news headline and multiple paragraphs <p>.

## 4. Applications

Corpus-based research studies facilitate the discovery of latent patterns and structures in a language (Mohammadi, 2019b). The Corpus of Law, Academic, and News (CLAN) contains three main genres that represent the structure of formal written Persian. The CLAN may be used in combination with the Corpus of Conversational Persian Transcripts (CCP) which represents informal spoken Persian (Mohammadi, 2019a). The combination of the two corpora can create a comprehensive representation of the Persian language yet allows for the juxtaposition of contrastive forms and genres. This creates the possibility of research studies that can pinpoint the structural

differences and/or similarities between the spoken and written language and further determine the linguistic features of various genres in Persian.

Moreover, the CLAN can be useful in multi-dimensional analysis of the language and other corpus-related research such as text analysis and genre analysis. The CLAN may be also useful in pedagogical developments for Persian as a second or foreign language. The CLAN can be specifically used for teaching writing skills in Persian as the corpus documents the structure of authentic language use in Persian texts. The Corpus of Law, Academic, and News (CLAN) is available for research and teaching purposes through the Linguistic Data Consortium.

---

## References

- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243-257.
- Megerdooian, K. (2004, May). Developing a Persian part of speech tagger. In *Proceedings of the 1st Workshop on Persian Language and Computer* (pp. 99-105).
- Mohammadi, A. N. (2018). Discourse markers in colloquial and formal Persian: A corpus-based discourse analysis approach (doctoral dissertation). Retrieved from <http://ufdc.ufl.edu/UFE0052106/00001>.
- Mohammadi, A. N. (2019a). Corpus of Conversational Persian Transcripts: Introduction. Philadelphia: Linguistic Data Consortium.
- Mohammadi, A. N. (2019b). Meaning potentials and discourse markers: The case of focus management markers in Persian. *Lingua*, 229, 102706.
- Rasooli, M. S., El Kholly, A., & Habash, N. (2013, October). Orthographic and morphological processing for Persian-to-English statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 1047-1051).
- Rasooli, M. S., Kouhestani, M., & Moloodi, A. (2013, June). Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 306-314).