THE UNIVERSITY *of* EDINBURGH

THE UNIVERSITY of EDINBURGH

A Corpus-based Study on the Use of Conversational Persian by Learners of Persian

Sepideh Daghbandan

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy in Education

University of Edinburgh

December 2022

Abstract

Conversational Persian, or Colloquial Persian as it is also referred to, is at its early stages of receiving attention in the field of Teaching Persian as a Second Language. However, research on the use of Conversational Persian by learners of Persian remains scarce. Therefore, this study aims to explore the use of Conversational Persian by language learners using a novel methodology in this field, namely, by using a corpus-based methodology.

To this end, after performing a pilot study, first, a spoken learner corpus, namely, the Learner of Persian Spoken Corpus (LoPSC) was compiled. LoPSC is the first spoken corpus collected from learners of Persian. It is also the first learner corpus in Persian focusing specifically on Conversational Persian. Data from LoPSC consists of approximately 40,000 words of transcribed audio recordings from conversations between advanced learners of Persian in the UK. After the compilation of LoPSC, to gain a better understanding of the use of the linguistic forms used by the learners, LoPSC was compared to a reference corpus, namely, the Conversational Persian Corpus. This corpus consists of 60,000 words of audio-transcribed recordings from conversations between first language speakers of Persian living in Iran. In addition to a comparative corpus-based methodology, this study also conducted interviews with learners of Persian to further investigate the factors influencing the differences between learner and first language speaker production of Conversational Persian.

The results from this study showed that the most significant difference between the use of Conversational Persian by learners and first language speakers of Persian was the word choice of these two groups of speakers. This difference in word choice was significantly reflected in the use of different discourse markers. In addition, the use of

corpus analytical tools showed significant differences in one category of Vague Language, namely, the use of Vague Category Markers. Learners and first language speakers not only used Vague Category Markers differently regarding frequency but variations in forms were also found. Finally, the findings from the interviews with the learners revealed that in addition to factors such as possible first language transfer and the influences of learner language itself, the learners also displayed certain negative attitudes toward using linguistic forms associated with Conversational Persian.

This study has three main contributions. First, it provides empirical findings in a novel context, namely, the use of Conversational Persian by learners. Second, this study also provides further empirical evidence on how learners use discourse markers and Vague Category Markers in Conversational Persian. Finally, as the first study to compile and analyse a spoken learner corpus in Persian, this study also provides insights into the challenges when compiling a learner corpus in this language, especially regarding the conversational register of Persian.

From the empirical and methodological contributions and findings of this study stem implications for textbook developers for Conversational Persian, Persian Language Instructors, learners of Persian and researchers in this field. This study has direct pedagogical implications by not only providing further insight into the actual use of Conversational Persian by learners but also their perceptions regarding this register. This insight, especially, highlights the forms that require further emphasis when developing teaching material for this specific register. The learner corpus developed in this study provides methodological implications by providing a dataset that can assist future research in this field.

Lay Summary

Conversational Persian, or Colloquial Persian as it is also referred to, is at its early stages of receiving attention in the field of Teaching Persian as a Second Language. However, research on the use of Conversational Persian by learners of Persian remains scarce. Therefore, this study aims to explore the use of Conversational Persian by language learners using a novel methodology in this field, namely, by using Corpus Linguistics. Corpus Linguistics as a methodology aids the analysis of Conversational Language by providing access to naturally occurring data. This data is then analysed using corpus software that provide information such as, the frequency of occurrences and the cooccurrence of linguistic forms.

To conduct a corpus-based study on the production of speech by learners of Persian, since there were no corpora available, the Learner of Persian Spoken Corpus (LoPSC) was compiled. LoPSC is the first spoken corpus collected from learners of Persian. It is also the first learner corpus in Persian focusing specifically on Conversational Persian. Data from LoPSC consists of approximately 40,000 words of transcribed audio recordings from conversations between advanced learners of Persian in the UK. After the compilation of LoPSC, to gain a better understanding of the use of the linguistic forms used by the learners, LoPSC was compared to a reference corpus, namely, the Conversational Persian Corpus. This corpus consists of 60,000 words of audio-transcribed recordings from conversations between first language speakers of Persian living in Iran. In addition to a comparative corpus-based methodology, this study also conducted interviews with learners of Persian to further investigate the factors influencing the differences between learner and first language speaker production of Conversational Persian.

The results from this study showed that the most significant difference between the use of Conversational Persian by learners and first-language speakers of Persian was the word choice of these two groups of speakers. This difference in word choice was significantly reflected in the use of different discourse markers. In addition, the use of corpus analytical tools showed significant differences in one category of Vague Language, namely, the use of Vague Category Markers, which are phrases such as *and things like that*. Learners and first language speakers not only used Vague Category Markers differently regarding frequency but variations in forms were also found. Finally, the findings from the interviews with the learners revealed that in addition to factors such as possible first language transfer and the influences of learner language itself, the learners also displayed certain negative attitudes toward using linguistic forms associated with Conversational Persian.

This study has three main contributions. First, it provides empirical findings in a novel context, namely, the use of Conversational Persian by learners. Second, this study also provided further empirical evidence on how learners use discourse markers and Vague Category Markers in Conversational Persian. Finally, as the first study to compile and analyse a spoken learner corpus in Persian, this study also provides insights into the challenges when compiling a learner corpus in this language, especially regarding the conversational register of Persian.

From the empirical and methodological contributions and findings of this study stem implications for textbook developers for Conversational Persian, Persian Language Instructors, learners of Persian and researchers in this field. This study has direct pedagogical implications by not only providing further insight into the actual use of Conversational Persian by learners but also their perceptions regarding this register. This insight, especially, highlights the forms that require further emphasis when

developing teaching material for this specific register. The learner corpus developed in this study provides methodological implications by providing a dataset that can assist future research in this field.

Acknowledgements

Throughout my PhD project, I have been fortunate to have met people whose support I am indebted to for bringing this thesis to completion. Space (nor my memory) allows the mention of every single name, so I will use this section to solely acknowledge a few names.

First, I would like to thank my principal supervisor, Dr Kenneth Fordyce, for his productive feedback, constructive criticism, and support. I am indebted to Ken for bringing, a much-needed balance, to my supervision team.

I would also like to extend my gratitude to my second supervisor, Dr Farah Akbar. Thank you, Farah, for your encouragement throughout the later stages of my project. I would also like to thank Dr Joan Cutting and Dr Brona Murphy for their feedback on earlier stages of this thesis. I would also like to thank my viva examiners, Professor Sylvie De Cock and Dr Vander Viana, for their insightful comments and feedback on this thesis.

As always, I am grateful for my parents, Soheila and Allahyar. First, I thank you both for financially supporting parts of my PhD project. I also (belatedly) thank you for ensuring I had the best of childhoods while you were in the early stages of pursuing an academic career. I am, and forever will be, in awe of you both.

Throughout the PhD process, I owe a great deal of my wellbeing to acts of genuine kindness from a group of wonderful and supportive people in Edinburgh. Dr Sabina Savadova and family, Iain Gardner, Dr Anthony Schrag, Sally Freedman, Dr Lucy Deacon, and, last but not least, the amazing PhD community of Moray House, thank you all for being my greatest source of comfort throughout my PhD years.

Finally, I would like to express my whole-hearted gratitude to the students at the University of Edinburgh's Islamic and Middle Eastern Studies. I am forever indebted to you for allowing me to listen in to your private conversations and for devoting your time so enthusiastically to this project. Without your interest, participation and appreciation, this thesis would have never had a beginning, middle or end. I hope that this project serves its purpose of helping Persian Language learners like you.

Table Contents

List of Tables and Figures

1. Introduction Chapter

1.1. Chapter Introduction

This introductory chapter first aims to provide the background that informed the rationale behind this study. This background includes a description of the Persian Language and the nature of Conversational Language in Modern Persian, and more specifically in the context of learning and teaching Persian as a second language. After providing this background the research aims, questions and contributions of this study are described. This chapter concludes by illustrating a map of the thesis in the form of brief summaries for each of the thesis chapters.

1.2. Background of the study

1.2.1. Conversational Language

In the description of language, Conversational Language has received less attention compared to the register of language more closely associated with formal written genres, such as academic and legal texts (Rühlemann, 2006). Such texts are referred to as the "standard written form" (Carter and McCarty, 2017; Crystal, 2003). Furthermore, since the descriptions of language are mainly based on this "standard" form; consequently, this form has informed the material used for teaching languages (Carter and McCarthy, 2017).

Whereas depending on the language and other social, historical and cultural factors, there could be several reasons behind Conversational language receiving less attention than the standard form, there are two main reasons that stand out. First, as Crystal (2003, p.235) states, since the taught form is usually presented in the written mode it is easier to capture and hence analyse compared to the spoken mode which

is “fleeting” in nature. Secondly, the language associated with the written form is associated more with “prestige” since it is used for “public official documents and literary language” (Rühlemann, 2006, p. 405). The early roots of the preference for the inclusion of “written grammar” as the choice of grammar to teach goes back to early works in lexicography such as that of Samuel Johnson in which spoken everyday conversations were considered “vulgar” and therefore avoided in the dictionary entries (Carter and McCarthy, 1995).

This focus on the prestige of the written standard form is also reflected in the terminology used for the description of Conversational Language. For example, whereas Leech (2000) promotes the description of a universal grammar for all kinds of registers and genres, Rühlemann (2006), propels for a completely separate description of conversations. Rühlemann’s main reason behind promoting such a separate description is due to the negative light that some common features of conversations are described in when the written language is used as the basis for description. Therefore, Rühlemann (2006) questions using frameworks and terminologies based on the written language for describing conversations (p. 385).

In the context of learning and teaching a second language, Carter and McCarthy (1995, p. 207) stress the importance of the inclusion of Conversational Language in teaching a second language. As they state, by teaching the features and characteristics of Conversational Language, learners will avoid sounding like a “book” (p. 207). In addition, by including Conversational Language in the context of second language teaching, learners will most importantly benefit from the interpersonal and interactive implications of using Conversational Language (Carter and McCarthy, 1995; Jücker et al., 2003). Two of the features of Conversational Language that occur with high frequency and contribute to the interpersonal and interactive functions of this

register are discourse markers and Vague Language, especially Vague Category Markers (Carter and McCarthy, 2017). Each of these features is described briefly in the following.

1.2.2. Discourse Markers and Vague Category Markers

Carter and McCarthy (2017, p.6) emphasise that one of the main differences regarding frequency, distribution and functions between conversational and the grammar associated with written genres is the occurrence of discourse markers. Discourse markers are words and phrases, such as *well, I know, I mean,* etc. that are not part of the clausal structure, but have significant interpersonal implications in the course of the conversation (See Section 2.4. for an overview of discourse markers).

Discourse markers reflect the essence of Conversational Language in that, first, as Watts (1989, p. 224) states there is a large discrepancy between speakers' frequent use of the discourse markers and their evaluation of discourse markers. That is, Watts (1989) found that although speakers use discourse markers frequently, their evaluation of discourse markers is negative. This asserts the notion of Conversational Language as a register with less perceived prestige than the "standard" form. Second, and as will be further expanded on in Section 2.4., discourse markers are a frequent feature of Conversational Language, since they reflect two characteristics of this register, namely, real-time processing and interactivity of Conversational Language.

However, despite the significance of discourse markers in Conversational Language, studies have shown that discourse markers are used less frequently by learners of a language compared to first language users (See section 2.6. on the use of discourse markers by learners). In addition, and most importantly, especially regarding the aims of the current study, studies on the use of discourse markers by learners of Persian remain scarce (See section 3.3).

Similar to discourse markers, as a central component of Conversational Language, vagueness has also been avoided in the description of language (Carter and McCarthy, 2017). In the introduction to her seminal work on vagueness, Channell (1994, p.1) states that "good" language was considered to be language that has clarity and precision and thus should avoid "vagueness, ambiguity, imprecision, and general woolliness". Channell also believes that this description of language has consequently affected the use of learners. On the subject of "vagueness" and awareness, Channell (1994) mainly emphasises the type of vague forms used by learners. That is, considering the variations in vague forms based on their level of formality, forms such as *and so on* may be more commonly used in more formal settings when compared with informal settings. Channell believes that this has implications for learners of a language by stating that even despite the grammatical, phonological and lexical correctness of advanced learners of a language, they still may appear as "rather bookish and pedantic" (p.21). That is, although the learner's goal may not be to attain a "native-like" level of proficiency but there is a preference to avoid judgements such as sounding "dogmatic, impolite, boring, awkward to talk to etc" (Svartvik, 1980, p. 171).

Another problematic area of the use of vagueness by language learners has been associated with the relatively lower frequency of vague forms compared to their L1 speaker counterparts. Although Thomas (1983) does not use the term vagueness or vague forms, she states that the preference for learners to give complete answers or to be over-explicit will lead them to do so in real-life situations which may lead them to sound insensitive, "petulant" or "positively testy" (p.16). Thomas (1983) associates the infrequent use of vague forms by learners to a pragmalinguistic failure. Thomas defines a pragmalinguistic failure as a form of pragmatic failure where there is an

inability on behalf of the hearer to recognise the illocutionary force behind the speaker's utterance when there is a mistaken belief about the pragmatic force of an utterance. According to Thomas, this mistaken belief can stem from two causes: the transfer of L1 norms to L2 and teacher-induced errors. It is with the latter form which she believes not using vague language can lead to a pragmatic failure.

Vagueness in language is divided into several categories (See section 2.5. for a definition of vague language and its categories). However, one of the categories that is especially important in Conversational Language and plays an important role in building rapport and interpersonal bonds between interlocutors is referred to as Vague Category Markers (VCMs) (Jucker, et al., 2003; O'Keeffe, 2014). Vague Category Markers are phrases such as *and things like* (See sections 2.5.2 and 2.5.3. for definitions and functions for VCMs).

1.2.3. Conversational Language and Corpus Linguistics

The advent of audio-recording and technology have made it possible for the spoken language to be analysed. As Adolphs and Knight (2010, p. 73) state "unscripted, naturally occurring conversations can be particularly interesting for the study of spoken grammar and lexis, and for the analysis of the construction of meaning in interaction". Therefore, one of the most prominent uses of corpora has been in determining the differences between what Biber, Johansson, Leech, Conrad, & Finegan (1999) label Conversational Grammar and what Carter and McCarthy (2015) label Spoken Grammar. Conversational Grammar as defined by Biber et al. (1999, p. 1280) refers to features of "grammatical features that are especially characteristic of conversational language, as compared with other registers"; that is features that occur only in conversational language or are more frequent in this register.

One of the types of corpora that has been particularly beneficial for learning and teaching a second language are learner corpora (Granger, 1998). Learner corpora capture naturally occurring data gathered from the language production of learners Regarding the spoken language of learners, learner corpora have subsequently led to significant findings regarding learner language, especially in the case of discourse markers (for example, Müller, 2005) and Vague Category Markers (for example, De Cock, 2004).

1.3. The Persian Language

Modern Persian belongs to the Western Iranian branch of the Indo-Iranian group of the Indo-European language family. It is a descendant of Middle Persian, the official language of the Sasanian Empire (3BCE-7CE) and Pahlavi (old Persian), the language of the Achaemenid Empire (6-4 BCE). Currently, Persian is spoken by more than 110 million people (Miller, et al., 2014). Persian is also the official language of three countries, namely, Iran, Afghanistan, and Tajikistan. Persian is sometimes known by its endonym “Farsi”, which was the term used by all its native speakers until the 20th century. Currently, due to political reasons, it is called Dari in Afghanistan and Tajiki in Tajikistan, whereas, in Iran, the term “Farsi” has remained the name used for this language.

Although the Persian spoken in Iran, Afghanistan and Tajikistan are mutually intelligible by its speakers, there are differences. For example, the Persian of Iran has borrowed many words from Arabic and French. Whereas, in Afghanistan, the Dari Persian is closer to the Middle Persian, and it has borrowed more words from English. Tajikistan, due to it being part of the ex-Soviet Union has borrowed many words from Russian. Regarding their orthographic systems, Persian has adopted the Arabic

writing system since the Arab conquest of Persia in the seventh CE, which is now utilised in Iran and Afghanistan, whereas in Tajiki Persian the Cyrillic alphabet is used. Persian has influenced other languages such as the Turkic languages, Armenian, Georgian and Urdu. It has also influenced Arabic, especially the Bahraini and Kuwaiti dialects of Arabic.

Research on Persian linguistics had previously centred on historical linguistics, mainly on ancient languages of the greater Iran, including Old Persian, Avestan, Pahlavi, and Middle Persian. Within the past century, Modern Persian linguistics has become an "active topic" of research (Shabani-Jadidi, 2018, p.2).

Modern Persian is divided into two main registers, namely, the Standard Written Form and Conversational Persian (Miller et al., 2014). Similar to Conversational English, Conversational Persian also has its unique features as a distinct register of Persian. However, due to the high quantity of the differences between Conversational Persian and the Taught Variety of Persian, the Persian Language has often been categorised as a diglossia (Fergusson, 1995). That is, Conversational Persian (or colloquial Persian as it is also labelled) was referred to as the "low variety" of language and the "standard" form as the "high variety". However, what constitutes this "high" and "low" variety has changed over time. As Megerdoomian (2016) and Shabani-Jadidi (2018) attest, in contemporary Persian, features of Conversational Persian appear in news articles and news reports which were genres previously associated with the High Variety of Persian. Therefore, as Mahmood Bakhtiari (2018) states although there are numerous differences between Conversational Persian and the standard variety of Persian, there is no indication of the contemporary Persian Language falling into the category of a diglossia.

This prevalence of Conversational Persian in other genres can be attributed to various factors. For instance, Nanbaksh (2011) looks at the use of the Tu/Vous System (*tu/shoma* in Persian) by Iranian speakers of Persian in Tehran. She concludes that there is an increase in the use of *tu* and that there is a more egalitarian approach to using this pronominal system. She attributes this change to the influence of Western norms on the Persian society, the increase in the education of Iranians and the egalitarian values promoted by the Iranian Revolution in 1979.

The growth of the scope of Conversational Persian in Iran has subsequently led to a growth in learning and teaching this register for and to second language learners of Persian enrolled in Persian language courses in universities both in Iran and around the world (Shabani-Jadidi, 2018).

1.4. Conversational Persian in the Context of Teaching Persian

The issue of prestige regarding Conversational Persian still holds true in terms of labelling, especially in the context of teaching Persian. This is still reflected in textbooks for teaching Persian. For example, although they acknowledge that there are no clear boundaries between Conversational Persian and the standard form, Ghayoomi et al. (2021) categorise the Persian Language into three categories of the Standard, sub-standard and super-standards (p.26), with the former representing the register used in the official documentation and newscasts and with the sub-standard referring to the Conversational Language and with the super-standard falling into the realm of literary texts.

Similarly, in his textbook for teaching Conversational Persian, Rafiee (2011) aims “to introduce spoken colloquial Persian by providing sufficient guidance for studying literary Persian” (p.xi). Rafiee’s book is based on the writers’ own intuitions and experiences in teaching the Persian Language. In his book on teaching “educated”

colloquial Persian, Rafiee (2011, p.x) presents two main general types of registers for the Persian Language, namely, Colloquial and Literary.

Table 1.1. shows an excerpt from Rafiee (2011) on his description of each of these registers.

Table 1.1

Comparison of the Colloquial and Literary form of Persian (adapted from Rafiee, 2011, p. xi)

Register	Level of formality	Context	example
Colloquial (spoken/written)	Formal	Addressing the elderly, or among participants in more formal settings such as business meetings	mikhahid bahashoun berid Iran? Do you want to go to Iran with them?
	Informal	addressing friends and close relatives	mikhay bahashoun beri Iran?
Literary (spoken/written)	Formal	Business correspondence, books, newspapers, news casting, emails and letters addressing seniors	Mikhahid ba **anha** beh Iran beravid?
	Informal	Personal correspondence addressing friends and relatives of similar or younger age; addressing peers or juniors	mikhahi ba **anha** be iran beravi?

As Table 1.1., the changes and the prevalence of Conversational Persian in all genres of Modern Persian are not reflected in textbooks.

1.4.1. Conversational Persian and Learners of Persian

In Chapter 3 of this thesis, I provide an overview of the studies on learners of Persian. This overview indicates that despite the growing attention to including teaching the conversational register as part of the university courses in Persian, there is a scarcity of studies specifically focusing on the use of this register by learners. Subsequently, studies focusing on salient features of the conversational register, namely the use of discourse markers and VCMs by learners of Persian have yet to be conducted.

1.5. Research aims, questions and contribution

Research Aims and Questions

Considering the above research gaps, this study aims to contribute to the growing field of teaching Conversational Persian by exploring how learners use this register, especially when compared to their L1 speaker counterparts. To this end, the study will answer the following research questions.

The four research questions addressed in this study are in Conversational Persian:

1. What are the significant differences between the forms used by learners and L1 speakers of Persian?
2. What are the differences in the pragmatic functions of the most significant forms used by learners and L1 speakers?
3. What are the differences between the use of discourse markers by learners and L1 speakers of Persian?
4. What are the differences between the use of Vague Category Markers by learners and L1 speakers of Persian?

To answer these research questions, I will use a mixed methods approach. This approach consists of using corpus analytical tools, employing Contrastive Language Analysis (CLA) and conducting semi-structured interviews. Before conducting the analysis of this study, since there were no available spoken learner corpora in Persian, the Learner of Persian Spoken Corpus (LoPSC) was compiled. LoPSC is a 40,000-word spoken learner corpus consisting of conversations from 18 English-speaking learners of Persian. For the analysis of this study, first, I will draw on results from a corpus-based CLA analysis. That is, I compare the LoPSC with a 60,000-word spoken corpus of conversational Persian from 30 speakers of Tehrani Persian. Finally, to gain a better perspective of their use of Conversational Persian, I also interviewed the learners after each recording session.

