Please cite this paper as: Mohammadi, A. N. (2019). Corpus of Conversational Persian: Introduction. DOI: 10.13140/RG.2.2.20630.09286/1.

Corpus of Conversational Persian: Introduction[1]

Ariana N. Mohammadi ✉

Abstract

The Corpus of Conversational Persian (CCP) is a collection of transcribed spoken language data extracted from ~20 hours of naturally occurring informal conversations in Iranian Persian, Tehrani dialect. The data were collected by twenty-two research participants who recorded their daily phone calls and face-to-face interactions in a variety of informal settings. The corpus contains forty-three freestanding XML text files that are validated against an internal DTD declaration. The corpus is annotated and is tagged for gender. This paper provides a description of properties, composition, transcription, markup, and standardization of the corpus.

Keywords: Corpus linguistics; Corpora; Corpus of spoken Persian; Gender-tagged corpus

1. Description

The Corpus of Conversational Persian (CCP) documents spontaneous, authentic, and naturally occurring informal interactions in Iranian Persian, Tehrani dialect. The CCP is part of the General Corpus of Persian (Mohammadi, 2018) and contains two sub-corpora: face-to-face conversations and phone calls. The corpus includes only texts (not

[1] *Content notice: The Corpus of Conversational Persian may contain coarse or offensive language. User discretion is advised.*

the speech) and is extracted from 1,201 minutes of conversation among twenty-two participants (twelve males and ten females).

The research participants recorded their daily phone calls and face-to-face interactions in a variety of informal settings. The conversations represent various interaction types (e.g., dialogue and group conversation), a wide range of settings (e.g., home, office, car, café and restaurant), different types of relationship among participants (e.g., family, couple, friend, acquaintance), and various communicative goals (e.g., joking, informing, explaining, arguing, describing, complaining, self-revealing, etc.).

Biber (1993) notes that mindful and effective sampling of texts in research corpora involves texts diversity in terms of demographic variation of speakers (e.g., sex, age, and occupation), wide range of communicative purposes, and inclusion of different topics. Other parameters that are important in defining diversity in corpora include representing addressees of different plurality, interactiveness, and degrees of shared knowledge.

The main focus in selecting texts in the Corpus of Conversational Persian is achieving diversity in terms of topics, settings, interaction types, and speakers' characteristics (e.g., age, gender, education, and occupation). Further, the corpus contains texts with various levels of interactiveness among participants and different levels of shared knowledge among speakers.

2. Transcription and markup

Stubbs (1983) argues that despite the chaotic appearance of transcribed conversational data, naturally occurring spoken language is highly ordered, as the coherent structure of spoken data is attained through repetitions, markers, and in-time synchronizations. As such, the transcription method used in the CCP is verbatim and includes all linguistic words and non-linguistic vocalizations.

All backchannelling cues, fillers, pauses, hesitations, repairs, and repetitions are included in the transcriptions. Pauses within speaking turns are shown with period(s), and each period indicates approximately one second of pause (untimed). Pauses between speaking turns, however, are not included in transcriptions. Moreover, the symbol " ، " shows continuative intonation in an utterance and the symbol " ؟ " shows the typical raising intonation at the end of an interrogative sentence.

Truncated words, repairs, or disfluencies are also transcribed and are shown with " - " within words (e.g., واژ- واژه 'wor- word'). Overlaps and interruptions are also transcribed in the sequential occurrence as they happen, and each turn is marked by the

speaker who has the floor during the conversation. Where two or more voices coincide, the transcription includes all audible words and phrases.

Non-linguistic vocalizations such as laughs, coughs, sighs, as well as relevant background sounds, are also included in transcriptions within box brackets. There is, however, no indication of voice quality such as loudness or speed of speech. Unintelligible utterances are further marked within box brackets as [نامفهوم] '[inaudible]' without any indication of the duration of the inaudible segment.

Each text file in the CCP is a freestanding XML document validated against an internal DTD declaration and encoded as UTF-8. All text files are marked with a header that indicates the file ID, genre, text type (e.g., phone call or face-to-face conversation), length of conversation, setting, and participants characteristics (e.g., name, age range, and gender). Each turn within the body of conversations is also tagged with the gender of the speaker.

To preserve the anonymity of transcripts, the names of natural persons who are not public figures, the name of private businesses, and other identifying information (e.g., addresses, affiliations, job titles, etc.) have been replaced by pseudonyms. To effectively optimize the anonymization process, the same pseudonym is not used for the same speaker in different text files, that is, a participant's pattern of association with different contexts, locations, and interlocutors cannot be retrieved through the examination of multiple files. Any resemblance between a pseudonym and an actual person or a business entity is coincidental.

3. Orthographic standardization

As Megerdoomian (2004) suggests, an important consideration in transcribing, processing, or tagging Persian texts is the nature of optional white space. In Persian texts, certain words (e.g., words that end with /a/,/d/,/r/,/z/) may appear before another distinct word without the white space. This results in processing of the two words as a single token (e.g., مردقوی). The optionality of the white space is also a point of caution in word affixes.