Research Contributions

This study aims to contribute to the growing field of teaching Conversational Persian by exploring how learners use this register, especially when compared to their L1 speaker counterparts. Therefore, this study will describe a lesser-examined variety of language. In addition, this study will provide information on how learners of Persian use discourse markers and Vague Category Markers. Finally, as a methodological contribution, this study will provide insight into compiling a spoken learner corpus for Conversational Persian. Subsequently, this study will provide perspective on conducting a contrastive corpus-based analysis with a learner corpus and reference L1 speaker corpus in Persian.

1.6. Map of the Thesis

In Chapter 1, I have aimed to provide an overview of the background to this study by drawing on the Conversational register, in general, and narrowing down the focus to the learners' use of Conversational Persian, specifically. I have also aimed at highlighting the importance of discourse markers and Vague Category Markers in Conversational Language before presenting the research aims, questions and contributions of this study.

Chapter 2 presents an overview of the literature related to the register under study in this thesis, that is, Conversational Language. This overview is divided into five sections. The first section of this chapter aims to provide a working definition for Conversational Language by drawing on the already existing literature. The next section presents the main pragmatic theories used in the analysis of Conversational Language. This section aims to provide a theoretical background to the two main components of Conversational Language, namely, discourse markers and vague language, which are further expanded on in the fifth and sixth sections of this chapter. Finally, an overview of the studies on learners' use of Conversational Language, with a specific focus on discourse markers and Vague Category Markers.

Chapter 3 looks into the Persian Language, specifically, by focusing on the studies conducted with learners of Persian. This part highlights the lack of systematic research on the use of Conversational Persian by learners, in general, and especially, in the area of using pragmatic markers and vague language.

In Chapter 4, I move on to introducing the methodological approaches used to answer the research questions of this study. After introducing the two methodological approaches, namely, Corpus Linguistics and Contrastive Analysis, the relevance of

the use of these two methodologies in the study of Conversational Language and learner language is further elaborated on. As with any methodological approach, there were certain limitations in using Corpus Linguistics and Contrastive Analysis for this study; therefore, the next section of this chapter presents the limitations in employing these methodologies. With the introduction, relevance and limitations of the methodological approaches to this study expanded on, the next two main sections of Chapter 4 present the data collection and data analysis phases of this study. The data collection section mainly focuses on presenting the two corpora used to answer the research questions. Subsequently, the data analysis section of this chapter describes the specific corpus tools used to answer the research questions of this study.

Chapter 5, in this chapter, a full description of the analysis used to answer the research questions of this study is illustrated. The initial analysis will draw on the results from using the keywords' analysis tool, which was used to find the most significant differences between the learner and reference corpora of this study. This initial analysis will then be expanded further to find any differences between the pragmatic functions used by the two groups of speakers. This analysis will also consist of focusing on the most frequent forms of discourse markers and Vague Category Markers found in the two corpora of this study.

Chapter 6, in this discussion chapter, I first provide a summary of the findings from the previous chapter by categorising the differences found between the two groups of speakers. I then proceed to consider the findings of this study in the light of previous studies. I conclude this chapter by exploring the possible factors that may have influenced the differences between the use of Conversational Persian between the learners and L1 speakers of this study.

Chapter 7 concludes this thesis by, first, presenting a summary of the findings in the form of addressing each of the research questions of this study. The contributions along with the implications of this study are also presented in this chapter. Finally, the limitations and a set of directions and suggestions for further research have also been provided in this concluding chapter.

1.7. Chapter Summary

In this introductory chapter, I presented the background that informed this study, and the subsequent research aims and research questions that form the framework of this study. I concluded this chapter by presenting the thesis structure in which I provided a summary of each of the chapters of this study. In the following chapter, I set out to provide a working definition for Conversational Language and the theoretical frameworks that have informed studies on this register. In the following chapter, I also present definitions for the two main features of Conversational Language, namely, discourse markers and Vague Category Markers.

2. Conversational Language, Discourse Markers, and Vague Category Markers

2.1. Chapter Introduction

The following chapter presents the definitions and theoretical frameworks that form the basis of this study. More specifically, in this chapter, I will focus on defining what is meant by Conversational Language, in this thesis. I will then proceed to provide the pragmatic theories underlying Conversational Language. Next, since this study focuses on two main features of Conversational Language, namely, discourse markers and Vague Category Markers, I will then provide an overview of the literature on discourse markers and Vague Category Markers. In this overview, due to the focus on Persian, the examples will be mainly drawn from the Persian Language. This chapter concludes with an illustration of the use of discourse markers and Vague Category Markers by language learners.

2.2. Conversational Language

2.2.1. Definition

The definition of Conversational Language depends on the definition of another notion in language, namely, register (Ruhlemann, 2006). Register itself is described by drawing comparisons with the concept of genre. Although register and genre are sometimes used interchangeably, and, at times, clear and distinct definitions of the two concepts are difficult to attain (Ruhlemann, 2006), Biber and Conrad (2009) present a distinct definition for the two concepts and present them as features of any given text. Biber and Conrad (2009, p.5) define a text as any naturally occurring language which is used for the purpose of communication. In turn, they define register

as the pervasiveness of certain linguistic features, which reflect the communicative purposes and situational context of a text (Biber and Conrad, 2006, p.5). Therefore, whereas register is situational in nature, the focus of the genre of a text is more on the conventional structure as opposed to a micro-level analysis of the linguistic characteristics of the text (Biber and Conrad, 2006, p.5).

Biber and Conrad's (2006) definition of register and genre has three main advantages over other definitions of these two concepts. First, since there is a clear-cut distinction presented for register and genre, confusing the two terms or using them interchangeably is avoided. For example, Carter and McCarthy (2006, p. 921) define register as "the style of speaking or writing that is used in particular fields of discourse or particular social contexts (e.g. academic writing, journalism, advertising, legal. Science and literary conventions)". This definition is confusing on three main grounds. First, there is no clear explanation of what constitutes style in this definition. Second, adding the notion of discourse to this definition adds an extra layer of confusion that causes difficulty especially when analysing texts, since discourse analysis and register analysis fall into two different categories. Finally, and most pertinent to this study, determining an operational definition for the purpose of analysis and consequently the scope of Conversational Language would not be possible without a clear definition that distinguishes register and genre.

As Biber and Conrad's (2006) definition shows, two main elements constitute a register, namely, the pervasive linguistic features and the situational context. The determination of the frequent and prevalent features of any given register is determined by quantitative measures. By providing the access to naturally occurring spoken language, Corpora have been used widely to study the features of conversations. Two of the most early and influential studies on Conversational

Language in English were informed by the analysis of the Cambridge-Nottingham Corpus of Discourse in English (CANCODE) (Carter and McCarthy, 1995) and the Longman Spoken and Written English Corpus (LSWE) (Biber et al., 1999). Both corpora have contributed to the production of two comprehensive descriptions of the grammars of the spoken and written registers of English, namely, the Cambridge Grammar of English (Carter and McCarthy, 2006) and the Longman Grammar of Spoken and Written English (LGSWE) (Biber et al., 1999). Since in this study, I also conduct a corpus-based approach to analysing Conversational Language, the role of Corpus Linguistics and corpus tools is described in the Methodology Chapter of this thesis (See Section 4.6)

As for the situational factors that distinguish the Conversational Register from other registers, Rühlemann (2006) argues that since conversations are situationally defined, any description of conversation would require a description of the "situational factors that determine the conversational situation" (p. 385). The two main situational factors that distinguish Conversational Language from other types of registers are what Biber et al. (1999, p. 1041-1052) and Leech (2000) refer to as "interactiveness" and "real-time processing".

Conversations are interactive since they are co-constructed by two or more interlocutors, dynamically adapting their expression to the ongoing exchange" (Biber et al. 1999, p. 1045). Therefore, interlocutors interact through a series of turns in conversations (Sacks, Schegloff & Jefferson 1974; Schegloff, 2000). In addition to making decisions based on the constraints of turn-taking, the interactiveness of conversations is associated with the "expression of feelings and attitudes"; therefore, the interactiveness nature of conversations stems from interpersonal factors (Leech, 2000, p. 697).

On the other hand, real-time processing of Conversational Language involves two features, namely, limited time for planning and simultaneous editing. Due to the ongoing nature of conversations, interlocutors only have a limited time to prepare what they intend to communicate. In addition, as Ruhlemann (2006) states the "limited ability to plan ahead is exacerbated by turn-taking, since speakers cannot simply reduce planning pressure by slowing down or pausing because then other speakers might 'usurp' their turn" (p. 393). In addition, due to the spontaneity and interactiveness of conversations, participants will need to make any desired changes to what they have already reiterated as the conversation continues (Biber et al., 1999; Ruhlemann, 2006).

Situational factors of conversations affect the linguistic choices that participants make. This is referred to as "speech management" which is the "externally noticeable processes whereby the speaker manages their linguistic contributions to the interaction and to the interactively focused informational content" (Allwood, Nivre and Ahlsén, 1990, p. 3). Pertinent to this study are discourse markers and Vague Language which are two main features of Conversational Language (Carter and McCarthy, 2017). These two features not only occur prevalently in this register but also aid speakers with the speech management process required to meet the situational factors that shape Conversational Language (See sections 2.5. and 2.6 for further descriptions of these two features).

2.2.2. Scope

In addition to an operational definition, setting the scope of Conversational Language helps with the clarification required for the analysis of this register. The first issue to cover regarding the scope of Conversational Language is the mode of conversation;

that is, spoken or written. Conversational Language is not only specific to the spoken mode but can also occur in the written mode, for example, through instant online messaging (Carter and McCarthy, 2017). Therefore, conversation may occur using both modes of interaction.

However, in earlier descriptions, Carter and McCarthy (1995) referred to Conversational Language as the spoken language and consequently using corpora and corpus tools provided a description of this “spoken language”, which they labelled “Spoken Grammar”. Spoken Grammar served to differentiate spoken conversations from the existing grammatical descriptions. These existing descriptions were based on registers mainly found in written genres, such as news reports and academic writing. Also using corpora and corpus tools, Biber et al. (1999) found that the grammatical differences in distribution and function between certain written genres, namely, fiction writing, news writing and academic writing were most significant when compared to spoken conversations. However, as Ruhlmann (2007, p.11) states using the term Spoken instead of Conversational falls short since register is confused with the mode of language. Therefore, in this study, the terminology that is used is Conversational instead of Spoken Language.

In addition, to the mode of interaction, there is also the notion of formality which needs further clarification when discussing the scope of Conversational Language. Formal is defined as “variation in speech or writing style in which choices of pronunciation, grammar and vocabulary are made which express a polite distance between participants, as in formal situations such as debates and official ceremonies” (Carter and McCarthy, 2006, p. 904). Informal is defined as “a term associated with variation in speech or writing style in which a more relaxed and colloquial choice of pronunciation, grammar and vocabulary is made, projecting a closer relationship

among participants" (Carter and McCarthy, 2006, p. 908). However, as Handford (2010) and Buttery et al. (2015) state conversation refers to everyday speech used by individuals in various settings that is interactional in nature and can appear as both formal and informal. Therefore, conversation is not only specific to casual and social interactions but speech in situations that are mainly formal in register, such as service encounters, or academic and business settings.

Nonetheless, in informal conversations, there is a higher frequency of the occurrences of forms more closely associated with the features of Conversational Language (Biber et al., 1999). For example, in one of the earlier works on Conversational Language, Carter and McCarthy (1995) use the Nottingham Corpus to explore the spoken grammar of English focusing on what they refer to as "informal and conversational" language as opposed to the more formal varieties of the spoken form of the English Language such as "broadcast talk" (p. 208). Consequently, in this study, I have also focused on informal conversations to ensure that there is a higher frequency of forms in order to allow for a better description of Conversational Language in the collected data.

Having presented an operational definition and scope for Conversational Persian, in the following section I proceed to describe the theories of language pertinent to this register of language, namely, pragmatic theories.

2.3. Pragmatic Theories

In this section, I describe three main pragmatic theories that have informed the study of Conversational Language, in general, and the two main features of Conversational Language, namely, discourse markers and Vague Category Markers, which are further analysed in this study. Therefore, the following section presents an overall depiction of the pragmatic theories, and in sections 2.4. and 2.5, I present a more detailed

account of how these theories inform studies on discourse markers and Vague Category Markers.

2.3.1. Cooperative Principle

Grice's Cooperative Principle forms the basis of its succeeding pragmatic theories. Cooperative Principle presented by Grice (1975) maintains that interlocutors in any given interaction adhere to the following four sets of maxims, namely, the maxims of quantity, quality, relation and manner. Grice (1975, p. 45-47) describes each of these maxims as such:

1. The maxim of Quantity, in which interactants should be as informative as is required for the purposes of the interaction, but not more informative.
2. The maxim of Quality, in which interactants should say only what they believe to be true or that for which they have enough required evidence and to show uncertainty in the case of the lack of evidence;
3. The maxim of Relation, in which interactants should make their contributions relevant to the purposes of the overall interaction;
4. The maxim of Manner, in which interactants should avoid obscurity of expression and ambiguity and should present their utterances in an orderly manner.

However, as Grice (1975) states, addressers usually do not adhere to these four maxims. The result of the violations of any of the maxims (or the "floating" of the maxim) would lead to the addressee resorting to use 'implicatures' to ascertain the speaker's intended meaning.

Grice's (1975) Cooperative Principle has been widely used as a means of elaborating on the However, Grice's theory has been the subject of criticism. Sperber and Wilson

(1986) are critics of Grice's Cooperation Principle. In the following, Sperber and Wilson's response to Grice's theory is explained.

2.3.2. Relevance Theory

Sperber and Wilson's (1986) criticism towards Grice (1975) manifests itself in the form of a modified version of Grice's Cooperative Principle, referred to as Relevance Theory. Based on Sperber and Wilson's (1986) Relevance Theory, Grice's maxim of relation acts as the umbrella maxim that includes all of Grice's other maxims. In other words, the maxims of quantity, quality and manner are all redundant, and the only maxim that explains the choices of interlocutors is the maxim of relation.

Sperber and Wilson (1986) base their theory on both Grice's Cooperative Principle and cognitive theories. Based on cognitive assumptions, Sperber and Wilson (1986) state that members of the speech community share a cognitive environment; that is, interlocutors share background information. Therefore, interlocutors have assumptions about what is to 'manifest to the other' (Sperber and Wilson 1986, p.41). However, the conversational goal for interlocutors is to be relevant while avoiding unnecessary cognitive effort. That is, the smaller the cognitive effort needed to interpret a given utterance, the greater its relevance would be in the conversation. As Wilson and Sperber (2004) state "the central claim of relevance theory is that the expectations of relevance raised by an utterance are precise enough, and predictable enough, to guide the hearer towards the speaker's meaning" (p.607).

To bring an example pertinent to this study, in the case of discourse markers, Blakemore (1987, p.88) presents the following widely stated example for the explanation of the use of discourse markers based on Relevance Theory.

*She slipped. **You see**, the road was slippery.*

As Blakemore states, *you see*, as a discourse marker, “does not add meaning but indicates that the description in the second utterance is an explanation for the information in the first utterance given the assumption that there is a connection between slipping and the road being slippery.”

Although widely used, there are two main criticisms towards the use of Relevance Theory for the description of discourse markers. First, analysing discourse markers by solely referring to Relevance Theory ignores other theories in pragmatics (Aijmer, 2002). Second, as explained, discourse markers perform various functions in conversations; therefore, although Relevance Theory may provide a description for rationalising the choice of interlocutors for using discourse markers, it fails to describe the specific and numerous functions that discourse markers serve. Considering these limitations in using Relevance Theory to describe discourse markers, the following section, first, presents another significant theory in pragmatics, namely, Brown and Levinson’s (1987) Politeness Theory.

2.3.3. Politeness Theory

Brown and Levinson (1987) claim that their model of politeness benefits the three fields of sociology, linguistic pragmatics and anthropology. In the case of linguistic pragmatics, they believe that the great amount of mismatch that exists between what is said and what is implicated can be attributed to politeness. Based on this assumption, they believe that concern with the representational functions of a language should be supplemented with attention to the social functions. In other words, the phenomenon of politeness is seen as one of the reasons that criticises the four maxims of Grice (1975). That is, Brown and Levinson (1987) present politeness

as one of the reasons that interlocutors may avoid the high level of efficiency in communication illustrated by Grice.

As Gumperz states in the introduction to Brown and Levinson (1987), the Theory of Politeness has become to be the "classical treatment on politeness in communication", since it "is basic to the production of social order, and a precondition of human cooperation, so any theory which provides an understanding of this phenomenon at the same time goes to the foundations of human social life" (Brown and Levinson, 1987, p. xiii).

Considering politeness phenomena to be universal principles in human interactions and reflected in human language, Brown and Levinson's goal was to present a universal model for politeness within the human language. They base their universal model on one singular concept, namely, the concept of *face*. Brown and Levinson's definition of face is mainly based on Goffman's (1967) concept of "face" and "from the English folk term, which ties face up with notions of being embarrassed or humiliated, or 'losing face'" (Brown and Levinson, 1987, p.61). Brown and Levinson hypothesise that every individual has two types of face, positive and negative. Positive face is defined as the individual's desire that their wants be appreciated and approved of in social interaction, whereas negative face is the desire for freedom of action and freedom from the imposition created by others.

In line with Goffman's notion of face, Brown and Levinson (1987) assume that facework involves the maintenance of every participant's face for the duration of the social interaction (as far as this is possible), it is therefore in the interests of all the participants to reduce face-threatening to a minimum. Based on this assumption, Brown and Levinson propose two politeness strategies which aim (a) at supporting or enhancing the addressee's positive face (hereafter referred to as positive politeness

strategy) and (b) at avoiding transgression of the addressee's freedom of action and freedom from imposition (hereafter referred to as negative face).

Therefore, Brown and Levinson postulate a set of five possibilities which are available to the speaker to do this, ranging from 'Don't do the Face Threatening Act (FTA)' to 'Do the FTA and go on record as doing so baldly and without any redressive action. If the participant goes on record as doing the FTA, they can soften the blow by carrying out two types of redressive action, (a) by choosing a strategy aimed at enhancing the addressee's positive face or (b) by choosing a strategy which will soften the encroachment on the addressee's freedom of action or freedom from imposition. Positive and negative politeness strategies take their names from the concept of positive and negative face; therefore, positive politeness strategies are directed towards strategies that reduce the impact of FTAs by aiming to address the interlocutors positive face and negative politeness strategies are used to reduce the impact of FTAs by attempting to boost the interlocutor's negative face.

The two main criticisms directed towards Brown and Levinson's Politeness Theory are based on two grounds. The first criticism towards Brown and Levinson's theory is the lack of consideration for the interlocutor in constructing politeness (Bousfield, 2008; Culpepper, 1996; Watts, 2003). This criticism in its turn has brought about theories of impoliteness and Discursive politeness as other proposed theories of politeness. For example, in the theories of impoliteness, the speakers' intentions are considered paramount to determining whether an utterance is polite or impolite and intentions can be determined through various aspects such as the context, interaction type, etc. (Bousfield, 2008).

The other category of criticism directed towards Brown and Levinson's Politeness theory is based on their definition of face. This criticism relates to the culture-

specificness of the concept of face in Brown and Levinson's theory. That is, Brown and Levinson's definition of face is believed to be Westernised and individualistic whereas in other cultures, face is considered to be collective rather than individualistic, such as the Japanese (Matsumoto (1988); Ide (1989)) and Chinese contexts (Gu (1990)).

In the same line and pertinent to this study, Brown and Levinson's (1987) Politeness Theory has received criticism in the Persian context by Koutlaki (2002, 2009). Using recordings of informal conversations between Persian speakers in Tehran, Koutlaki (2002) aimed to find the extent to which Brown and Levinson's theory of politeness applies to Persian Politeness. The results of her study showed that offers and expressions of thanks have reverse functions to those proposed by Brown and Levinson. Based on these results Koutlaki (2002) challenged the concepts of face presented by Brown and Levinson and proposed an alternative concept of face which was more specific to the Iranian Context.

For her alternative definition for the concepts of face, Koutlaki (2002) also draws on the definition of Goffman (1972) who conceptualises face as a person's ''most personal possession and the center of his security and pleasure'', which, however, ''is only on loan to him from society'' and ''it will be withdrawn unless he conducts himself in a way that is worthy of it'' (Goffman, 1972, p. 322). However, Koutlaki also emphasises on Goffman's statement that a person is not only concerned with their own face, but, every single individual, in addition to their own self-respect, is expected to show consideration for the feelings and wants of others and to make efforts to uphold the face of others to show emotional identification with others and their wants (Goffman, 1972, p. 324). Therefore, Koutlaki's definition of face differs from Brown and Levinson's concept of face in that "face is oriented towards an ideal social identity, or

public face" in contrast to Brown and Levinson's face as an individual's self-image (Koutlaki, 2002, p. 5).

Koutlaki (2002) also proposes two replacements for Brown and Levinson's positive and negative faces, namely, *sˇæxsiæt* (personality) and *ehteram* (respect). As Koutlaki states, sˇæxsiæt is similar to Brown and Levinson's notion of positive face although stating some very important differences. For example, Brown and Levinson conceptualise positive face as a person's individual want to be desired, respected and liked, and to have their wants shared by others. However, Koutlaki states that "in an Iranian setting, giving sˇæxsiæt to an addressee has to do with society's injunctions about paying face, and also with group face wants." That is, "behaving in line with societal values is of paramount importance". In Koutlaki's concept of face, sˇæxsiæt is linked to social values and thus is related to public face, as opposed to Brown and Levinson's concept of private face, which is rooted in the individual's wants and desires. As Koutlaki states, "the concept of public face is in line with the closely-knit ties that exist both among the members of the nuclear and of the extended family (the circle of friends and acquaintances), in that unacceptable behaviour reflects badly on one's entire family" (p. 8)

As an example, Koutlai (2002) states the following instance.

> "In Iran it is common for a shopkeeper to nominally refuse payment with the formulaic expression qabeli nædare ("it's not worthy of you"") but this is never meant literally. Such ritual refusals serve a dual purpose: they anoint the speaker's face (sˇæxsiæt) because they show generosity and sincerity, but they also enhance the addressee's face in that she is presented as a person of high standing (sˇæxsiæt) through the show of ehteram (respect)" (p. 5).

Having briefly presented the three main pragmatic theories that inform studies on Conversational Language, the next two sections of this chapter elaborate on two of the main features of Conversational Language and the focus of the second and third research questions of this study, namely, discourse markers and Vague Category Markers.

2.4. Discourse Markers

Although prevalent in the written language, discourse markers are more frequent and varied in structural forms in the spoken language, especially in Conversational Language (Fung and Carter, 2007). However, different definitions exist as to what constitutes a discourse marker. In addition, different terminologies exist for referring to this term and consequently, different forms have been considered as discourse markers. For example, the scope and range of discourse makers may vary from small non-lexical cues such as *mhm* (Norrick, 2009) to phrases such as *y'know* and *I mean* (Schiffrin, 1987). Therefore, in the following, I provide the definition for discourse markers used in the current study, and the model used for analysing this concept.

2.4.1. Definition and terminology

The multiple definitions for discourse markers also reflect the use of different labels used for this concept. That is, there is also a lack of consensus on the label used to define these set of elements. For example, discourse particles (Schourup, 1985; Aijmer (2002); Fischer (2006)), discourse markers (Hansen, 2006; Lewis, 2006a; Waltereit, 2006), and pragmatic markers are some of the labels used in literature (Aijmer and Simon-Vandenbergen (2006)).

Alternatively, the choice of terminology also reflects the definition of these terms. For example, Aijmer and Simon-Vandenbergen (2006) make the distinction between discourse markers and pragmatic markers regarding the functions in language, by stating that discourse markers are a subgroup of the broader group called pragmatic markers. That is, whereas pragmatic markers show both textual and interactional functions, discourse markers only show textual functions, such as signalling textual coherence. This resonates with Green's (2006) classification of English discourse markers into two functional categories, namely, attitudinal discourse markers and structural discourse markers. Attitudinal discourse markers "indicate something about how the speaker feels about what is being said" (Green, 2006, p. 118). On the other hand, structural discourse markers are used by speakers "to indicate a structural boundary in the discourse, and a hint of how what will follow relates to what went before" (Green, 2006, p. 119). Similarly, Brinton (2010) makes a distinction between two major functions of discourse markers, namely textual and interpersonal functions. The problem with this type of classification is the issue of the overlap of the textual and interpersonal functions. That is, in most, if not all cases, discourse markers display both textual and interpersonal functions (Fung and Carter, 2007). Therefore, in this study, I will continue to use the term discourse marker to refer to:

> *words and phrases that are not part of the clausal structure (Carter and McCarthy, 2017), but perform both textual and interpersonal functions (Aijmer and Simon-Vandenbergen, 2006; Green, 2006).*

2.4.2. Schiffrin's Model for Discourse Markers

As Aijmer (2002) states, despite the lack of agreement on the terminology and precise definitions for discourse markers, there is nonetheless a common consensus about

the description of discourse markers in that they are context-specific. That is, the meaning of any given discourse marker can only be interpreted by its linguistic and situational context. This sets the interpretation of the meaning of discourse markers in the realm of pragmatics in opposition to semantics (Aijmer, 2002). That is not to say that discourse markers do not have semantic meanings, but mainly due to the process of grammaticalisation and pragmaticalisation, the meaning of discourse markers is interpreted based on contextual and situational clues instead of their literal semantic meaning. Consequently, the theories around the description of discourse markers are based on theories in pragmatics.

Three of the main pragmatic theories used for analysing discourse markers were explained in section 2.3; however, as mentioned in this section one of the disadvantages of these three pragmatic theories is their lack of cultural specificity. Whereas discourse markers are a strong indicator of the culture of a given speech community, they are also unique to each variety of language and consequently untranslatable from one language to another (Aijmer, 2002; Mohammadi, 2018). Therefore, considering the limitations of pragmatic theories in specifying both the cultural specificity of contexts and the nuance functions of discourse markers, Schiffrin (1987) proposed a model for the description of discourse markers in their given contexts.

Schiffrin's study was the seminal work on discourse markers in the English language, in which she proposes an integral approach to studying discourse markers. In this study, she looked into the following discourse markers *and*, *so*, *or*, *but*, *because*, *then*, *oh*, *well*, *I mean*, *y'know*, and *now*. Regardless of the findings for each individual discourse marker, Schiffrin's (1987) approach to studying discourse markers has one

main advantage over other approaches. That is, Schiffrin uses the bottom-up approach in studying discourse markers.

According to Aijmer (2002), there are two approaches to studying discourse markers in any given conversation. First, there is the top-down approach, in which discourse markers are determined based on the discourse structure. This method has one main disadvantage in that there is a neglect of the discourse markers in their given context. The next approach, which is the approach that Schiffrin uses, is the bottom-up approach. In this approach, the forms are defined and identified before conducting the analysis of the study. Therefore, the bottom-up approach explores each individual discourse marker in its surrounding context. This is especially important in the study of discourse markers which are to a very large extent context and culture-specific (Aijmer, 2002), and is, therefore, the approach used in the current study.

Since this study focuses on the Persian Language specifically and considering the culture-specificity of discourse markers, the following section explores the use of discourse markers in Persian.

2.4.3. Discourse Markers in Persian

There have been three studies on the use of Persian discourse markers. All three studies have used a corpus-based approach and Schiffrin's (1987) model to explore Persian discourse markers. These three studies are Zowghdar and Dabir Moghadam (2002), Alami (2016) and Mohammadi (2018). Each of these studies are expanded on in the following.

Zowghdar and Dabir Moghadam (2002) compared the use of the discourse markers *amâ* in Persian and its equivalent *but* in British English. The British examples were taken from Schiffrin (1987) and the Persian examples were from TV shows in the form of interviews which were mainly on the subject of sports. Zowghdar and Saffar

Moghaddam found that *amâ* displayed varying functions based on its turn position. For example, *amâ* functions to present a new topic only when used in the turn initial position or mid-turn position. However, overall, they reported similar functions for but and *amâ.* For example, both forms were used as a politeness strategy in order to save the speaker's face.

Broadening the scope of Persian discourse markers, Alami (2016) looked at the discourse markers occurring in conversations held between 50 speakers of Iranian Persian (22 male and 28 female). She found 254 occurrences of discourse markers from a 3,105-word corpus. This adds to 8.18 % of the data being discourse markers. The discourse markers that Alami found were *na/na baba* (the highest frequency discourse marker with 33 occurrences), *dige, are/bale/ yani, vali, hala/alan, bebin/nega kon, aslan*. Similar to Zowghdar and Dabir Moghadam (2002), Alami (2016) also looked at the position of the discourse markers as she believed that this affected the function of discourse markers. For example, *bebin (look)* was only used in the initial position indicating its main function as what Fraser (1996) labels as an “attention-getter” in discourse.

Whereas Zowghdar and Dabir Moghadam (2002) looked at *amâ,* Alami (2016) explored the conversational form of *amâ,* namely*, vali*. She found that making a clearcut distinction between *vali* as a conjunction and *vali* as a discourse marker was difficult, since it mainly appears in the mid-position turn showing a form of contrast between two prepositions and therefore has an "intermediary" (Alami, 2016, p. 259) existence between the two functions. However, similar to *amâ*, as a discourse marker, *vali* presents the notion of "counter-expectancy" and can be used as a means of continuing the talk and for giving more information (Alami, 2016, p. 259).

Mohammadi (2018) looked at formal, functional and distributional patterns of discourse markers in Persian using a corpus-based discourse analytical approach. She used a 2-million-word corpus of formal and informal Persian and focused on the discourse markers with a high keyness value in the Conversational subcorpus of her corpus (Mohammadi, 2019) (See Chapter 4 on Methodology for a full description of this corpus). Mohammadi found that the discourse markers in Persian have their origin in formal written Persian; however, they serve a small number of functions. Nonetheless, it is in spoken colloquial Persian in which they are grammaticialised and serve more functions. Similar to the two previous studies on Persian discourse markers, Mohammadi (2018) also explored the positions of discourse markers in Persian. She found that contrary to English, some Persian discourse markers are not only prohibited from the initial position but can only appear in the final position (p. 14-15).

Having presented an overview of discourse markers, in the next section, I present the next main feature of Conversational Persian, namely Vague Language. In this section, I specifically focus on Vague Category Markers a specific category and the focus of the fourth research question of this study.

2.5. Vagueness in Language

Phrases, such as, *I think* and Vague Category Markers have been attributed to the notion of vagueness in language (Channell, 1994). Channell states that vagueness can be expanded to an extent for it to include any form of language use. However, for the purpose of this study, a more comprehensive and narrow framework is needed to make the analysis of vagueness plausible. Therefore, in this section, I will aim to provide a working definition for this term in this study.

Channell (1994) was the first to systematically and extensively explore vagueness in language. In her study, which looked at the occurrence of vagueness in different

registers and genres of British English, she concluded that due to the prevalence and multiple functions of vagueness in language, “a complete theory of language must have vagueness as an integral component” (Channell, 1994, p.5). Regarding the theories of languages used for accounting for vagueness, Lakoff (1975) stated that vagueness cannot be justified with the existing models in semantics for meaning interpretation, namely, the truth-conditional model of semantics. Lakoff stated that the tripartite semantic model which is based on the interpretation of the meaning of utterances as being either true, false, or lacking a truth value was incomplete and introduced the concept of *fuzziness* of meaning in semantics. Lakoff’s (1975) concept of fuzziness was taken by the mathematician Zadeh (1965), who introduced Fuzzy Logic in mathematics[1].

Based on his theory of fuzziness in language, Lakoff (1975) also introduced the term “hedges” which he labelled as words and expressions that make utterances fuzzy or decrease the degree of fuzziness. Words and phrases such as “sort of”, “kind of”, and “technically speaking” fall into this category. Hedges have been closely associated with the notion of vagueness. This association of hedges with vagueness stems from the study of Prince et al. (1982). Prince et al. looked at the hedges in the conversations between doctors in the Intensive Care Unit (ICU) and found that hedges are frequently used within this setting. They found hedges being used between 150 and 450 times per hour which is equivalent to a hedge being produced every 15 seconds. Prince et al. (1982) associated this frequency of occurrence as a means for doctors to avoid generalizing while interacting with patients.

[1] Semantic models are also based on mathematic models; therefore, it is common for mathematic models to be introduced into the field of semantics.

However, Channell (1994) believes that for the purpose of the analysis of vagueness, Lakoff's hedges should be divided into two categories where only one of the categories is considered as vague language. Chanell proposes that hedges should be divided into two separate categories of shields and approximators. Shields show the commitment of the speaker to the truth of an utterance while approximators introduce fuzziness within the preposition. Examples of shields are phrases such as "I think", "I believe", "It seemed", "It appeared" and examples of approximators are "approximately". Channell (1994) claims that while approximators fit in with her definition of vagueness, shields do not. That is, as Carter and McCarthy (2006) state Shields show the epistemic stance of the speakers. On the other hand, in a study of vague forms used by American, Persian and Chinese speakers of English, Sabet and Zhang (2015, p. 48), who refer to shields as "subjectivizers" identify them as vague forms. In this study, I also take the stance of Sabet and Zhang (2015) and consider the expression, *I think* as performing multiple functions. That is, although at times it does show the stance of the speaker, it can also create vagueness in an utterance.

Overall, Lakoff's concept of fuzziness in semantics suggests that utterances also display a degree of truth rather than being either true, false, or nonsense, as suggested by the Truth-conditional semantics. Therefore, for the interpretation of the meaning of vagueness in language, other theories of language other than semantic theories need to be taken into account.

Channel (1994) proposed the use of pragmatic theories for the interpretation of the meaning of vague forms in language. Channell stated that Grice's (1975) Cooperative Principle and Brown and Levinson's (1978) Theory of Politeness can be used for the justification and interpretation of vagueness in language (See Section 2.3). For

instance, Channell (1994, p. 32) gives the following two examples that demonstrate how Grice's Cooperative Principle can be used to account for the occurrence of vague forms. The first example is that of a person replying to what time they would arrive home as being *about 6 o'clock* when the use of *about* indicates that they are not entirely sure of what the exact time might be due to factors such as workload, traffic, etc. Therefore, the speaker would be adhering to the maxim of quality in this case.

The second example from Channell (1994) shows how the use of vagueness can be explained when the floating of a maxim happens. In this example, the quantity maxim is floated when a linguist is presenting at a conference.

> We've got *about five or six* of them but I'm only going to talk about three of them today. (Channell (1994, p.33)

The addressees would assume that the linguist who has done the research would know the exact number of informants (*them* in this extract refers to the number of informants in the linguists' study). However, by flouting the maxim of quantity, the addressees are able to recognise that the speaker is trying to shift the focus of attention from the number of informants to more important information, such as the three informants that he is planning on talking about in the conference.

Therefore, there are two notions central to the concept of vagueness. First, vague forms are context-bound for their interpretation of meaning. That is interpretations of meaning can be made when words appear in context. This echoes Tarniyokova (2009, p. 119) who considers vagueness strategies and manifestations employed by language users to be "partly universal but to a considerable degree language and culture-specific." In addition, as Zhang (2011, p. 579) further asserts "the interpretation of VL is socially and culturally co-constructed". Second, speakers use vague forms

strategically to perform communicative goals (See Section 2.6 below on functions of vagueness in language). Channell (1994, p. 20) labels this strategic use of vague forms in language as a purposeful and "unabashed" decision. This conscious decision of the speaker in choosing vague forms to fulfil a communicative purpose has also been emphasized in other definitions of vagueness in language (Carter & McCarthy, 2006; Zhang, 2011).

Carter and McCarthy (2006, p.928) define vague forms as "words and phrases with very general meanings (thing, stuff, or whatever, sort of, or something, or anything) which deliberately refer to people and things in a non-specific, imprecise way. Also stating that, "purposefully, vague is very common in informal spoken language." Imprecision as a synonym for vagueness has also been used in other studies of vague forms in language (Crystal and Davy, 1975; Jucker, Smith, & Lüdge 2003; Zhang, 2015).

On the other hand, the notions of generality and indeterminacy have been referred to in other studies such as Crystal (2008) who provide the example of the word *parent* for generality and the phrase *a balloon* in the sentence *Mary saw a balloon* for indeterminacy. Although Crystal (2008) and other studies that synonymise vagueness with terms such as generality and indeterminacy are grounded in the field of linguistics, these concepts stem from studies in the field of psychology and specifically to the concept of category membership in psychology by Rosch (1975). Rosch concluded from a series of experiments that perceptual categories such as shapes and colours have what she phrased as an "internal structure"; i.e., perceptual categories have a set of the best exemplifiers for the category with other members of decreasing similarity to this core meaning. Having reached this conclusion, she also found that

this held true for psychological concepts. For example, she found that the category of "birds" had an internal structure with the best examples of the category being birds, such as robins and eagles and with surrounding examples in decreasing order of chickens, ducks, geese, penguins, pelicans, bats, etc. Therefore, a robin or an eagle may be considered as having more characteristics that would put them in the category of "bird" when compared to pelicans and bats. Rosch's theory of categorisation is especially salient in the explanation of Vague Category Markerwhichat will be further discussed in Section 2.6.1.

A final notion regarding the definition and scope of vagueness is the concept of the *resolvability* of the vague form. That is, whether an exact interpretation or a precise reading of vague forms is possible. In the context of the Persian language, which is the focus of this study, Parvaresh and Tayebi (2014, p. 568) categorise forms as vague only if they remain *unresolvable* when they appear in the context. As an example, in the following extract adapted by Parvaresh and Tayebi (2014, p. 569), the Persian word *chizi* (translated by Parvaresh and Tayebi (2014) as *thingy*) is not categorised as vague, since the referent appears immediately in the context and is therefore resolved.

> Speaker: ….daram donbale *chizi* migardam. Donbale in marker sabze!
>
> Speaker: . . . I'm looking for the *thingy* (.) for the green marker!
>
> (Parvaresh and Tayebi, 2014, p. 569)

However, in this study, I have also considered vague forms that are resolvable in meaning based on two grounds. First, when discussing resolvability a question that comes up is "resolvable by whom?" That is, is the vague form resolvable by the speaker or hearer? Channell (1994) states that she defines meaning based on what

is perceived by the hearer. She states two reasons for making this decision. First, because the structure of the conversation is based on the subsequent actions of the hearer. Second, as Channell (1994, p.26) states “hearers’ meaning” is more open to empirical observation whereas finding the speakers’ intentions would be difficult. In addition, methodological problems may appear when asking speakers about their intentions to determine the function of vague forms. For example, speakers may conceal their true intentions, and, in certain cases, may also find being asked about their intentions as intrusive. Therefore, decisions on functions of vague forms are generally based on cues from the context.

One of these cues is based on the hearer’s reaction to an utterance. That is, the speaker’s intentions can be construed as deliberately withholding information when a speaker avoids questions asked by the addressee to clarify a certain vague expression. This can be further explained using Zhang’s (2020) example of a conversation between an officer in the Australian customs and a passenger who has been stopped for questioning regarding a suspected illegal activity.

> Passenger: Yeah, he give me ***some*** money.
>
> Officer: He gives you some money.
>
> Passenger: Yep.
>
> Officer: How much money?
>
> Passenger: (pause) ***Six, seven hundred*** …
>
> Officer: Okay, up to seven hundred dollars?
>
> Passenger: Yeah. (adapted from Zhang (2020, p. 117-118)).

In this example, in order to minimise the seriousness and potential penalty of the illegal activity he is a suspect of, he resorts to the vague expressions “some money” and “six,

seven hundred" instead of providing a precise figure even after the officer has asked him for a precise amount of money.

Second, this *unresolvability* of vague forms conflicts with widely referred to functions of vague forms, such as using vague forms for lexical gaps due to short-term memory loss or using vagueness to withhold information (Channell, 1994; Zhang, 2020) (See Section 2.6.2 for further functions).

Having looked at the definitions and theories underlying the concept of vagueness, as a conclusion to this section, a working definition used for vagueness in the current study is provided.

> *For the purpose of this study, a suffix, word or phrase that creates imprecision and* under *specificity is referred to as a vague form. This imprecision and/or under specificity may be resolved; i.e., there may be an exact meaning that could replace this vague form. Vagueness is used purposefully by language users to serve communicative goals. In addition, vague forms are highly dependent on the context for their meaning. For example, although the vague quantifier "about", is used to show approximation, "about 12 o'clock" and "about 12 chairs in the room", the vague form "about" shows different approximations or intervals of numbers based on the context in which it occurs.*

Based on this definition of vagueness, there are various categories of vague forms; however, what is salient to this study is that different vague categories show different variations in frequency and the type of vague forms used. For example, vague quantifiers appear more in academic talk and Vague Category Markers appear more in informal conversations between familiars (Channell, 1994). Therefore, based on this

frequency of the vague forms and the functions that Vague Category Markers fulfil in conversational language, these forms will be the main focus of the next section.

2.5.1. Vague Category Markers: Definition and Categories

Vague Category Markers (VCMs) refer to phrases such as *or something*, *and everything*, *and things, and stuff*, etc. which are added to *exemplars* (See section 2.6). For instance, *sports and things like that* is an example of the use of an exemplar (*sport*) and the vague category marker *and things like that* to form an ad hoc category (Barsalou, 1983; Overstreet and Yule, 1997). The spontaneity of these ad hoc categories and the "context-dependent nature" of these categories for the interpretation of meaning reflects an important feature of their use in real-time face-to-face interactions (Cheng and Warren, 2001 p. 385) which is a feature of Conversational Language (See section 2.1.).

Vague Category Markers (VCMs) in this study refer to "a relatively homogeneous set of forms consisting of a conjunction plus a noun phrase" that are added to other phrases or clauses (Overstreet, 1999, p.12) for the purpose of creating an ad hoc category (O'Keeffe, 2014, p.19). The meanings of these ad hoc categories are "socio-culturally grounded and are co-constructed within a social group that has a shared socio-historic reality". (O'Keeffe, 2004, p.6).