Transcription of the texts in the Corpus of Conversational Persian is primarily based on standard Persian orthography. To preserve the consistency of transcriptions within and between the text files, the following orthographic standards are used throughout the corpus:

- The suffixes /-ha/ 'plural' and /-tar/ 'comparative' (and their inflections) are treated as part of the noun to which they are attached. The suffixes /-ha/ and /-tar/ are only treated as separate items where the adjacent noun ends with /consonant + h/.

Example: **آدمها، دادگاهها،** بچه **ها، بزرگتر،** مرفه **تر**

- The suffix /-ha/ when used as an emphasis marker is always separated from the adjacent verb or noun.

Example: بهت گفتم **ها**

- The suffixes /-am/, /-i/,/-ast/, /-eh/, /-im/, /-id/, /-an(d)/ (denoting copula) and /-am/, /-at/, /-ash/,/-moon/, /-toon/, /-shoon/ (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/.

Example: خسته **ام،** کجا**ست،** خونه **تون،** شهر**شون**

- The progressive morpheme /mi-/, and its negative form /nemi-/, are always attached to the adjacent verb.

Example: **میرود، میخورد، نمیماند**

- The prefixes /nâ-/ and /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 *bi-xerad* 'unwise' versus *b*â-*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 *ârash bi k*â*r zendegi mikoné* 'arash lives without a job', /bi/ is treated as a separate item.

Example: ادب **نا**راضی، **بیخرد، بی**

- Likewise, /ba-/ as a binary adjectival prefix is attached to its adjacent word. However, the preposition /ba/ is always treated as a separate lexicon.

Example: من **با** اشکان همکارم، تا اونجا که میدونم پسر **با**ادبیه

4. Composition of the corpus

The CCP is composed of forty-three separate conversations (i.e., phone calls or face-to-face interactions) in different settings and with varying lengths. Each text file is extracted from an audio source which has been obtained by one of the speakers in the conversation. Table 1 represents the composition of the corpus and shows the properties of each text file in terms of its type, length, setting, number of participants, and gender of participants.

Table 1. Corpus of Conversational Persian

ID	Type	Length	# of Participants	Gender of participants	Setting
CCP01	phone call	87:19	2	male / female	phone
CCP02	phone call	21:26	2	female	phone
CCP03	phone call	27:24	2	female	phone
CCP04	phone call	00:42	2	male / female	phone
CCP05	phone call	27:15	2	female	phone
CCP06	phone call	44:50	2	female / male	phone
CCP07	phone call	29:17	2	male	phone
CCP08	phone call	25:14	2	male	phone
CCP09	phone call	03:36	2	female / male	phone
CCP10	phone call	21:29	2	male	phone
CCP11	phone call	45:33	2	male	phone
CCP12	phone call	05:02	2	female / male	phone
CCP13	phone call	19:36	2	female	phone
CCP14	phone call	24:36	2	male / female	phone
CCP15	phone call	17:38	2	female	phone
CCP16	phone call	18:03	2	male / female	phone
CCP17	phone call	29:30	2	male	phone
CCP18	phone call	18:07	2	male	phone
CCP19	phone call	29:48	2	female	phone
CCP20	phone call	29:56	2	female	phone
CCP21	phone call	01:19	2	female	phone
CCP22	phone call	28:58	2	male / female	phone
CCP23	phone call	32:03	2	female	phone
CCP24	face-to-face	90:17	3	male	car
CCP25	face-to-face	43:51	2	female / male	car
CCP26	face-to-face	18:08	4	female(3) / male	home
CCP27	face-to-face	57:38	3	male / female(2)	car
CCP28	face-to-face	16:57	3	female(2) / male	home
CCP29	face-to-face	22:35	3	male(2) / female	café
CCP30	face-to-face	26:55	5	male(3) / female(2)	Restaurant
CCP31	face-to-face	21:30	2	male	car
CCP32	face-to-face	27:14	3	male	car
CCP33	face-to-face	54:05	2	male / female	car
CCP34	face-to-face	93:49	3	male /	office
CCP35	face-to-face	28:35	2	male / female	home
CCP36	face-to-face	04:02	2	female	home
CCP37	face-to-face	10:37	3	female(2) / male	home
CCP38	face-to-face	04:47	2	female	home
CCP39	face-to-face	17:08	3	male / female(2)	home
CCP40	face-to-face	20:55	3	female(2) / male	home
CCP41	face-to-face	15:10	2	male	car
CCP42	face-to-face	06:44	2	female / male	car
CCP43	face-to-face	31:16	2	male	car

As Table 1 shows, the CCP contains various texts, topics, settings, and interaction types. The corpus, therefore, may be useful in a variety of corpus-related research areas. The CCP has been particularly compiled and annotated with the following research areas in mind: text analysis, discourse analysis, sociolinguistics, cultural studies, gender studies, and pragmatics. Moreover, the corpus may be useful in corpus-based teaching and learning and the development of authentic teaching materials for courses of Persian as a second/foreign language.

5. Conclusion

The present paper describes the properties, composition, transcription, markup, and standardization of the Corpus of Conversational Persian. The corpus 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.

Megerdoomian, 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.

Stubbs, M. (1983). *Discourse analysis: The sociolinguistic analysis of natural language*. Oxford: Basil Blackwell Publisher Limited.