Various labels have been used for describing Vague Category Markers. For example, Ball and Ariel (1978) and Ward and Birner (1993) refer to them as tags; Dines (1980) and Macaulay (1991) use 'terminal tags'; Biber et al. (1999) refer to these forms as "coordination tags"; 'general extenders' are used by Overstreet and Yule (1997); 'generalised list completers' is the term used by Jefferson (1990); and extension particles by Dubois (1992). Overstreet (1999) uses the term General Extender

because she believes that “general” refers to something unspecific, and she uses the term “extender” “because they extend otherwise complete utterances” (p. 3). However, others have used the term vague for their labels of these forms. For example, Channell (1994) uses the term ‘vague category identifiers’; Cheng and O’Keeffe (2015) refer to them as vague tags (p. 360); and O’Keeffe (2004) uses ‘vague category markers’. In this study, similar to O’Keeffe (2004), I have opted for using the term Vague Category Markers (VCMs), since this label encompasses the vagueness associated with the interpretation of the meaning of these terms and the formation of ad hoc categories referred to by Barslou (1983) and Overstreet and Yule (1993).

Regardless of the terminology used for describing VCMs, there are two aspects regarding these forms that previous studies have in common. First, there is a consensus that certain forms of VCMs are used more frequently based on the level of formality of the text. For example, Biber et al. (1999) state that the most frequent of these forms in the conversational data found in the Longman Spoken and Written Corpus are *or something*, *and everything*, *and things* and *and stuff*. Similarly, Dave (2001) found *and stuff* to be significantly frequent in British Conversational Language.

Regarding phrasal structure, VCMs are generally formed using a conjunction and a noun phrase (Overstreet, 1999). Overstreet presents the following structure for forming VCMs in English:

conjunction + (I don’t know) + (preposition) + noun phrase

In addition, Overstreet (1999) divides VCMs into two categories based on the conjunctions that they are formed with. That is, VCMs formed with the conjunction *and*

are labelled as adjunctive VCMs and conjunctions formed with the conjunction *or* are labelled as disjunctive VCMs. However, Overstreet (1999) also states that VCMs may appear without conjunctions as shown in Extract 1 or Extract 2. Both extracts are adapted from Overstreet (1999, p. 11).

Extract 1

I show myself about eighty feet out, ***something like that***.

Extract 2

it's just about, you know, questions like, you know, are you still coming the twelfth, do you need me to meet you somewhere*,* ***blah blah blah****.*

In Extract 1 and Extract 2, *something like that* and *blah blah blah* are VCMs that appear without a conjunction.

Persian VCMs also have multiple similarities with English VCMs. The next section provides an overview of Persian Vague Language studies, with a specific focus on VCMs in Persian.

2.5.2. Persian Vague language and Vague Category Markers

Since Channell's study for the English Language, vagueness has also been explored in other languages including Persian. As Parvaresh & Sheikhan (2019) state, vagueness is a frequent feature of the Persian Language, and it is especially salient in Conversational Persian.

Parvaresh and Tayebi (2014) are the first to conduct a study on Vague Language in Persian. Based on a 160,000-word spoken corpus of informal conversations in Persian among friends and family members, they identified 9 categories for Persian Vague Language. Each of these categories are described in the following.

The first category of Persian Vague Language identified by Parvaresh and Tayebi (2014) are vague quantifiers, such as, “bish az” (more than), and “chand (ta)” (some). Other categories that they found were also reported in studies on English Vague Language. These categories include vague nouns, such as “chiz” (thing); VCMs, such as “va injoor chizha” (and things like that); vague adverbs, such as, “ye jayi” (somewhere), vague hedges, such as, “ye jurayi” (sort of), “ta hadi” (to an extent). Among these categories, vague nouns were found to have the highest frequency of occurrence in their recorded conversations.

Parvaresh and Tayebi (2014) have also reported differences regarding the forms of Persian and English Vague Language. The first of the differences are Vague Language forms labelled as “rhyming words” by Parvaresh and Tayebi (2014, p. 572). Rhyming words refer to non-existent words that rhyme with a proceeding noun, such as the non-existent word “metab” in the phrase “ketab metab” (roughly translated into “book mook”). This rhyming word always starts with either a /b/ or /m/. Although Parvaresh and Tayebi (2014, p. 572) identify “rhyming words” as completely “unique” categories in Persian, although rhyming words do not follow the typical structure of adjunctive VCMs that is a conjunction and a noun phrase (Overstreet, 1999, p. 3), the rhyming word acts similarly to an adjunctive VCM in that it forms an ad hoc category (Overstreet, 2019) with the noun phrase that it has been added to. As an example, in the expression “ketab metab”, “metab” refers to expressions such as “ketab and the things (like that)”. Nonetheless, it is different from a VCM, because it is more dependent on the noun phrase that it proceeds; i.e., the rhyming word is formed based on the noun phrase proceeding it. Therefore, in terms of forming an ad hoc category, Persian “rhyming words” and adjunctive VCMs in English are the same, and their difference lies within their formation. Moreover, although not found in English, “rhyming

words" have been observed in other languages, such as Turkish and Hebrew (Parvaresh & T. Tayebi, 2014).

Regarding VCMs, specifically, in a study of the Persian VCMs used by speakers of Persian in colloquial speech, Parvaresh et al. (2012) found that Persian speakers also used the "conjunction + noun phrase" structure to form VCMs. Tables 2.1 and 2.2, respectively, show the list of Persian adjunctive and disjunctive VCMs found by Parvaresh et al. (2012) from a 104,000-word corpus of Conversational Persian compiled from data between familiars.

Table 2.1

Adjunctive VCMs in Persian (adapted from Parvaresh et. al (2012))

Form	Frequency	Percent
va inâ (and stuff)	91	37.6
va az in harf hâ (and of such talks)	21	8.67
va in chiz hâ (and such things)	17	7.02
va nemidunam az in harf hâ (and I don't know of such talks)	16	6.61
va hame chiz (and everything)	12	4.95
va az injour harf hâ (and of such sort of talks)	9	3.71
va az in chartvapart hâ (and of such nonsense)	8	3.3
va az injoor chiz hâ (and of such sort of things)	8	3.3
va az in mozakhraf hâ (and of such flams)	7	2.89
va az inqabil chiz hâ (and of such group of things)	7	2.89
va az injoor k âr hâ (and of such sort of issues)	7	2.89
va in va un (and this and that)	7	2.89
va va va (and and and)	6	2.47

va az in jafangijât (and of such hot air)	5	2.06
va az indast harf hâ (and of such kind of talks)	5	2.06
va az indast khabar hâ (and of such kind of reports)	4	1.65
va az in qalat hâ (and of such big talks)	3	1.23
va az in kâr hâ (and of such issues)	3	1.23
va az in jaryân hâ (and of such drifts)	3	1.23
va az in qertibâzi hâ (and of such shows)	2	0.85
va az in khozabalât (and of such rubbish)	1	0.41
Total	**242**	**100**
Total forms	**21**	

Table 2.2

Disjunctive VCMs in Persian (adapted from Parvaresh et al. (2012))

Form	Frequency	Percent
yâ chizi (or something)	22	29.41
yâ chi (or what)	18	26.47
yâ nemidunam chi (or I don't know what)	15	22.05
yâ harchi (or whatever)	5	7.35
yâ harjâ (or wherever)	4	5.88
yâ ye hamchin chizi (or something of that sort)	4	5.88
Total	**68**	**100**
Total forms	**6**	

Based on the VCMs from Tables 2.1 and 2.2, Parvaresh et al. (2012, p.265) claim that the structure of Persian VCMs is identical to the English structure of VCMs stated by Overstreet (1999) (See Section 2.6.1).

An investigation of the frequency and structural distribution of VCMs by Parvaresh, et al., (2010) reported that adjunctive VCMs are more frequently used in Persian than disjunctive VCMs. That is, Persian speakers prefer VCMs beginning with *and* to ones beginning with *or*. This is in line with Cheshire's (2007) finding on native British English, but in contrast with Overstreet's (2005) finding of native American English. Another finding by Parvaresh et al. (2012) was that disjunctive VCMs in Persian were less likely to occur after prepositional phrases, and that 'both Persian and English disjunctive general extenders show smaller variability of forms compared with their adjunctive counterparts' (p. 33). A final difference that Parvaresh et al. (2012) reported was that whereas Persian speakers demonstrated a tendency to use VCMs both clause- finally and clause- internally, they appeared in English only in clause- final position.

The significance of the prevalence and frequency of certain forms of VCMs based on the level of formality and the conjunction used in their formation is related to both the functions of VCMs and consequently to how learners use these forms in their conversations. Both issues will be discussed, respectively, in the following.

2.5.3. Functions of Vague Category Markers

As per the definition of vagueness discussed in Section 2.6, vagueness is used 'to avoid an excess of precision', or to achieve 'imprecision' or 'imprecise language use' (Crystal and Davy, 1975, p. 112–114). This imprecision created by vagueness contributes to 'naturalness and the informal, convergent tenor of everyday talk' (McCarthy, 1998, p. 108–118). As Leech (2000, p. 695) states vagueness allows 'a

speaker to take refuge in strategic imprecision'. This "strategic imprecision" that Leech (2000) refers to 'enables polite and non-threatening interaction' (Carter, 2003, p. 92) and 'softens expressions so that they do not appear too direct, or unduly authoritative or assertive' (Carter and McCarthy, 2006, p. 202).

In addition, Zhang and Parvaresh (2019) believe that speakers resort to using vague forms when they believe that an exact word or phrase may be complicated for the addressee. They believe that providing a vague utterance that requires a less cognitive process for the hearer may be more suited to the purpose of communication. Although Zhang and Parvaresh (2019) categorise this function as "giving the right amount of information" in their taxonomy of vague functions, Cheng (2001) labels instances in which the speaker uses vague forms to avoid burdening the addressee(s) with the cognitive process needed for interpreting a specific word or phrase as an "accommodation strategy" (p. 94). Having observed the spoken conversations in English between L1 speakers of British English and L1 speakers of Cantonese, Cheng (2001) found that both speakers used vague forms as an accommodation strategy. The following extract from Cheng (2001, p, 94-95) shows an example of the use of this accommodation strategy. In this dialogue that occurs between an L1 English Speaker, "A", and an L1 Cantonese speaker, "B", speaker "A" is believed to use the vague noun "*stuff*" to first refer to a general set of salted food before moving to more specific examples (i.e., dry salted plums and prunes). Cheng (2001) states this as a strategy on speaker A's part to avoid forming comprehension problems for speaker "B".

1. A: yea well all my life in Thailand they don't eat that much you know
2. what I mean is the food is light (.) you have rice and vegetables
3. and soups things like that and then all of a sudden

4. to move to the States and have some meat and potatoes (.) you can't

5. eat very much of (.) already gained weight even though I don't eat

6. very much because the food is so um rich and they have sugar in

7. everything

8. B: yes right so you usually take dessert after meal

9. A: not usually

10. B: not usually

11. A: = I don't like sweets when I was a teenager I ate a lot of chocolate

12. B: ah huh

13. A: but I lost the sweet tooth and I really like salty *stuff* you know the

14. dry salted plum and prunes

15. B: oh yes

16. A: those are my favourites that's what I eat for candies

17. B: wow that's very salty

18. A: yea

(Adapted from Cheng (2001, p, 94-95))

Cheng and Waren (2007) state that the use of vagueness to perform the accommodation strategy can only be seen in intercultural communications. However, studies with first language users have also shown the use of this strategy. For example, and more specific to the context of this study, in a study of the functions of

Persian vague forms, Parvaresh and Tayebi (2014) found that Iranian speakers of Persian also used vagueness to accommodate their interlocutors.

Another function of vagueness in language is using vagueness to withhold information, some have categorised this function as a non-cooperative function compared to other categories of vagueness (Parvaresh and Tayebi, 2014; Zhang, 2011) which are generally perceived as showing cooperation between interlocutors (Cheng, 2007). Zhang (2011) was one of the first to address this extensively. In the case of Zhang (2011), vague language was used in what she called “tension-prone” situations, such as the example of the customs of Australia presented above. Therefore, vague forms would be used to “deliberately avoid conveying correct/accurate information to manipulate the situation to the speaker’s advantage” (Zhang, 2011, p. 577).

Specifically focusing on the functions of VCMs, studies also point to the higher frequency of these in informal situations (Crystal and Davy, 1975). This is justified in the notion of the interpersonal functions that VCMs serve especially with how they are used for positive and negative politeness strategies. Terraschke (2007) showed that even in conversations between strangers there are more chances of an increase in FTAs; therefore, the speakers resort to using more vague language to decrease the threat of the FTA. In addition, they use more VCMs to produce "an amicable situation" (Terraschke, 2007, p.151) by drawing on commonalities between the speakers.

In addition to the use of VCMs for performing politeness strategies, especially the use of VCMS for mitigating face threats (Cheng and O’Keeffe, 2015, p. 383-389), VCMs perform social, interpersonal and interactional functions by making conversations sound less formal and less direct. Therefore, the use of VCMs indicates a high level of interactivity, particularly in highly context-dependent conversations where responsibility for meaning-making is shared among speakers (Cheng, 2007). For

example, O'Keeffe (2003) found that speakers in Irish Radio Talk shows create a pseudo-informal atmosphere by using VCMs.

In Conclusion, studies provided in the previous sections point to the prevalence of discourse markers and vague language in Conversational Language. Albeit discourse markers and vague language are both features of other registers, studies show that they are more frequent in terms of both formal types and functions (Ruzaite, 2018). Carter and McCarthy (2017) state that discourse markers and vague language are two important components of Conversational Language. This was also shown to hold true for the Persian language regarding both discourse markers (See Sections 2.4.3 and 2.5.3) Considering the saliency and significance of discourse markers and vague language in Conversational Language, research on the use of these two components by learners helps inform the learning and teaching of Conversational Language.

Although there is a lack of research on the use of discourse markers and VCMs by learners of Persian, there is a rich and growing literature on these two features by learners of other languages, and especially learners of English. Although differences are found in these studies, there are also similarities that point to certain characteristics that are specific to learner language. The following section aims to provide an overview of these studies.

2.6. Second Language Learners, Discourse Markers and VCMs

One of the extensive studies on the use of discourse markers by learners was conducted by Muller (2005). Muller's aim was to investigate how American speakers of English and German speakers of English used four English discourse markers, namely, *well*, *so*, *you know* and *like*. To this aim, Muller (2005) compiled the Giessen-Long Beach Chaplin (GLBCC) Corpus. The GLBCC Corpus consists of two sub-

corpora, namely, a sub-corpus spoken solely by American speakers of English and the other corpus spoken by German speakers of English. Muller divided the function of discourse markers into textual and interactional. In addition, based on previous studies that pointed to a higher frequency of discourse markers in informal conversations, Muller compiled the corpus from spoken conversations between students discussing a film they had just watched. Furthermore, in line with Thomas (1985), Muller (2005) associated the use of discourse markers with the speakers' pragmalinguistic behaviour. Overall, Muller's (2005) study points to the following conclusions that set out the main recurring themes in the study of discourse markers used by learners.

1. Muller (2005) examined whether the number of hours of contact with L1 speakers of English affected the use of discourse markers by the learners. She found that the number of hours of contact had the most effect on the learners' production of discourse markers. The number of contact hours with L1 speakers has also been shown to affect the type of discourse markers used by English speakers of other languages. For example, Secova (2017), in her study of the use of pragmatic markers by French speakers of English and English speakers of French found similar results. This attests to previous research from Hays (1992) who stated that discourse markers were learnt in contact with native speakers. The importance of the number of hours of contact can in term reflect the low frequency of the appearance of discourse markers in language textbooks which is reflected in the second conclusion of Muller (2005).

2. In addition to the number of hours of contact the number of occurrences of the discourse markers in textbooks also affected their frequency of use and their similarity to the functions that the American speakers used. In the case of the German speakers, this was true about the discourse marker *well*, which appeared more frequently in the

investigated coursebooks by Muller compared to the other discourse markers under study and was therefore used with more frequency and with similar functions by the German speakers.

3. In terms of the comparison between the frequency of use and the range of functions used by both groups of speakers, Muller found that the frequency of use and the range of functions used were significantly higher for the American speakers. This holds in contrast with the study of different discourse markers in other studies. For instance, in the study of the use of discourse markers by Chinese learners of English in comparison to L1 speakers of English, Fung and Carter (2007) found that the learners used certain discourse markers such as *I think* more frequently in their conversations when compared with their L1 counterparts.

The case of the occurrence of certain discourse markers, especially, the discourse marker, *I think*, and its high frequency of occurrence has been reported widely reported in other studies. For example, In the Persian context, in a corpus-based study comparing the Vague Language used by American, Chinese and Persian speakers of English, similar to Fung and Carter (2007), Sabet (2013) also found that both the Chinese and Persian speakers of English who were both learners of English used the expression *I think* significantly higher than the American speakers of English. Sabet (2013) observed that the learners in his study used *I think* to soften the blow of confrontations since the learners tended to use it more in clusters such as, *but I think* and *you know*, *I think*. Whereas the L1 speakers used it more for the purpose of agreement. However, Sabet (2013) does acknowledge that this use of the more frequent use of *I think* in the learner corpora in his study (i.e., a Chinese and Persian Spoken Learner corpora) was due to the learners adhering more to the use of *I think*

instead of other forms since L1 speakers had other expressions in their repertoire to use.

4. Muller (2005) also reported that German speakers used the discourse marker *well* for a function that was not used by American speakers, namely, to introduce a summary/conclusion and to continue expressing an opinion. In contrast, some functions did not appear in the German corpus, for example, the use of *you know* to introduce quoted material.

However, Muller's (2005) study does not encompass all the nuances of the use of discourse markers by learners. A frequently reported difference between the use of discourse markers by L1 and learners is the use of novel forms by learners. For instance, Zvěřinová (2016) found that the Czech speakers of English used *and so on so* which was not used in the L1 data. Similarly, in the context of Persian, as reported by Parvaresh et al. (2012, p. 277), Persian learners of English used VCMs such as "and and and" and "and this and that". These forms are in contrast to the conjunction + noun phrase/determiner phrase + (like that) of forming VCMs in English. Parvaresh et al. (2012) attributed this finding to transfer from Persian to English, since the equivalents of these forms were found in Persian, namely, *va va va* and *va in va an*.

As to the reason behind the differences between the use of pragmatic markers between language learners and L1 speakers, as Muller's (2005) study shows, several factors could influence this difference. Transfer from the learners' L1 has been considered as the main drive for these differences. A study that explores the possibility of the influence of L1 transfer on the use of pragmatic markers focusing solely on VCMs is Terraschke (2010). Terraschke (2007) found that in a corpus of conversations between German and New Zealand speakers of English, there was a significantly higher frequency of the use of VCMs by the German speakers. In addition. There was

also a significantly higher frequency of the VCM, *or so*, by the German speakers. This led to the study conducted by Terraschke (2010) which explored this form further.

Terraschke (2010) uses three corpora to compare the use of the English VCM, *or so* and its German equivalent *oder so*. The three corpora comprise of a total of 224,00 words. The first set of data is from New Zealand speakers of English, the second set is from German speakers of German and the final set is from conversations between German and New Zealand speakers of English. All the data were collected from participants who were considered to be "non-familiars". The data consisted of 60 dyads of conversation from 60 participants (30 German and 30 New Zealand). Each conversation lasted between 20-30 minutes and the participants were left in a room to interact and be recorded by a recorder. Terraschke concluded that L1 transfer (that is, from German to English) was the main factor behind this discrepancy. Similarly, Aijmer (2015) also found that certain forms were preferred more than other forms; therefore, leading to the use of a higher frequency of certain forms of VCMs by the learners compared to their native counterparts. As an example, she found that *or something* was the preferred choice of VCM for the Dutch, German and Swedish learners. Aijmer attributed this finding to the influence of the learners' mother tongue.

However, Terraschke also points to the need to consider each form separately to provide a more accurate description of the use of discourse markers by learners. In addition, Terraschke (2010) acknowledges that focusing solely on L1 transfer ignores all the other possibilities that are influential in the discrepancies between language learners and L1 speakers' use of discourse markers. These possibilities include "lack of attention to teaching pragmatic devices in formal language learning settings (Sankoff et al. 1997 and Muller, 2005), access to native speaker interactions (Bardovi-Harlig and Dorneyi, 1998), the learners' knowledge of L2 pragmatic conventions and

general linguistic proficiency (Kasper, 1998) and their ability to notice certain language features (Schmidt, 1990)" (p. 451). This assumption is based on the notion that L1 transfer is the most commonly reported. This assumption could be wrong because it can be the notion which is most easily discernible based on the methodologies used for such studies.

In addition, as Terraschke (2010) points out this could be the result of learner speech in general. For example, as she notes apart from the small number of occurrences of *and tralala*, there were no other forms of VCMs that were directly translated and used from German to English in cross-cultural communications. As an example, there was not a high frequency of the occurrence of the most frequent adjunctive in German, *und so (and so)* in the English data. However, it was with the disjunctive *oder so* that the learners failed to recognise the differences in use. Nonetheless should be noted that although "and so" did not appear as a frequent form in the German English data, other forms which did have equivalents in English, such as "and so on" (the equivalent in German being "und so weiter") did appear as the most frequent VCM in the German English-speaking data whereas it appeared with a significantly lower frequency in the NZ data. Therefore, the issue here is that when there is an equivalent there seems to be a mismatch between the use of the VCM.

Similarly, regarding the influence of “learner speech” in general, a comparison of the discourse markers used by learners and their L1 speaker counterparts leads to the notion that in terms of functions, they are far less used for interpersonal functions compared to other functions such as used as fillers or pauses for floor keeping devices. For example, Cheng and Warren (2001) found that Cantonese speakers of English used discourse markers to mark propositional uncertainty more than as politeness strategies. Similarly, Aijmer (2004) suggests that learners used discourse markers as

fillers and hesitation markers to buy time for planning their conversation instead of using them for politeness devices.

Regarding the notion of the frequency of adjunctive and disjunctive VCMs, while some studies may agree on the frequency of the VCMs, other studies differ. Terraschke (2007) found that German speakers used more variation in the production of disjunctive VCMs compared to NZ speakers of English. This variation may differ regarding which VCM is used. For example, Terraschke's (2010) study showed that the use of *or so* was different in terms of both the frequency and the functions of use in the New Zealand and German learner corpus. This difference was attributed to the "wider semantic scope of "or so" 's German equivalent, *oder so,* as evidenced by the German corpus (p. 40). The frequency of occurrence is more in the German data. However, the flexibility of the structure of the VCMs is more in the NZ data. This could be related to the learner data, but it could also be related to the finding from Overstreet (2005) that German VCMs are less flexible and varied compared to English VCMs.

A final note considers whether the mismatch in the use of "or so" led to any failures in communication. Terraschke (2010) found that as evidenced by the hearer's reaction, there seemed to be no breakdowns in the communication. She attributed this to two possible factors. First, the NZ speakers had enough contextual information to interpret what the German speaker had said.

The second reason that Terraschke presents contradicts her first reason. That is, she draws on Ostman (1982, p.153) by saying that since VCMs do not influence the propositional content of the utterance then they could be simply ignored by the NZ interlocutor. This is in line with the definition proposed for discourse markers in that they are not part of the propositional content of the utterance. Nonetheless,

Terraschke's (2010) is in line with other studies with learners conducted in intercultural settings (Cheng and Warren, 2007).

However, communication breakdown may also happen when vague forms are used to withhold information Although not specific to solely the function of withholding information, communication breakdown manifests itself in the form of a request for clarification by the addressee or the presentation of a less "vague" expression by the addresser. For example, Gassner (2012) looked at the use of vague forms by L1 and L2 users of English and stated that although L2 speakers' use of vagueness could lead to communication breakdown whereas this was not the case for L1 speakers. However, Parvaresh (2015), in a study of the use of vague forms by Persian speakers, demonstrates that communication breakdown does indeed happen in conversations between L1 speakers and is not specific to second language learners and intercultural settings.

Learners have also been reported to produce forms that are not used or recognised by L1 speakers. For example, Channell (1994) found that learners used what she referred to as non-standard" forms, such as "and that" in British English. Similar results have been identified in other studies. For example, Parvaresh, et al. (2012) found Persian speakers of English use VCMs formed by using only conjunctives, such as "va va va" (and and and), which have not been found in English.

Similarly, differences in the functions of VCMs used by learners and L1 speakers have also been found. Focusing on the Persian context, Parvaresh et al. (2012) reported a new function of this vague category in the Persian corpus but missing in the non-native speaker data: a general extender used by an interlocutor to express outrage at what another interlocutor had mentioned. Unlike native speakers of English who attached intensifying effects to general extenders, the Persian speakers did not assign this

function to the same category of vague language, either in their L1 or in English as an L2. The learners' dominant use of disjunctive general extenders was the result of uncertainty in word choice, which is a case that occurs with a very low frequency in the Persian corpus.

2.7. Chapter Summary

In this chapter, I have presented the definition and scope of Conversational Language used in this study. Since two features of Conversational Language, namely, discourse markers and Vague Category Markers are the main focus of this study, definitions for each of these features were also included. To set the context for the Persian Language, examples in this section were also drawn from studies on Persian discourse markers and Vague Category Markers. Finally, since this study explores learners' use of Conversational Language, in general, and the use of discourse markers and Vague Category Markers, specifically, this chapter concluded with an overview of the literature in this area. As this overview showed, there is a lack of research on the use of Conversational Persian by learners. Consequently, there is a lack of research on features more closely associated with the Conversational register, such as discourse markers and Vague Category Markers. The following chapter expands on this notion and provides an overview of the studies that have been conducted so far with learners of Persian.

Studies on Learners of Persian

1.1. Chapter Introduction

The current chapter aims to provide an overview of the studies on learners of Persian. This overview starts by first presenting the scope of these studies. This scope is followed by presenting a taxonomy of the most common features of language use attributed to learners of Persian. This chapter concludes by illustrating the situation of Conversational Persian in studies with learners of Persian.

3.2. Studies on Learners: Scope

The studies conducted with learners of Persian are divided into two main categories (Ghaffari, 2020). The first category belongs to studies on Iranian speakers of Persian who have Persian as a second language. That is, since Iran has L1 speakers of other languages, such as Turkish Azari, Laki, Kurdish, etc., these speakers start learning Persian in school or at later stages of life. The second category of studies with learners of Persian involves studies with adult learners of Persian enrolled in Persian courses. These studies are mainly conducted with learners of Persian enrolled in Persian (Ghaffari, 2020) university courses, such as the current study. Although these two categories are different regarding their participants, there are similar trends in terms of the focus of research and their findings. Therefore, an overview of studies in these two categories is presented in the following.

3.2.1. Iranian Learners of Persian

These studies are fewer in number compared with second language learners of Persian from other countries. However, similar to studies with non-Iranian second language learners of Persian, these studies primarily show three trends that are similar to their counterpart studies. First, these studies are conducted based on error analysis.

This error analysis involves judging learners' errors based on the intuition of one or two L1 speakers of Persian. This intuitive judgement mainly comes from the researcher (s) of such studies (Khanbabazadeh, 2016).

The second similarity is the lack of focus on the spoken language and, more specifically, on the Conversational register. For example, Golbaghi (2001) looks at the Persian of Laki speakers based on the error analysis of their vocabulary use in written assignments. Similarly, in written assignments, Ahmadian (2004) explores the errors of Kurdish speaking Iranians, Kamju (2011) looks at the morphological errors of Mazani speakers of Persian, and Khanbabazadeh (2016) reports on the syntactic errors of Taleshi speakers.

Finally, pragmatics is the less studied field in this area, and based on a survey of these studies Ghaffari (2020, p. 547) reports that the syntactic errors are much higher than morphological and semantic errors regarding their frequency of occurrence. Ghaffari also found that the source of the errors in such studies was reported to be from L1 transfer.

3.2.2. Non-Iranian Learners of Persian

Similar to the previous category, these studies also show a preference for using error analysis in written assignments. For example, Eslami (2013) looked at the errors of intermediate Russian speakers of Persian; Motevalian and Ostovar's (2013) explored the syntactic errors and Taherzadeh et al. (2016) looked into the morphological errors of intermediate Arab speakers of Persian; Motevalian and Malekiyan's (2014) focused on the syntactic errors of Urdu speakers of Persian; finally, Mirdehghan et al. (2014) focused on the errors of elementary level German speakers. The majority of these studies are based on James's framework of error analysis (1998). This framework categorises the errors of written language, specifically, into four main groups, namely,

spelling, mechanical, morphological and grammatical. The findings from these mentioned studies mainly point to the grammatical errors having the highest frequency regarding the occurrence of errors. Spelling, morphological and mechanical errors are respectively of lower frequency compared with the other categories (Ghaffari, 2020).

Regarding the source of errors, such studies have reported both interlingual and intralingual sources. For example, in the study with Urdu speakers of Persian, Motevalian and Malekian (2014) report that interlingual syntactic errors are more frequent. Amongst the intralingual semantic errors, the errors related to semantic relations are more frequent than the errors related to morphological collocation. Also considering the semantic errors related to interlingual errors, the frequency of the errors because of loan translations is much higher than the errors because of direct borrowing from L1. As for the syntactic errors, the errors related to verbs and prepositions have the highest occurrences, respectively. In addition, there are but there are some errors which are considered ambiguous as to whether they are interlingual or intralingual in nature. That is, these errors can be attributed to both mother tongue language interference and target language at the same time. However, similar to the previous category, the majority of such research shows that intralingual errors are of higher frequency (Ghaffari, 2020).

The point of divergence of the studies on non-Iranian learners of Persian compared with the Iranian Learners of Persian is the higher focus on studies on phonology (Falahati, 2020). That is, whereas the latter category of studies focuses mainly on syntax, phonological aspects of the learners' language production are more frequent in the latter case.

In a review of the studies on the phonological aspects of Persian Language learners' speech production, Falahati (2020) found that the majority of such studies were

informed by Contrastive Analysis (Lado, 1957) and the influence of the learners' L1 on their pronunciation in Persian. These studies were conducted with learners from various L1 backgrounds, including English (Majd, 2002); Russian (Bābāyi, 2014); French (Osati, 2015); Chinese (Sām, 2011); Turkmen (Qarebāqi, 2005); Italian (Osati 2015); Arabic (Sām, 2011); Kurdish (Dārābi, 2001; Mehdi Zādeh, 2008); Danish (Ābediyān Kāsegari, 2016); Azeri Turkish (Morādkhāni 2008); Spanish (Osati, 2015); and Urdu (Mirdehghān, 2009; Najafi Eskandari, 2016); Japanese (Moqadamkiyā, 2009; Hosseyni, Bijan Khān); Mandarin (Sadeghi and Mansoory Hararehdasht, 2016). Despite, the larger number of research in the area of phonology in second language learning of Persian phonology, Falahati (2018) reports the lack of research that looks at the L2 production of the learners. That is, these studies are gathered from experimental settings instead of looking at data from naturally occurring instances. This notion will be further explored in section 3.3. on Learners of Persian and Conversational Persian. However, in the following, I present a summary of the main features of the Persian Interlanguage using Ghaffari's (2020) study on learners of Persian with an L1 English background.

3.2.3. Ghaffari's Taxonomy of Features of Persian Interlanguage

For research on English speakers' interlanguage of Persian, considering different levels of proficiency and especially regarding higher levels of proficiency, Ghaffari's (2020) ongoing research has the largest number of participants, to date. Currently, Ghaffari's involves 128 students at the Elementary level (A1- A2), 69 students at the Intermediate level (B1- B2), and 27 students at the Advanced level (C1- C2). These students are between 19–25 of age (Ghaffari, 2020, p.549). Ghaffari's learners were learning the language in both academic and non-academic settings. He looked into

the interlanguage of the learners regarding its phonological, semantic, syntactic and morphological characteristics.

Phonological Aspects

Regarding the phonological differences, Ghaffari (2020) found 4 features for the interlanguage of English speakers of Persian. These features were attributed to L1 transfer and the over-generalisation of a particular rule. These features are outlined in the following.

1. Three Persian phonemes that do not exist in English and therefore produce difficulty for English-speaking learners. These phonemes are: /q/, /x/, /ʔ/
2. In contrary to the first character, there is a phoneme in English, namely, /w/, that does not exist in Persian; therefore, at times leading the learners to use this phoneme when pronouncing Persian words that are somewhat similar to words in English containing this phoneme. For example, the form for the 3rd person singular pronoun in the formal register is /vey/ in Persian which due to its similarity is pronounced as /wei/.
3. The third character is related to the phoneme /h/. The English speakers of Persian find it difficult to pronounce this phoneme when it appears in the mid or final syllable position. Therefore, words such as /mehr/ in Persian are pronounced as /mer/ by the learners.
4. The final characteristic is related to the overgeneralization of specific rules of pronunciation as in the case of /tʃe/ in Persian. Whereas /tʃe/ in pronounced as /tʃi/ in spoken Persian, this rule does not apply when a noun phrase is followed by /tʃe/ in a sentence expressing surprise or an interrogative sentence (Ghaffari, 2020, p. 550-551).

Morphological Features

Regarding the morphological Persian interlanguage 5 common features were presented for English speakers of Persian. These features are outlined below.

1. There are certain borrow words from English to Persian such as the case of sports names, such as *football* and *basketball*. However, the learners use English words for names of sports that are not borrowed from English in Persian such as the word for *chess* (*shatranj*).
2. There are certain words in Persian that do not have the same semantic mapping as in English. For example, in Persian, there are separate words that refer to maternal and paternal uncle (daei and amu, respectively) whereas in English uncle is used for both cases. Therefore, English speakers of Persian would use daei or amu interchangeably to refer to "uncle" as they would do in English.
3. In certain cases, some phrases are translated word-by-word when they have fixed idiomatic expressions. For example, "to look for something" is translated word-by-word into Persian (*baraye didan*) whereas there is a fixed expression for it, namely, "donbale chizi gashtan". The same applies to forming comparatives. For example, the comparative for beautiful appears frequently as *bishtar ziba* (more beautiful) whereas it should be zibatar, since the suffix –*tar* is used to form comparatives in Persian and not *bishtar* (*more*) as in English.
4. Confusion over when to use "harf zadan" and "goftan" which are two ways to form "say" in Persian. For example, "nemidoonam chi bayad harf bezanam" is a common phrase used in the interlanguage of English speakers of Persian.
5. The rules for forming comparatives are generalised. For example, to form the comparative and superlative in Persian the suffixes –*tar* and –*tarin* are added,

respectively. However, in the case of some words such as *khoub* (*good*), this is not the case. That is, the comparative of this word is *behtar* and the superlative is *behtarin* whereas, in the Persian interlanguage, these forms appear as *khoubtar* and *khoubtarin*.

Semantic Features

Regarding the semantics of Persian Interlanguage, two features are described.

First, is the use of *shodan* (to become) and *budan* (to be) interchangeably. This is attributed to the verb "to be" which is used for both purposes of *becoming something* gradually and *being something at the moment*. Second, the use of some words is the result of overgeneralisation. For example, the word *ghalam* is taught as an instrument for writing; however, the learners overgeneralise this to refer to *pencil* (*medad*) and *pen* (*khodkar*), which all have equivalents in Persian.

Syntactic Features

Ghaffari (2020) states the following as the four main features of Persian Interlanguage regarding syntax.

1. Word order: The usual word order for Persian is SOV and since English is an SVO language sometimes the learners use this structure to form sentences. Therefore, this feature of the interlanguage is attributed to L1 transfer. Another closely related area to word order is the order of adjectives and nouns. In Persian, adjectives follow nouns; however, in the Persian Interlanguage adjectives precede the noun. Ghaffari (2020) attributes this to transfer as well since in English adjectives precede nouns.

2. Verb omission: The equivalent of "to be" is frequently omitted in the Persian interlanguage.

3. Pronouns: First, there is an extensive use of the subject pronoun. This is reported to be the result of transfer since Persian is a pro-drop language whereas English is not.

4. Indefinitive marker /-i/. This feature can be divided into the three categories of omission, addition and overgeneralisation. In the first category, the indefinite marker is not added when needed. This is attributed to L1 transfer, since English does not have such a marker. In the second case, it is added when there is already a marker of indefiniteness such as the adjective “yek” (one). Third, it can be added to cases when the noun is definite and marked by determiners indicating this definiteness such as “in ketab”. This is attributed to the learning process.

Since syntax is the most widely studied area for learners of Persian (Falahati, 2020, p. 2), findings from other studies expand on the features from Ghaffari’s (2020) study. For example, Moore and Sadegholvad (2013) looked at the syntactic errors of heritage learners of Persian. Moore and Sadegholvad use Montrul’s (2012) definition of heritage learners as the children of immigrants born in the host country or immigrant children who arrived in the host country sometime in childhood” (p. 26). The heritage learners of Moore and Sadegholvad’s study were lower-level language proficiency university students enrolled in Persian courses. The results from their study showed that, despite their low proficiency, the learners showed no problems in comprehending and producing the spoken language, especially when compared with the non-heritage learners of Persian. This is a common feature for heritage learners of Persian which will be further expanded on in section 3.3.

However, a number of syntactic errors with the language of the learners were reported. The most significant error in terms of frequency was with producing compound verbs. As Moore and Sadegholvad (2013) explain, the majority of the verbs in Persian are

produced using an adjective, adverb or noun and a verb. Such as *nejat dadan* (*rescue*). However, the Heritage learners had difficulty in producing these types of compound verbs by drawing on the dominant (in this case English) to form phrases such as *nejat kardan* (p. 87). Therefore, there was an overreliance on the verb *kardan* (*to do*). Megerdoomian (2012) found similar results in her study of heritage learners of Persian, albeit with higher proficiency levels.

Moore and Sadegholvad (2013) attributed this overreliance to two factors. First, the majority of compound verbs are produced using *kardan* (p. 86). Thus, learners may generalise this notion. The other possibility that Moore and Sadegholvad (2013) put forward is that the verb *to do* is also considered to be a semantically empty verb in English. Therefore, there could also be the element of transfer from English to Persian involved.

Moore and Sadegholvad (2013) also found that the continuous and simple past tenses were used interchangeably by the learners (p.88). Moore and Sadegholvad attributed this to transfer from colloquial American English to Persian. This assumption was made based on the researchers' intuition and the interviews they had with their participants.

In this section, I have highlighted the most common features of Persian Interlanguage based on previous studies. However, there is one main deficiency with these features that reflects the scarcity of research on the use of the Persian Conversational register by learners of Persian. For example, as Ghaffari (2020) mentions regarding word order, learners tend to use the SVO word order instead of the SOV due to transfer from English, which also has the SVO order. However, the Persian word order is flexible, especially in Conversational Persian (Karimi, 2008). Therefore, differences in

word order do not necessarily signify a feature that is exclusively characteristic of the learners. Therefore, the following section looks at studies that have considered Conversational Persian, specifically.

3.3. Learners of Persian and Conversational Persian

As mentioned in section 1.2.3., studies specifically focusing on the use of Conversational Persian by learners remain scarce. Nonetheless, with the growing scope of the use of Conversational Persian, this register of Persian is starting to gain attention in the field (Shabani-Jadidi, 2021). However, this attention is mainly directed towards a comparison between second language learners and heritage learners of Persian.

There are different definitions for heritage speakers (Shabani Jadidi, 2018); however, in the case of Persian, and especially when compared to second language learners, definitions all point to a higher exposure to Conversational Persian (Sedighi, 2010). Therefore, heritage speakers of Persian are reported to "often demonstrate native-like pronunciation, are typically able to carry out conversations on everyday topics in Persian, can understand rapidly spoken conversational language, and are familiar with the sociocultural behaviour of the heritage language community" (Shabani, Jadidi, 2021, p. 55). In addition, Sedighi (2009) also found that heritage learners outperform second language learners in various ways including their performance on formulaic speech and language chunks.

Subsequently, heritage learners have also been reported to demonstrate the disadvantage of "high level vocabulary and have difficulty moving from one register or variety of the language to another in order to use contextually appropriate language." (Shabani- Jadidi, 2021, p. 55). Megerdoomian (2009) states that this once again highlights exposure to Conversational Persian. That is, since heritage learners are

more exposed to the Conversational register through their family background, they tend to use the features associated with Conversational Persian more and consequently transfer these features to other contexts.

These features of Conversational Persian are in turn reflected on only certain aspects of this register, namely, morpho-syntactic features associated with the level of formality of this register, such as the plural verb endings *–id* and *in* for formal situations and the singular verb ending *–i* for informal situations. Therefore, other aspects of Conversational Persian, especially features that are closely associated with this register, such as the use of discourse markers and Vague Category Markers remain to be explored.

3. 4. Chapter Summary

This chapter provided an overview of the studies on learners of Persian. As this overview showed, although there is a growing number of studies on the language use of learners of Persian, studies specifically focusing on the use of Conversational Persian by learners are in the early stages of receiving attention. In addition, the existing studies mainly focus on certain aspects of the language, with a high preference for exploring the morpho-syntactic features of learner language. Therefore, there is a lack of studies on the pragmatic interlanguage of Persian.

4. Methodology Chapter

4.1. Introduction

This chapter is divided into five main parts. First, this chapter describes the learner corpus compiled for the purpose of this study. This description mainly incorporates two sections: the factors considered for the design of the learner corpus, and the measures applied for compiling the corpus. Next, the ethical considerations applied to this study are discussed. This section explains how these ethical considerations were used for all the stages of data collection and data management. The third part of this chapter centres around the transcription of the learner corpus. The choice of the transcription script and a description of the transcription conventions constitute the two main parts of this section. The fourth section of this chapter describes the reference corpus used in this study. Finally, this chapter concludes with a description of the main corpus tools used for analysis.

4.2. Learner of Persian Spoken Corpus (LoPSC)

Corpus Linguistics as a methodology involves the collection of texts which are designed and compiled to represent a certain variety of languages with the aim of answering specific research questions (McEnery et al., 2006). There are various types of corpora available. One of the widely used corpus types in the field of second language learning and teaching is the learner corpus. For the compilation of a learner corpus, written and/or oral productive data is collected from learners of a language. Since their emergence in the 1980s, learner corpora have provided insight into language learners' actual language use as opposed to other predominant research

methods such as the use of elicited data from learners collected in experimental settings. Therefore, learner corpora have provided and continue to provide new perspectives and invaluable insights about the difficulties that learners face when learning a new language. Despite the importance of learner corpora in providing a rich resource of data, learner corpora for learners of the Persian Language remain scarce. In the following, the considerations taken for the design of the corpus, and the process of compiling the corpus are explained.

4.2.1. Design of LoPSC

The design of any corpus reflects the research purpose(s) (McEnery et al., 2006). The purpose of this study was to find the forms used by learners of Persian in spoken colloquial Persian. Therefore, the corpus type is a learner corpus (i.e., a corpus representing the interlanguage variability), the target language of the corpus is Persian, and the mode of the corpus is spoken. Whereas the choice for the corpus type and target language is self-explanatory, a note should be made here regarding the choice for the spoken mode.

Corpora are presented in written, spoken or multi-modal modes or a combination of two or all modes together (O'Keeffe and McCarthy, 2020). Since including as much context as possible helps in the analysis of the spoken form, Rühlemann (2019) advises using multi-modal corpora for the analysis of the spoken form. However, the use of multi-modal corpora comes with its own set of disadvantages. Knight (2009, p. 97-98) refers to two shortcomings of the use of multi-modal corpora, namely, exacerbating Labov's (1972) notion of the Observer Paradox and the restrictions placed on participant movement. As for the notion of Observer Paradox, although this applies to all studies where participants are being recorded, as Knight (2009) notes,

Observer's Paradox is amplified when in addition to voice recorders, video cameras are also added for the purpose of compiling a multi-modal corpus.

In addition to the effect that video cameras have on Observer's Paradox, to allow video recording to take place, participant movement would need to be restricted. As will be explained further in the section on Authenticity (See 4.2.1.2.) of the conversations and the section on the Main Phase of Data Collection, the aim of the study was to capture near-authentic conversations between the learners. Therefore, setting restrictions such as time, location, number of interlocutors, etc. were avoided. Hence, limiting participant movement to enable video recording would also decrease the chances of near-authentic conversations occurring.

An additional reason that can be added to Knight's (2009) observation on the possible disadvantages of multi-modal corpora, is the issue of the impracticality of using video recorders in all settings. That is, installing cameras in all locations would not have been feasible. For example, one of the conversations took place in a park, another conversation in a café, and the remaining conversations took place in communal areas in different university locations. In such settings, using cameras would only be possible if numerous permissions were granted beforehand, and the consent of others in addition to the participants was ensured (See section 4.3).

Therefore, although the advantages of multi-modal corpora in analysing spoken corpora are fully acknowledged, due to the disadvantages mentioned, in the initial phases of this study, the decision was made to opt out of the compilation of a multi-modal corpus. However, although contextual information such as facial expressions and gestures are left out of this study, instances in which participants refer to objects in their setting, such as showing a photo on their mobile phone, were marked and

included in the transcription of the recorded data (See section 4.4 on Transcription for further explanation on this issue).

Regarding, the other criteria taken into consideration for the design of LoPSC, these criteria both reflect the criteria involved in designing corpora in general and the criteria involved in designing learner corpora, specifically. The two criteria discussed here are the authenticity and representativeness of the LoPSC.

4.2.1.1. Authenticity

Gilquin (2021) believes that some genres are difficult to achieve in designing and compiling learner corpora. For example, she mentions having a spontaneous conversation with a friend would be difficult to capture in learner corpora, because this would more likely happen in the learners' first language than in the intended target language of the learner corpus. Therefore, Gilquin (2021) believes that authenticity in communication, as defined by Sinclair (1996) as being "genuine communications of people going about their normal business" is not possible in learner corpora. Therefore, Gilquin states that what is authentic to the classroom setting, and consequently in learner corpora, should be redefined. For example, Gilquin points out that tasks such as role playing and writing argumentative essays are authentic to the language classroom but not authentic elsewhere.

However, Gilquin's (2020) statements, which represent the trend in learner compilation, can be rejected in certain contexts. For example, in the case of writing argumentative essays, first language speakers may also engage in writing argumentative essays. Therefore, the authenticity of certain tasks is not necessarily confined to language classrooms and learner corpora.

More specific to this study is Gilquin's (2020) claim on the unachievable task of collecting spontaneous and informal conversations from learners of a language on the

basis that learners would not converse in their second language to conduct informal conversations with friends but would resort to their first language for this purpose. However, this is not necessarily the case, since different learners may have different first language backgrounds and may still be friends; therefore, conversations among friends may happen in the second language. Therefore, Gilquin's (2021) statement that learner corpora only reflect a certain number of limited genres is not a reflection on the nature of interlanguage, but only the limitations that learner corpora compilers place on interlanguage.

As for the learners of this study, although they did share a common first language (English); as a demand from their program, they were highly encouraged to speak in Persian outside of the classroom Therefore, even prior to data collection of this study, the participants would have conversations in Persian among themselves outside of the classroom.

Overall, in the case of this study, although there is an acknowledgement that factors, such as the Observer's Paradox mentioned in 4.2.1, subtract from the authenticity of the collected data, holding informal conversations in Persian outside of the classroom was not introduced by this study. To further ensure the authenticity of the data, the choice of their interlocutors, the topics of discussion, and the amount of time and location for each conversation were also decided on by the participants. The topics decided by the participants included but were not limited to topics on holidays, future plans, family, and politics. (See the section on the Main Phase of Data Collection for more information on how each conversation took place.)

Having discussed the issue of authenticity in the LoPSC, the following section provides a description of the other criterion used in this corpus, namely, representativeness.

4.2.1.2. Representativeness

Leech (2007) compares attaining representativeness in corpora to obtaining the Holy Grail; that is, although having a representative corpus would be impossible to realize, attempts are still made to achieve it. The issue of representativeness in corpora, regardless of the type of corpus is important, because every corpus is a sample of a language variety (O'Keeffe and Carter, 2020). That is, any language variety can be considered as the population of the study, and since studying the entire language variety or population is usually not possible, studying a sample of the population in the form of a corpus remains the only feasible option (Biber, 1993). Therefore, ensuring that this sample of the population is as representative as possible will conclusively lead to making more sound claims based on corpus findings.

As for this study, the population would be all the spoken forms of Conversational Persian produced by learners of Persian. The sample corpus compiled to represent this population was the spoken forms of Conversational Persian produced by 18 advanced learners of Persian. At the time of data collection, these learners were in year three or four of Iranian studies or a similar program in a UK university. (Table 4.1 in the section on the Main Phase of Data Collection shows the information on the individual participants.)

The choice of the number of participants was related to the context of the study, which will be further discussed in the section on data collection. However, the other two remaining criteria, namely choosing advanced learners of Persian and university students in their third or fourth year of study, are related to the design of the corpus and therefore, will be discussed in the following.

As Granger (2021) states, interlanguage is heterogeneous, since it involves a wide range of variables; therefore, making certain variables, such as the proficiency level

of the learners, constant could help with making claims and generalisations about the findings of learner corpora. However, in this study, the reason behind choosing language proficiency as a constant variable was not to make claims for a certain proficiency level. The reason for choosing learners from an advanced level was interconnected with choosing students in their third or fourth year of undergraduate study. That is, in courses for teaching Persian, the teaching of Standard Persian precedes teaching of the spoken Conversational form; therefore, it is in the advanced level or in their third or fourth year of study that students are introduced to Conversational Persian. Hence, choosing a cohort of first- and second-year students who have not been introduced to the forms of Conversational Persian could have lowered the chances of observing the use of spoken Conversational forms by learners, and, consequently, undermined the representativeness of this study.

Another issue that is closely related to the issue of representativeness of corpora is the size of the corpus. Corpus Linguistics as a field is associated with quantitative analysis, and traditionally a larger size corpus was and still is associated with a higher probability of the corpus being representative for the sake of generalisability of the corpus (McEnery, Brezina, Gablasova, & Banerjee, 2019). Nonetheless, similar to other aspects involved in the process of corpus design, the decision made for the size of the corpus reflects the research purposes. For example, with general corpora, such as the 11.5 million word corpus of British National Corpus (BNC) 2014(Love, Dembry, Hardie, Brezina, & McEnery, 2017), in line with the BNC 1994 (Leech, 1993), the aim of the corpus designers was to compile a well-balanced corpus ensuring that the BNC was truly representative of the entirety of the contemporary British English Language in all forms of production and all genres.

In the case of learner corpora compiled in large sizes, as an example, the Trinity Lancaster Corpus (Gablasova, Brezina, & McEnery, 2019) can be referred to. This 4-million-word corpus was compiled mainly with the aim of including as many learners of English with different L1 backgrounds as possible. The compilers of Trinity Lancaster corpus, in line with the traditional approach to corpus linguistics, designed this corpus with the assumption that having larger amounts of data would make the claims based on the findings of this corpus easier to justify.

On the other end of the spectrum regarding corpus size, are smaller-sized corpora, such as O'Keeffe's 55,000-word Irish Radio Phone-in Show Corpus (O'Keeffe, 2004). Such corpora add to the value of more quantitative corpus tools, such as Keywords and collocations, by making use of one of corpus linguistics' most valuable, and often overlooked properties, that is, the provision of co-text which helps with enabling a more nuanced analysis of the data and adds a qualitative dimension to analysing data with corpora (Aijmer & Rühlemann, 2015; O'Keeffe, Clancy, & Adolphs, 2020). In line with such studies, this study also used this method of analysis.

In addition to allowing for an in-depth exploration of the co-text with corpus linguistics, a small-sized corpus also allows for a manual inspection of the corpus data. In the case of this study, which looked at an underrepresented language variety, a manual inspection of the data in addition to the use of corpora tools would allow for noticing forms that would be difficult to notice otherwise. For example, in the case of spoken Persian, a relevant example would be the use of rhyming vague forms; that is, the combination of a word and a rhyming non-word, such as *ketab metab* (*book metab*) (See Section 2.5.3). Although these forms are highly frequent in colloquial spoken Persian (Ghomeshi, 2018; Parvaresh & Tayebi, 2014), they would be impossible to notice without a manual search of the corpus. Therefore, since the main aim of this

study was not to pursue the generalisability of the findings, but to present a snapshot of an underrepresented language variety, namely, the spoken Conversational forms produced by learners of Persian, a manual search of the data would allow for forms such as rhyming vague forms to be searchable.

A final note on the size of the corpus is that although the size of the corpus has been talked about in terms of its importance, a specific guideline does not exist on what would constitute an ideal size for the corpus in number of words. The nearest to such a guideline is Biber (1993). Based on a set of statistical tests, Biber states "frequency counts for common linguistic features are relatively stable across 1,000-word samples"; therefore, for studying words that are considered highly frequent, a small size corpus would be adequate. In the case of this study, as will be shown in the analysis section, high frequency occurring words in Conversational speech are observed; therefore, a small-sized corpus would suffice for the purpose of this research.

Having discussed the criteria applied for designing the LoPSC Table 4.1. summarises the elements involved in the design of the LoPSC.

Table 4.1.

Design Criteria for the LoSPC

Type	**Mode**	**Target Language**	**Task Type**	**Topic**	**Level of Proficiency of learners**
Learner Corpus	Spoken	Persian	Informal Conversations	Various everyday topics	Upper intermediate and advanced

The following section describes the next stage of compiling this corpus, i.e., data collection.

4.2.2 Data collection for the LoPSC

This section describes the steps and procedures taken for compiling the corpus, which were: requesting for participation, which will be discussed under sampling; conducting a pilot study; and the main stage of data collection.

Sampling

As was discussed in the section on the design of the corpus, the aim of the study was to collect spoken data from conversations in colloquial Persian between learners of Persian enrolled in their third or fourth year of undergraduate studies in an Iranian Studies program or other similar programs.

This study was based in the UK. This decision was made to ensure that two conditions were met. First, the spoken data was recorded using the device provided to assure the comprehensibility of the audio. Second, and most importantly, to facilitate enabling some form of rapport with the participants. The reason for seeing the need for this was to compensate for the lack of shared background knowledge that comes with being an outsider to a group. Although this shared background knowledge plays an important role in analysing spoken conversations, texts from corpora are unable to provide this insight making this one of the limitations of corpus-based analysis (Ruhlemann, 2019). Therefore, with the aim of taking a step to overcome this limitation, one to two-hour in-person meetings were organised with the participants. It is worth noting that the date, time, duration and location of these sessions were organised by the participants. In addition, the participants believed these sessions would help them to gain additional practice in speaking in Persian. Hence, these sessions were mutually beneficial to both sides, with the participants being able to get the practice in Persian they thought

would help them, and for the analysis of the study to benefit from a better understanding of the world of the participants outside of the limited space of the corpus.

After using various channels of asking for participation, such as contacting heads of relevant schools, tutors on Persian courses, and individual students, 20 participants, out of a total of 23 students who met the requirements for participation, showed interest in participating in the study. Two of these students took part in the pilot study (see section on Pilot Study for further information) and the remaining 18 students provided the data that was used to compile the LoPSC. Relevant metadata gathered from the students after the recording of each session is provided in Table 4.2 (See Appendix I and Appendix II for the information sheet and consent form used for collecting this data from the participants).

Table 4.2.

LoPSC Participant Information

Speaker	L1	Age	Gender	Other Languages Spoken	Travel to Iran (if applicable, duration of stay)	Average number of hours practicing spoken Persian with a L1 Persian speaker
1	English (British)	25	F	Arabic (advanced), French (advanced). Spanish (advanced), German (advanced)	Yes (2 months)	1 hour
2	English (British)	22	F	French (A-levels), Arabic (intermediate)	Yes (numerous visits, total of 10 months-longest stay: 3 months)	7 hours

3	English (American)	22	F	French (lower-intermediate), Arabic (intermediate)	No	3 hours
4	English (British)	24	M	French (upper-intermediate), Spanish (upper-intermediate), Arabic (upper-intermediate), Afrikaans (upper-intermediate)	No	1 hour
5	English (British)	22	M	French (intermediate)	Yes (2 months)	2 hours
6	English (British)	22	M	Arabic (intermediate), Spanish (upper-intermediate), French (intermediate)	Yes (2 months)	3 hours
7	English (British)	21	F	Arabic (upper-intermediate), Spanish (intermediate), French (intermediate)	No	2 hours
8	English (British)	21	F	Arabic (intermediate), Ancient Greek (basic reading and writing), Urdu (upper intermediate speaking), Spanish (upper intermediate), Gujarati (only spoken, advanced)	Yes (2 months)	3 hours
9	English (British)	22	F	Arabic (intermediate), Spanish (upper-intermediate), French (intermediate)	Yes (2 months)	3 hours
10	English (British)	22	F	Arabic (intermediate), Spanish (upper-intermediate), French (intermediate)	Yes (2 months)	3 hours
11	English (British)	22	F	Arabic (intermediate), Spanish (upper-intermediate), French (intermediate)	Yes (2 months)	3 hours
12	English (British)	22	F	Arabic (intermediate), Spanish (upper-intermediate), French (intermediate)	Yes (2 months)	3 hours
13	English (British)	22	F	Arabic (intermediate), Spanish (upper-intermediate), French (intermediate)	Yes (2 months)	3 hours
13	English (British)	22	M	Arabic (intermediate), Spanish (upper-intermediate), French (intermediate)	Yes (2 months)	3 hours
14	English (British)	21	F	Arabic (upper-intermediate), Spanish (intermediate), French (intermediate)	Yes (2 months)	2 hours

15	English (British)	22	F	Arabic (upper-intermediate), Spanish (intermediate), French (intermediate)	Yes (2 months)	2 hours
16	English (British)	21	F	Arabic (upper-intermediate), Spanish (intermediate), French (intermediate)	No	4 hours
17	English (British)	23	F	Arabic (upper-intermediate), Spanish (intermediate), French (intermediate)	Yes (2 months)	2 hours
18	English (British)	22	F	Arabic (upper-intermediate), Spanish (intermediate), French (intermediate)	No	3 hours

Before moving on to the main stage of data collection, a pilot study was conducted. The following section provides the reasons for seeing the need to have a pilot study and a description of how this pilot study took place.

Pilot Study

A pilot study was used for two main reasons: to check the recording devices for audio quality and to see how the participants would interact with the given instructions.

Of the 20 participants, 2 of the students who were in their fourth year agreed to take part in the pilot study. Both participants were male and between the age of 21 to 24. The participants had been classmates and friends for four years. In line with the data collection stage, the time, date and location of the recording were chosen by the participants.

Participants were first asked to go through the information sheet describing the project (See Appendix I) and to sign the consent sheet (See Appendix II) if they agreed to take part in the pilot study. In addition, to the instructions being noted in the information sheet, they were also once again informed that the data collected from the pilot would not be used in the main study, and it would only serve the purpose of better informing the data collection stage of the study. It was also reiterated that the data would be

securely stored and disposed of after three months (See section on Ethical Considerations, Data Management and Storage for further information).

After the learners were given the time to ask any questions related to the pilot, the consent forms were collected, and the recording device was set in place and tested. After ensuring the device was working, the participants were then left alone to start their conversation. To avoid interrupting their conversation, the participants were asked to make a telephone call to me once they felt that their conversation was over. Their conversation lasted for approximately an hour.

Following their conversation, a semi-structured interview followed where the participants were asked to talk about their experience with the session. This interview was not recorded and only notes were taken from this interview. The interview lasted for about 30 minutes. During this interview, the participants were asked the following main questions:

1. Overall, how did you find the experience?
2. How did you find the instructions given before the recording?
3. How did you find the experience of having a conversation while being recorded?

Based on the answers to these questions, the participants of the pilot study had a positive attitude toward their conversation session. However, a comment was made by one of the participants about the size of the recording device. The participant mentioned that the recorder made the session “look like an exam” due to its size. However, the participant mentioned that after he had started the conversation, he began to forget he was being recorded. Nonetheless, the recording device was changed to a smaller device for the main data collection phase.

Having described the pilot study, and the changes that were made to the main phase of data collection based on the interview from the pilot study, the following section describes the data collection for the LoPSC.

4.2.3. Main Phase of Data Collection

The process of data collection was similar to the pilot study; that is, the participants allocated a date, time and location for their conversations to take place. The participants also chose whom they preferred to have as their interlocutor during the conversations. They also determined the duration of the conversations.

Before recording the sessions, participants were given the information sheet describing the project and the process of data collection (See Appendix III). They were also given the time to ask any questions before the start of the recording. Participants were then asked to sign a consent form (See Appendix IV) if they agreed with the data collection to proceed.

A total of 8 sessions of recording took place. The information on these sessions containing the number of participants, their gender, age, and location of the conversation can be found in Table 4.3.

Table 4.3.

LoPSC Recorded Session Information

Sessions	Participants	Session Duration (minutes)
1	Speaker 1, Speaker 2, Speaker 3	31
2	Speaker 4, Speaker 5	30
3	Speaker 6, Speaker 7	51

4	Speaker 8, Speaker 9	50
5	Speaker 10, Speaker 11	63
6	Speaker 12, Speaker 13, Speaker 14	66
7	Speaker 15, Speaker 16	48
8	Speaker 17, Speaker 18	54

The audio-recorded sessions were then copied from the recording device to a password-secured OneDrive folder. The audio files on the device were then deleted. Information on the measures taken for data storage and management is described in the following section on ethical considerations.

4.3. Ethical Considerations

Throughout this study, ethical guidelines based on the British Association for Applied Linguistics (BAAL, 2016) and British Educational Research Association (BERA, 2018) were adhered to. Two of the main criteria in both these ethical guidelines are to ensure that the consent of the participants has been obtained prior to conducting the study and to respect the participants' will to withdraw from participation in the study. Both these criteria were met in this study: participants were given consent forms to sign before beginning the recording, and the participants received a written statement and verbal instructions that they could withdraw from the study at any stage of the study (i.e., prior, during and after data collection). Participants were also given a copy of their signed consent forms.

Another main and mutual criterion in both BAAL (2016) and BERA (2018) is ensuring participant anonymity and confidentiality. The anonymity of participants is achieved by

not using their real names (BAAL, 2016, p. 5). In this study to achieve participant anonymity, all the participants were referred to as “Speaker 1”, “Speaker 2”, …, “Speaker 16”.

In the case of confidentiality, which is to ensure participants are not identified in any way (BAAL, 2016, p. 5, 6), the following steps were taken:

1. Third parties mentioned in the conversations were anonymised using “Name of person 1”, “Name of person 2”, etc. The reason for choosing to assign numbers was to prevent distorting the meaning of the conversations. That is, when several third parties were mentioned in one conversation, to allow distinguishing between these individuals, numbers were also assigned.
2. Name of locations mentioned were also anonymised using “Location 1”, “Location 2”, etc. The same logic explained above for using numbers in addition to the anonymised label “Location” applies here. These measures were taken to ensure the confidentiality of the data did not come at the cost of causing confusion during data analysis.

To clarify, the encoding used for the purpose of anonymisation did not interfere with the data analysis for two reasons. First, for the transcription of the data, the Persian script was used, and since the encodings used for anonymisation and confidentiality appeared in English, these encodings could be easily noticed and taken out, if they appeared in the results (See Section 4.4 for further explanation of the transcription of the data). In addition to the differences in the scripts used for anonymisation and the transcription of the data used for analysis, the encodings were also included in angle brackets. This measure was taken since the corpus analytical software used in this study (#LancsBox 6.0) treats tokens placed in angle brackets as tags and does not

include such tokens in the data analysis (See Section 5.2. on Keyword Analysis for further explanation on the analysis used in this study.)

Regarding data management, this differed based on the type of data collected. The collected data, in this study, can be categorised into four categories: recorded conversation from the pilot study, notes from the pilot study interview, recorded conversations for LoPSC, and metadata collected from participants. As mentioned in the section on the Pilot Study, the recorded conversations and the interview notes from the pilot study were stored in a password-secured OneDrive folder. These two sets of data were deleted after the aims of the pilot study were met. (See section on Pilot Study for further explanation on the aims of the pilot study.)

As for the recorded conversations for LoPSC, the audio files were also stored in the same password-secured OneDrive folder. The metadata collected from the participants with their real and anonymised names were also stored in the same folder.

Finally, as a requirement from the University of Edinburgh's Moray House School of Education and Sports, the Research and Knowledge Exchange Ethics Committee of this school approved this study.

4.4. Transcription

The transcription of the data is divided into two sections. The first section involves a discussion of the paralinguistic features included in the transcripts. The second section addresses the choice of orthographic script used in this study.

4.4.1. Paralinguistic Features

The paralinguistic features that were added to the transcription of the data were chosen for two reasons. First, to avoid the loss of contextual cues that could provide

insight during the analysis of the data, certain paralinguistic features were included in the transcription of the audio data, such as the inclusion of laughter.

In addition to providing contextual cues, the inclusion of certain paralinguistic features was required to answer the research questions of this study, specifically regarding research questions 3 and 4 on discourse markers and vague category markers. For example, pauses are considered in the transcription because they are considered important for distinguishing discourse markers from other forms (Fung and Carter, 2007). Similarly, the inclusion of pauses helps to distinguish vague category markers from phrases with similar structures but with different functions (Channel, 1994). Therefore, pauses were also included as paralinguistic features in the transcription of this study. To avoid the inclusion of paralinguistic features when using corpus tools for statistical purposes similar to the encodings used for anonymisation and confidentiality, all paralinguistic features were included in angle brackets. See section 4.2.4. for the rationale for using angle brackets.

4.4.2. Choice of Script

In corpora of spoken Persian, audio transcriptions of recordings generally appear in three formats: the Persian script, the Romanised Persian version of the Persian Script or the phonetic alphabet. As an example, the extract below shows a sample from transcribed audio recordings of the spoken corpus CallFriend Farsi (Graff, et al., 1970). Audio recordings in CallFriend Farsi have been transcribed using the Persian Script and a Romanised version of the Persian Script, as shown in the extract below.

Extract: Examples of transcriptions of Persian using the Persian and Romanised version of the Persian script

A: آره ویدیُوش هم بود همه چیز بود

A: *Are vidiyoSH ham bud hameCHi bud*

Similar to other aspects of designing a corpus, the choice of transcription depends on the research questions that the corpus aims to answer. However, comparing two corpora with two different orthographic transcription conventions and scripts would be impractical and, in certain cases, impossible. Since learner corpus-based studies fall into the category of comparative studies, the scripts used for the learner and reference corpus should be identical to facilitate comparisons and query searches across the two corpora.

In addition, to ensure that the script used for transcribing the audio recordings is identical for learner corpus-based studies, there are also obstacles that would arise when working with the Persian language, and Conversational Persian, specifically. These obstacles are, but under no circumstances limited to, the innate ambiguity in the Persian script and differences between colloquial Persian and "standard" Persian. Each of these obstacles is described in brief below.

4.4.3. Ambiguity in the Persian Script

Due to its nature, the Persian script may lead to ambiguities in recognising the written form. One of the main causes of this ambiguity is omitting diacritics in the Persian Script. These diacritics act as replacements for certain vowels in Persian. For example, the extract above from CallFriend Farsi shows the diacritic ُ replacing the vowel /o/. Therefore, omitting them would lead to ambiguities in recognising the written form. For example, Ghayoomi, et al. (2022, 29) give the example of the form کند in Persian. Without the diacritics, this form can be recognised and pronounced as any of the following forms /*kand*/ '*picked'* (Verb, Past Tense), /*kanad*/ 'picking up' (Verb, Present Continuous Tense), /*konad*/ '*doing'* (Verb, Present Tense), /*kond*/ '*slow'* (Adv), and /*kond*/ '*blunt'* (Adj).

In learner-corpus-based studies, transcribing recordings using the Persian script would lead to two problematic areas. First, performing a search query using any of the corpus tools would not be possible. Second, idiosyncrasies of learner language will not be picked up in transcripts.

One solution to the ambiguities of the Persian script would be to use the Romanised script for representing Persian sounds. However, using the Romanised Script for the Persian Language carries one main disadvantage in that there is no agreed-upon standard form between transcribers in presenting Persian in the Romanised Script. As an example, the /u:/ sound in Persian is found to be transcribed in various forms, such as "*oo*", "*ou*", etc. As shown in the extract above from CallFriend Farsi, /u:/ is transcribed as "*u*". Therefore, although using the Romanised scripts can be convenient in many ways, the lack of an agreed upon standard form for this script would also lead to the same problems associated with the Persian script, namely, difficulties in performing query searches with corpus tools, especially in comparative studies such as this current study.

The decision to use the Persian Script was based on two reasons. As Adolphs (2010) states, decisions formed for the transcription of the data are best chosen based on the research questions and the reusability of the transcripts. Since as with other spoken corpora, the transcription of audio data is a time-consuming process, although the decisions made for the transcription of LoPSC were made primarily with the research questions in mind, the reusability of the transcripts was also considered as an important factor in the final decision for choice of the orthographic system. Therefore, the choice of the Persian script over other scripts would allow for more comparisons of the LoPSC with other corpora.

4.4.4. Differences between Conversational and written Persian

Although its usage is expanding, Conversational Persian generally refers to the language variety used in everyday spoken conversations. Colloquial Persian is placed in contrast to “Standard” Persian which is the form used in teaching Persian in Iranian schools and Persian language learning textbooks. The “Standard” and colloquial form of Persian show a variety of differences especially in the phonological and morphological representation of words. As an example, depending on the context in which it occurs, the equivalent for *if* in Persian (اگر) appears in various forms in Persian, namely, *agar* in “Standard” Persian and *age* in colloquial Persian (Ghayoomi et al 2021). In addition, more than one form of the same word can be used in colloquial Persian. For example, the equivalent for *and* in Persian may appear as *va* and *o* in everyday conversations.

Another difference between Conversational Persian and the “Standard” form of Persian is the substitution of the phoneme /u:/ for the phoneme /a:/ in the LoPSC. This substitution is a common feature of Conversational Persian (Miller, 2011), and it is reflected in the written script. For example, the equivalent of *bread* in Persian, *nan,* appears as /na:n/ in the Standard form and /nu:n/ in the Conversational form. This is also reflected in the Persian script as نان and نون in Conversational Persian.

In corpora of spoken Conversational Persian, transcribers are faced with multiple options regarding the choice of script. One option would be to transcribe the words in verbatim. In such a corpus, *va* and *o* would both occur. This type of approach to transcribing colloquial Persian would lead to the same problems associated with the ambiguity of words in the previous section, namely, difficulties in search queries and annotating the corpus. A solution would be to use one form for the representation of all the forms regardless of their pronunciation. Therefore, the form for *and* would be

transcribed as either *va* or *o* across the entire corpus. This was the option for transcription that was opted for in the current study.

4.5. Reference Corpus: Corpus of Conversational Persian

4.5.1 Rationale for Using a Reference Corpus

The majority of studies with learner corpora use a corpus of L1 speaker language, which is also referred to as the reference corpus, in their method of analysis (Flowerdew, 2015, p.469). The preference for the use of this method is to identify non-standard forms or errors and to identify items that are “underused” or “overused” (Granger, 2015, p.19).

The concept of the “overuse” and “underuse” of certain forms by learners of a language has received much criticism. As an example, Gries and Deshors (2014, p.114) believe this to be a “decontextualised” way of looking at interlanguage. However, this study takes the stance of Granger (2004, p. 132) in that this form of analysis “overuse” and “underuse” “are not meant as being evaluative but purely descriptive”. In addition, as Granger observes “the study of over- and underuse has been a real eye-opener in learner corpus research, because it has shown that the foreign-soundingness of learner language, especially at advanced levels of proficiency, is to be attributed as much (or perhaps even more) to differences in the frequency of use as to downright errors” (Granger, 2004, p.132).

In addition, Gries and Deshor’s (2014) concept of decontextualised overuse and underuse does not apply to this study. This is due to the small size of both the reference corpus and the LoPSC, which enables the manual observation of the concordance lines, and hence, avoids falling into the pitfall of “decontextualising” the data by allowing individual forms to be considered within the co-text of occurrence.

Having explained the reasoning behind including a reference corpus in the analysis of this study, the following section provides a description of the chosen reference corpus for the analysis.

4.5.2 Description of the Corpus of Conversational Persian

Table 4.4 shows the list of the spoken Persian corpora available. This table also includes the text types used for the compilation of each corpus. The size of each corpus is also presented as stated by the designers; that is, in number of words, tokens or number of hours of recording, etc.

Table 4.4.

List of Available Spoken Corpora in Persian

Name of Corpus	Medium	Text type/Task type	Size
FARSDAT (Bijankhan et al., 1994)	spoken	Read aloud sentences from newspaper extracts by native speakers of Farsi	2000 utterances each read aloud by 300 native speakers
OGI (Farsi Sub-corpus) (Muthusamy et al., 1999)	spoken	Answers to a set of questions in phone conversations	Approximately 2.5 hours (ongoing project)
Persian Speech Corpus	spoken	Read out news extracts by a singular speaker	2.5 hours
PERSICA [2] (Eghbalzadeh et al., 2012)	spoken	News reports	1,454,745 words
Large FARSDAT	spoken	Reading aloud of newspaper extracts	73 hours
CALL friend (Farsi Database) [3]	spoken	Telephone conversations between family members	1,590 hours of recording

TFarsdat (Bijankhan, 2003)	spoken	Telephone conversations in the form of interviews	2 hours (ongoing project)
Farsi Linguistic Database (FLDB) (Assi, 1997)	written & spoken	selection of contemporary Modern Persian literature, formal and informal spoken varieties of the language, and a series of dictionary entries and word lists	3,000,000 words
Corpus of Conversational Persian Transcripts (Mohammadi, 2019)	Spoken	Informal telephone and face-to-face conversations between Iranian Persian speakers living in Tehran	Approximately 1200 minutes of recording

The Corpus of Conversational Persian (CCP) (Mohammadi, 2019) is a 126,298-word spoken corpus of naturally occurring informal conversations in Persian. According to Mohammadi (2019, p. 6), the CCP was compiled and annotated for research in the areas of "text analysis, discourse analysis, sociolinguistics, cultural studies, gender studies, and pragmatics". The corpus represents the Tehrani dialect, which is also referred to as the Standard Dialect of Iran (Miller, et al., 2014). Textbooks and the current language programs are all informed based on the Persian which is referred to as the Tehrani dialect. The Tehrani Dialect refers to the Persian which is predominantly spoken in Iran's capital city of Tehran. This has also informed the choice of using speakers of the Tehrani dialect as the point of comparison with the learners. That is to stress and clarify that the group of speakers which are referred to as L1 speakers throughout this study do not act as a benchmark which indicates the correct use of Persian, but as a reference point for comparison in order to provide insights into learners' choice of forms.

The CCP is a sub-corpus of the General Corpus of Persian (Mohammadi, 2018). The General Corpus of Persian consists of words and is compiled from written and spoken texts in Persian. Table 4.5 shows the composition of the General Corpus of Persian.

Table 4.5.

Composition of the General Corpus of Persian

(adapted from Mohammadi, 2018, p. 68)

Text type	Topics	Number of words
Phone calls and face-to-face conversations	various	126,298
interview	socio-political/sociocultural	124,250
political speech and sermons	national/international/religious	124,874
political debate and public forums	presidential/socio-political/socioeconomic/art and culture/ideological	125,370
daily news	national/international	167,042
movie and play script	miscellaneous	167,891
children's story, adult storytelling, stand-up comedy	miscellaneous	167,535
published abstracts	humanities/social sciences/ natural sciences	199,183
legal text	constitution/ civil and criminal	201,407

front page newspapers	miscellaneous	100,910
novel	miscellaneous	100,101
online posts	cooking blogs/personal diaries/ romantic posts	100,802
comments (reactions to videos)	social issues/political issues/ comedy/general	101,401
online Q&A	legal/marriage and family/medical/psychological religious	100,021

As Table 4.5 shows, from all the sub-corpora of the General Corpus of Persian, the CCP was chosen as the reference corpus, since it was the most comparable with the Learner of Persian Corpus.

The CCP consists of 43 separate conversations between 22 individuals. All the conversations were recorded in Tehran, and each conversation was transcribed in separate XML files. Table 4.6 shows the composition of each separate file, which includes the number of words, the number of participants, and the gender of the participants.

Table 4.6

Composition of the CCP (adapted from Mohammadi, 2019, p.5)

File name	Number of words	Number of participants	Gender of participants
CCP01	11,873	2	male / female
CCP02	3,552	2	female

CC P03	4,033	2	female
CC P04	146	2	male / female
CC P05	3,702	2	female
CC P06	5,674	2	female / male
CC P07	3,963	2	male
CC P08	3,166	2	male
CC P09	494	2	female / male
CC P10	2,440	2	male
CC P11	5,993	2	male
CC P12	688	2	female / male
CC P13	2,836	2	female
CC P14	3,711	2	male / female
CC P15	2,773	2	female
CC P16	2,700	2	male / female
CC P17	4,314	2	male

CC	P18	2,741	2	male
CC	P19	4,701	2	female
CC	P20	4,788	2	female
CC	P21	240	2	female
CC	P22	4,688	2	male / female
CC	P23	4,786	2	female
CC	P24	8,382	3	male
CC	P25	4,986	2	female / male
CC	P26	1,913	4	female(3) / male
CC	P27	6,736	3	male / female(2)
CC	P28	1,877	3	female(2) / male
CC	P29	2,621	3	male(2) / female
CC	P30	3,726	5	male(3) / female(2)
CC	P31	1,816	2	male
CC	P32	3,418	3	male

CC P33	4,599	2	male / female
CC P34	8,610	3	male
CC P35	3,163	2	male / female
CC P36	503	2	female
CC P37	1,481	3	female(2) / male
CC P38	494	2	female
CC P39	1,891	3	male / female(2)
CC P40	2,251	3	female(2) / male
CC P41	2,200	2	male
CC P42	840	2	female / male
CC P43	3,702	2	male

From this Corpus, the face-to-face conversations were chosen for the purpose of this study. That is, the files including telephone conversations were left out of this study, and the face-to-face conversations were only included. Therefore, files 24 to 43 of the Conversational Corpus of Persian were only considered for the purpose of this study. These files will from now on be referred to as the Reference Corpus (RC) throughout this thesis. The RC of this study consists of 65209 words.

The reason for deciding to only include the face-to-face conversations of the Conversational Persian Corpus was based on ensuring comparability between the two data sets (i.e., the learner corpus and the reference corpus). Since the learner corpus of this study was based on face-to-face conversations, the reference conference also consisted of face-to-face conversations to allow for comparability of the two data sets. Excluding the telephone conversations was especially of relevance to this study since previous studies have reported differences between the use of discourse markers in face-to-face and telephone conversations. For example, Crystal and Davy (1975) report a decrease in the use of discourse markers in telephone conversations compared to face-to-face conversations. On the other hand, Urbanova (1999), for example, found the opposite to be true, especially regarding the increase in the use of forms signifying vague meanings. Urbanova (1999, p. 10) attribute this to "the lack of personal contact in telephone conversations" which in turn results "in a relatively high degree of tentativeness", which is reflected in the higher frequency of vague forms in telephone conversations when compared to face-to-face conversations.

4.6. Interviews

In addition to the corpora used, I also conducted semi-structured informal interviews with the learners involved in providing data for the LoPSC. These interviews were conducted after the learners had finished recording their conversations for the compilation of LoPSC. Each interview lasted between 15 to 20 minutes. The answers to the interviews were anonymised and stored in a separate file. Please refer to Appendices III and IV for further detail on the storage of the interview answers.

The interviews served two purposes in this study. First, the questions of these interviews, which served as the metadata for this study, included general information regarding the age of the participants and further information, such as the learners' first

language and similar information which has been included in Table 4.2. regarding participant information for LoPSC. The remaining questions covered in the interviews were related to the learners' perceptions towards using Conversational Persian and their knowledge and use of features that contrast Conversational Persian from other registers of the language.

As mentioned, the interview data were collected from the same participants that had participated in the main phase of the data collection. The interviews took place after the data collection sessions ended and were recorded using the same recording device used for the data collection.

In addition to these questions, the students were also asked a set of questions designed at gathering further information that could potentially justify certain decisions that learners made regarding the use of Conversational Persian. These questions included the types of material the learners used for learning Persian, in general, and Conversational Persian, specifically. Two other questions were also included in the interviews, namely, what the students thought were the most important aspects of Conversational Persian and under what circumstances (if any) they would use Conversational Persian. These interview questions were included since in addition to serving as the means for the collection of metadata, these interviews were collected as a means of possible triangulation (Baker, 2016) for the factors influencing the linguistic choices made by learners when using Conversational Persian. I expand further on the answers provided by the learners in sections 5.5 and 6.5.

4.7. Corpus Tools

In Corpus Linguistics, statistical measures used for analysis are available through corpus analytical software. Corpus analytical software provide a platform for analysing

corpora by using various corpus tools. In turn, corpus tools analyse the data with built-in statistical measures. Similar to other aspects of corpus-related research, the choice of tool and statistical measure depends on the research objectives of the study. The main corpus tools used in this study were keyword lists, frequency lists, concordance lines, collocations and N-grams.

Keyword Lists were initially used to answer the first research question of this study (What are the significant differences between the forms used by learners and L1 speakers of Persian?). Keyword Lists allow for two or more corpora to be compared. In addition, to keyword lists, due to certain aspects of the Persian language, an additional check of the differences using frequency lists was required (See section 5.4. for further explanation of this point). The keyword lists and frequency lists are interconnected in that frequency lists provide the basis for keyword analysis. In brief, frequency lists provide the frequency of different forms that appear in a given corpus. Frequency lists provide both the absolute (or raw) frequency and the relative (or normalised) frequency of forms. The relative frequency mainly serves as the means for comparison across corpora. Since this study compares the frequency of two corpora, both the raw and relative frequencies have been included.

Keywords Lists Keywords lists are created using the frequency lists of two corpora. One corpus is used as the corpus under study and the second corpus is used as the reference corpus for means of comparison. By using the frequency lists of two corpora, the keywords' list tool creates what are known as keywords' lists. Keywords' lists can show positive keywords (that is, words that appear with a significantly higher frequency in the first corpus compared with the reference corpus) or negative keywords (that is, words that appear with a significantly lower frequency in the first corpus when

compared with the reference corpus.). For this thesis, the corpus under study was the LoPSC and the Conversational Persian Corpus acted as the Reference Corpus.

The following research questions of this study were answered using a mixture of corpus tools. Collocations, described as “lexical co-occurrence, i.e. the co-occurrence of words with other words” (Gries and Durrant, 2021, p. 142), were mainly used to determine and compare the pragmatic functions of the identified forms used for the purpose of data analysis, which was directed towards answering research question two, but was also used when determining the pragmatic functions of the discourse markers and vague category markers (See section 5.2.1.3 for an illustration of how collocations were used in this study).

As will be further expanded on in Chapter 5 sections 5.2. and 5.3, key n-grams were also a corpus tool used in this study, since there was an assumption that certain identified keywords were part of larger units of language. These larger units of language which in this study I refer to as n-grams are defined as “contiguous sequences of words of various length [that] are known by many names, including formulaic sequences, lexical bundles, or n-grams (where n = the number of words in the sequence, e.g., bigrams, trigrams, etc.).” Miller goes on to separate n-grams from other multi-word units by stating the following:

> “Conceptually, what separates these units from other MWUs is that, rather than being recognized and interpreted as semantically ‘complete’ units (as with, for example, phrasal verbs or idioms), formulaic sequences may appear semantically or structurally incomplete and are thus typically categorized functionally (e.g., framing: the existence of a; quantifying: in a number of) or structurally (e.g., PP-based: as a result of; NP-based: the nature of the)” (Miller, 2021, p. 81)

In terms of the statistical measures drawn for the calculation of key n-grams, it is seen as an “extension” of the keywords approach. (Rayson and Potts, 2021, p. 132).

The final corpus tool that will be discussed in this chapter are concordance lines. This tool provides information about the forms in the context in which they occur. This is an important aspect of using corpora, in that words are not considered in a vacuum but can be observed in real-life situations. In this sense, concordance lines provide the researcher with qualitative means of analysing the corpus data in addition to the quantitative analysis provided by other corpus tools.

In Chapter 5, I illustrate how each tool was used to analyse the data in this study. However, before moving on to the analysis of the data, a note is to be made regarding the corpus analytical software used for this study. The software used for the purpose of this study was #LancsBox 6.0 (Brezina et al., 2020). This software was chosen for two reasons. First, the software was highly compatible for analyzing the Persian script. That is, although UTF-8 was used for encoding the text, due to the nature of the Persian script, especially regarding the prominent use of half spaces, the generation of keyword lists and collocations with certain software, such as Sketch Engine (Kilgarriff et al., 2014), was not possible. The second reason for the choice of #LancsBox 6.0 is related to the more extensive list of measures available, especially for generating keyword lists and collocations. For example, there are 8 statistical measures for calculating collocations and 5 measures for the calculation of keywords. This was especially important for the purpose of this study since, first, both corpus tools were used in this study and therefore having more statistical measures for calculating each tool would allow for a more robust choice final decision to be made regarding the choice of statistical measures (for example, see section 5.2.2.3. for an

explanation of the choice of the collocation measure used to calculate the collocation score).

4.8. Chapter Summary

In this chapter, I, first, provided a description of the compilation of the LoPSC. This description included the factors considered for the design and the measures applied for compiling the corpus. Since a reference corpus was also used in this study this reference corpus was also described in detail. In addition, the ethical considerations applied to all the stages of data collection and data management were also discussed. The third part of this chapter centred around the transcription of the learner corpus. Finally, this chapter concluded with a description of the main corpus tools used for analysis, namely, frequency lists, keyword lists, and concordance lines. The analysis of the corpora using these tools will be discussed in the following chapter.

5. Analysis and Results Chapter

5.1. Introduction

This chapter explains the analysis and results used to answer the research questions of this study.

1. What are the significant differences between the forms used by learners and L1 speakers of Persian?
2. What are the differences in the pragmatic functions of the most significant forms used by learners and L1 speakers?
3. What are the differences between the use of discourse markers by learners and L1 speakers of Persian?
4. What are the differences between the use of Vague Category Markers by learners and L1 speakers of Persian?

Therefore, the first section of this chapter explores the first research question of this study, by conducting a Keywords Analysis using the LoPSC and RC. The remaining sections of this chapter further investigate the results of this analysis to provide answers to the three remaining research questions.

5.2 Keywords of the LoPSC

Table 5.1. shows the five top positive keywords of the LoPSC. (A complete list of the first 50 positive keywords for the LoPSC appears in Appendix V).

Table 5.1.

LoPSC top five keywords

Type	Frequency LOPSC	Relative Frequency LoPSC*	Dispersion LOPSC	Frequency RC	Relative Frequency RC	Dispersion RC	Keyness Score**
amâ (but)	293	97.66	0.37	3	0.44	3.25	74.13
ou (he/her/him/her)	192	64.00	0.62	0	0	0	68
baleh (yes)	485	161.66	0.5	69	10.14	1.01	28.8
barâyeh (for)	270	90.00	0.33	43	0.0632	1.53	28.75
kami (a little/ a bit)	47	15.66	0.6	1	0.1471	4.36	27.84
mikonam (I do)	272	90.06	0.31	73	10.73	0.79	18.99

*Note: The relative frequency was calculated per 10.000 words.

**Note: The keyness score was calculated using Simple Maths (Kilgarriff, 2014).

According to Table 5.1, the five top positive keywords for the LoPSC were *amâ* (but), *ou* (he/she/him/her), *baleh* (yes), *barâyeh* (for), *kami* (a little/a bit), *mikonam* (I do), with the respective keyness scores of 74.13, 53, 28.8, 28.75, 27.84, and 8.99.

According to Table 5.1, *ou* (he/she/him/her) occurs with a raw frequency of 192 times in the LoPSC whereas it does not appear in the RC. As Ghomeishi (2018) observes, in colloquial Persian, the standard form for the third person singular pronoun used to refer to humans (i.e., *ou*) is replaced by the demonstrative pronoun *oun* (that). This was also the case in the RC; that is, *oun* was always used as a replacement for *ou*. Extract 1 shows an example of the use of *oun* to refer to a person.

Extract 1

من فایلهای خام صدا رو میدم کیا، ***اون*** کاراشو میکنه.

Man filhâye khâm sedâ ro midam kiâ, ***oun*** kârâsho mikoneh.

I will give the raw files to Kia, ***he*** will do its work.

In Extract 1, *oun* is used as the pronoun to refer to Kia, who is the mutual friend the speakers were talking about in this conversation.

Therefore, *ou* as a positive keyword was not further analysed in this study since there was a clear preference for the learners of the LoPSC to use this formal written form whereas the speakers of the RC opted for the colloquial form.

In addition, *barâyeh* (for) was not considered for the analysis of this study since the lower frequency of *barâyeh* (for) in the RC was the result of its representation in different orthographic and morphological forms in the RC. That is, regarding its orthographical variations, *barâyeh* (for) was transcribed in its colloquial spoken form *barâ.* As for the morphological variations of *barâyeh* in the RC*,* the colloquial form *vâseh* (for) was used by the speakers of this corpus. *barâ* occurs 62 times (9.11), and *vâseh* occurs 82 times (1205) in the RC. With the addition of the 6.32 of the occurrences of *barâyeh* (for), the equivalent of the English *for* occurs 27.48 times in the RC. Therefore, although there is a difference between the frequency of use of *barâyeh* (for) between the LoPSC and the RC, there is no significant statistical difference to place *barâyeh* as one of the five top keywords in the LoPSC.

The two last forms of Table 5.1, *kami* (a little/ a bit) and *mikonam* (I do), occurred most frequently as part of two of the top 5 key N-grams of the LoPSC, namely, *yek kami* (a little/a bit) and *fekr mikonam* (I think)[4]. Therefore, these two forms will be further explained in section 5.2 on Key N-grams.

With the exclusion of *ou* and *barâyeh,* and the explanation of *kami* and *mikonam* left for further sections in this chapter, the occurrences of the remaining two keywords of Table 5.1, namely, *amâ* and *baleh* are further described in this section.

5.2.1. *amâ*

Based on Table 5.1, *amâ* (but) is the token with the highest positive keyness score of 74.13. This indicates that it appears significantly higher in the LoPSC compared to the RC.

amâ appears with a raw frequency of 293 (97.66) in the LoPSC. However, it occurs 3 (0.44) times in the RC, with 2 out of the total 3 hits appearing in only one of the texts; thus, justifying the high coefficient of variation of *amâ* in RC (i.e., 3.25 from a maximum of 4.36).

To examine the reason for this discrepancy in the use of *amâ* in the two corpora, the following is divided into two sections. First, the positioning and functions of *amâ* in the LoPSC and the RC are described. Next, since *amâ* appears with only 3 occurrences in the RC, the possible substitutions for this word used by the speakers of the RC are explored.

Before providing a description of the positions and functions of *amâ,* a note is to be made regarding the coding of functions for the chosen forms in this study. The coding took place during two sessions. That is, each occurrence of a particular form was coded in two separate coding sessions. Each session occurred with a weekly interval. This decision was made since as stated in Chapter 2 section 2.4.1., when providing categories for functions of discourse markers and vague category markers, an overlap of functions is at times inevitable. The same applied to the identification of functions for this study. That is, at times, certain forms would correspond to two or more functions. Therefore, choosing two coding sessions with a time interval in between coding would strengthen the reliability of the coding decision. In instances where an overlap did occur, coding would take place a third time.

The decision to correspond each occurrence with a singular pragmatic function was made based on an initial inter-rater reliability test using Cohen's Kappa Statistics. According to Cohen (1960), the agreement between two raters using Kappa Statistics falls between 0 to 1, with different intervals indicating different levels of agreement. That is, the value 0 would indicate no agreement between the raters. Whereas 0.01–0.20 indicates slight agreement, 0.21–0.40 as a fair level of agreement, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement between the raters. The first and second coding of the functions of this study received a score of 0.77. Therefore, based on Cohen's proposed measurements, there was a high level of agreement between the two codings. Nonetheless, in instances in which an agreement was not made, coding did take place a third time.

In any case, this study does acknowledge that an overlap of functions did occur with the forms; however, as explained, with the use of the Kappa Statistics, this overlap was not statistically significant to require double or multiple illustrations of functions for each form.

Position and functions of amâ in the LoPSC

Table 5.2 shows the positioning of *amâ* in the LoPSC.

Table 5.2

Positions of amâ in the LoPSC

Turn	Raw Frequency	Relative Frequency	Percentage
Turn initial	55	18.33	18.77
Turn mid	226	75.33	77.13
Turn final	12	4	4.06

Total	**293**	**97.66**	**100***

*The actual addition of the percentages is 99.96 and not 100. This is due to rounding the individual percentages for the sake of brevity in presentation.

According to Table 5.2, the majority of the cases of *amâ* occurred in the turn-mid position (i.e., 77.13%). This is followed by cases of *amâ* occurring in the turn-initial position with 18.77%. Finally, the least number of occurrences of *amâ* were in the turn-final position with 4.06%. However, the concordance lines showed that *amâ* occurred in the turn-final position only when the speaker was interrupted by the other speaker(s) in the conversation. Extract 2 shows *amâ* occurring in the turn-final position as a result of Speaker 5 interrupting Speaker 4.

Extract 2

hmm <Speaker4>غیرقانونی هست و اگه مشکل برای آنها میشه ***اما***

<<Speaker5> وسواس است

<Speaker 4> hmm gheyre ghanoono ast va ageh moshkeli baraye ona mish ***am***

<Speaker 5> vasvâs ast

<Speaker 4> hmm it is illegal and if a problem for them ***but***

<Speaker 5> It's obsession

In Extract 1, Speaker 4 is talking about a political scandal that happened in France. In previous turns of the conversation, he could not find the equivalent of *scandal* in Persian. While he continues to talk about the political scandal, in turn (2), Speaker 5 interrupts Speaker 4's turn with a word that he believes is the equivalent to *scandal* in Persian, namely, *vasvâs*. Then, the two speakers digress from the topic of the French Scandal to talk about the word *vasvâs.*

Therefore, the learners of the LoPSC only used *amâ* in the initial or mid-turn position. This selection of turn position for *amâ* is similar to its English equivalent, the connective *but* (Fraser, 1996).

5.2.1.1. Positions and Functions of amâ in the LoPSC

The functions for amâ were manually coded based on the functions of its English equivalent, *but*. The pragmatic functions for *but* are to signal protest, as a contrastive marker, as a mitigative marker, as a topic management device, to indicate surprise, to show a change of mind, and to draw the listener's attention (Fraser, 1996; Peterson, 1986; van Dijk, 1979). These functions were found in the LoPSC except for signalling protest. In addition to the functions mentioned, three additional functions for *amâ* were used in the LoPSC, namely, showing an attitude, showing surprise and presenting a stance. Each function will be further explained in the following with extracts from the LoPSC. Table 5.3 also shows the frequency of each function in the LoPSC[6].

Table 5.3

Frequency of functions of amâ in the LoPSC

Functions	Raw Frequency	Relative Frequency
Contrastive device	53	17.66
Mitigating device	27	9.00
Opener (introduction of new topic/ terminating a topic)	61	20.33
Change of focus (away or towards a topic and then back again)	82	27.33
Change of mind (seen in narratives)	14	0.466

Attention getter	13	0.433
Showing an attitude	15	5
Showing surprise	17	0.566
Presenting a stance	13	4.33

amâ as a contrastive marker

According to Table 5.3, *amâ* is used as a contrastive marker with a frequency of 17.66. Therefore, it is the third most frequent function for *amâ* in the LoPSC.

A connective can have the pragmatic function of a contrastive marker where it is used to contradict a prior statement (Fraser, 1996). This statement can be either spoken by the other speaker(s) in the conversation or a speaker may also use *amâ* to contradict their statement. Extract 3 shows an example of the use of *amâ* in the LoPSC.

Extract 3

<speaker 4>دامداری hmm ولی برای میلیونها نفر شاید

<speaker3> uhum

<speaker4> روش امرار معاششون است

<speaker3> ah

<speaker4> <laughs> ما این جوری نیستیم در انگلستان ***اما***

<speaker 4> damdari val baraye miliyoonha nafar shayad

<speaker 3> uhum

<speaker 4> ravesh emrar maasheshoon ast

<speaker 3> ***amâ*** ma in joori nistim dar inglestan <laughs>

<speaker 4> farming for millions of people may be

<speaker 3> uhum

<speaker 4> their way of life

<speaker 3> ***but*** we are not like that in England <laughs>

In Extract 3, Speaker 3 contradicts what was said by Speaker 2 in the prior turn by using *amâ*. Speaker 2 presents the need for farmers with cattle to continue their livelihood as a reason why vegetarianism would not be feasible. However, Speaker 3 contradicts this statement by stating that this does not hold true in the UK.

amâ as a mitigating device

amâ was also used as a mitigating device in the LoPSC with a frequency of 9. Connectives are used as mitigating devices to

Extract 4

<speaker8> چون فکر میکردم ساعت ۱۲ آن جلسه تمام بشود و همشون برون

<speaker9> آره

<speaker8> به نظر رسید که مثل مهمونی بود

<speaker9> آره آره نه میفهمم

<speaker8> بله

<speaker9> موندیم دیگه اما من همین فکر کردم که ما باید بریم دیگه تموم شد (name-of-person2) اتفاقاً من و آره

<speaker8> (Incomprehensible)

<speaker9> نه عیبی نداره. ***اما*** یک پیغام بهش بفرستید، خب؟

<speaker8> chon fekr mikardam saat 12 on jalaseh tamam beshavad va hameshoon beravan

<speaker9> areh

<speaker8> be nazar resid keh mesle mehmooni bood

<speaker9> areh areh na mifahmam

<speaker8> baleh

<speaker9> moondim digeh. Ama man hamin fekr kardam keh ma bayad berem digeh tamoom shod areh. etefaghan man o (name-of-person2)

<speaker8> (Incomprehensible)

<speaker9> ***amâ*** yek peygham behesh befrestid, khob?

<speaker8> because I thought that meeting would end at 12 and everyone would leave

<speaker9> yeah

<speaker8> it seemed like a party

<speaker9> yeah yeah I understand

<speaker8> yes

<speaker9> we just stayed. But I thought that we should leave as well, it's finished yeah. In fact, (name-of-person2) and I

<speaker8> (Incomprehensible)

<speaker9> ***but*** send (her) a message, OK?

Speaker 9 and the other speaker in the conversation, Speaker 8, had both attended a party organised by a mutual friend; however, Speaker 8 had left the party earlier without saying goodbye to the host. The host was surprised that Speaker 8 had left abruptly. So, in Extract 4, Speaker 9 suggests that Speaker 8 send a text to the host as an apology. However, Speaker 9 mitigates the possible intrusive force of her suggestion by using *amâ* as a mitigative device.

amâ as a topic management device: opener and change of focus

As shown in Table 5.3, *amâ* was frequently used in the LoPSC as a topic management device, namely, to introduce a new topic (hereafter referred to as opener) or to change the focus of the conversation either away or back to the topic being discussed by the speakers.

***amâ* as an opener**

amâ as an opener was the second highest function used in the LoPSC with a frequency of 20.33. Extract 5 shows an example of the use of *amâ* as an opener. In Extract 5, Speaker 13 uses *amâ* to change the topic of the conversation from explaining his bad experience with a shop assistant at a shop to the aesthetics of the same shop. Speaker 13 also shows this change of topic in the last turn by saying “And what’s up?”.

Extract 5

<speaker 13> خیلی عجیبه وقتی که کسی که یعنی او میخواد فرش فرش هایش را بفروشد اما نمیدونم اما (IC)

<speaker 12> آره <laughs>

<speaker 13> خوب نیست

<speaker 12> آره فکر میکنم علاقه به هیچی نداره یعنی این جوری هست که به نظر میرسه که واقعاً خوشحال نیست میدونی؟

<speaker 13> hmm سرش خیلی شلوغ بود hmm احتمالاً چون که آخر هفته

<speaker 12> آره ممکنه

<speaker 13> که مدیره اون جوری نباشد(name-of-person7) بله اما فکر میکنم که

<speaker 12> آره نه او خیلی مهربونه

<speaker 13> آره عجیب <laughs>

<speaker 12> آره <laughs>

<speaker 13> ***اما*** خیلی قشنگ آن فروشگاه

<speaker 12> آره

<speaker 13> من خیلی دوست دارم. و تو چه خبر؟

<speaker 13> (Incomprehensible) kheili ajibeh vaghti keh kasi keh yani ou mikhahad farsh farshhayash ra beferoshad ama nemidoonam ama

<speaker 12> areh <laughs>

<speaker 13> khob nist

<speaker 12> areh fekr mikonam alagheye be hichi nadareh ysni injoori hast keh be nazar mireseh keh vaghan khoshhal nist, midooni?

<speaker 13> saresh kheili sholoogh bood hmm ehtemalan chon keh akhar hafteh hmm

<speaker 12> areh momkeneh

<speaker 13> baleh ama fekr mikardam (name-of-person7) injoori nabashad

<speaker 12> areh na ou kheili mehrabooneh

<speaker 13> <laughs> areh ajibeh

<speaker 12> areh <laughs>

<speaker 13> ***amâ*** kheili ghashand an foroshgah

<speaker 12> areh

<speaker 13> man kheili doost daram. va to cheh khabar?

<speaker 13> (Incomprehensible) it's very strange that somebody who wants to sell their carpets carpets but I don't know but

<speaker 12> yeah <laughs>

<speaker 13> it's not good

<speaker 12> yeah I think he doesn't like anything. I mean it seems as though he really isn't happy , you know?

<speaker 13> he was really busy hmm it was probably because it was the weekend hmm

<speaker 12> yeah it might have been

<speaker 13> yeah but I think the manager (of the shop) (name-of-person7) is not like that

<speaker 12> yeah no he is very kind

<speaker 13> <laughs> yeah it is very strange

<speaker 12> yeah

<speaker 13> ***but*** (it is) really nice that shop

<speaker 12> yeah

<speaker 13> I like it. And what's up?

amâ for change of focus

The highest frequency function for *amâ* in the LoPSC (with a relative frequency of 27.33) was to change the focus of the conversation either away or back towards the topic being discussed after a digression from the topic. Extract 6 shows an example of this function wherein Speaker 16 moves back towards the topic of the heating of her house after a digression from the topic to find the equivalent of *heating* in Persian.

Extract 6

<<Speaker 16 توی خونه ی من نمیدونم heating چیه به فارسی ***اما*** خراب شد یعنی کاملا خراب شد

<Speaker 16> toye khouneh man nemidoonam heating chiyeh be farsi ***amâ*** kharab shod yani kamelan kharab shod.

<Speaker 16> in my house I don't know what heating is in Persian ***but*** it's broken (I mean) it's completely broken.

amâ to show a change of mind

amâ was used with a relative frequency of 0.466 to indicate a change of mind by the speaker about what they had said in previous statements. Extract 7 shows an example of this function.

Extract 7

اینجا دوستان صمیمی کمی دارد hmm ***اما*** نه چرا که دارد

inja doostane kami darad hmm ***amâ*** nah chera keh darad

here he has a few friends hmm ***but*** no why he does

***amâ* to draw the listener's attention**

According to Table 5.3., *amâ* was used with a relative frequency of ***0.***433 by speakers to draw the listener's attention. Extract 8 shows such an example.

Extract 8

<Speaker 13> فردا میخواستم با (name-of-person) صحبت کنم ***اما*** میدونی که چرا هر ساعت قبل از جلسه لغو میشه اگه کس دیگر hmm hmm نپیوندن؟

<Speaker 13> farda mikhastam ba (name-of-person) sohbat konam ***amâ*** midouni keh chera har saat ghabl az jalaseh laghv misheh ageh kase digar hmm hmm napayvandan?

<Speaker 13> tomorrow I wanted to speak with (name-of-person) ***but*** do you know why every hour before the meeting it disconnects if someone else hmm hmm does not connect?

In Extract 8, Speaker 13 is talking about a meeting he has on an online platform called Chatterbox. To draw his listener's attention to a question he has regarding connecting to the platform he uses the connective *amâ* before asking his question.

amâ to show an attitude

In addition to the functions identified in previous literature for contrastive connectives, *amâ* was also used by speakers of the LoPSC to show an attitude. This attitude was

either in the form of disapproval, disappointment or annoyance. According to Table 5.3., this function occurred with a frequency of 15

Extract 9 is an example of the occurrence of such a function in which the speaker is showing disapproval or annoyance at her mum for not considering talking further about how she is meant to fund her travels.

Extract 9

<<Speaker7 به مامانم درباره ی پول و این جور چیزا صحبت کردم ***اما*** فقط گفت که

ما درباره ی آن بعدن درباره ی آن

<Speaker 7> beh mamanam darbarehye pool o injoor chiza sohbat kardam

amâ faghat goft keh ma badan darbarey an sohbat mikonim

<Speaker 7> I talked with my mom about money and those sort of

things ***but*** she only said that we will talk about that later

Extract 10 shows another example of this function in the LoPSC. In Extract 10, Speaker 6 is expressing her annoyance towards a restaurant due to a negative experience she had with the quality of the food there.

Extract 10

<<Speaker6 میتونن به درستی اینو بپختن ***اما*** نمیدونم تصمیم میگیرن که این کار رو نمیکنم

<Speaker 6> mitonan be doresti ino bepokhtan ***amâ*** nemidoonam tasmim

migiran keh in kar ro nemikonam

<Speaker 6> They can cook this the right way ***but*** I don't know why they decide not

to do this

***amâ to* show surprise**

amâ was also used by the speakers of the LoPSC to express surprise. Extract 11 shows an example in which Speaker 13 expresses his surprise in finding out that unlike the British, Iranians do not frequently meet with other Iranians when living abroad.

Extract 11

Speaker13>> من فرض میکنم که ایرانیها دوست دارن با ایرانیهای دیگر ملاقات کنن

اما اون جوری نیست

<Speaker 13> man farz mikonam keh iraniha doost daran ba iranihaye

digar molaghat konan *amâ* ounjouri nist

<Speaker 13> I thought that Iranians like to meet other Iranians but it is not like that

The function of expressing surprise using contrastive connectives was previously observed in previous literature. However, the surprise mentioned was targeted towards what the other speaker had said and occurred in the initial turn position (van Dijk, 1979). This is in contrast to what was found in the LoPSC. That is, expressions of surprise occurred within mid-turn positions and the speaker directed this surprise towards their own previous perceptions, as was shown in Extract 11.

amâ to present a stance

Table 5.3. also shows the occurrence of *amâ* to present the speaker's epistemic stance. This function occurred with a frequency of 13 in the LoPSC.

Extract 12 is an example of such a function in which Speaker 15 signals his hesitance about his previous statement regarding whether or not Vitamin D can be found in almond milk by using *amâ.*

Extract 12

<Speaker 15> نمیدونم درباره ی شیر almond فرض میکنم چون که هم جایگزین شیر گاو است اما نمیدونم

<Speaker 15> nemidoonam darbareye shir almond farz mikonam chon ke ham jaygozin shir gav ast ama nemidoonam

<Speaker 15> I don't know about almond milk I imagine because it is a substitute for cow's milk as well but I don't know

As this section has shown, the speakers of the LoPSC used the contrastive connective *amâ* with a high frequency and for a variety of functions. The next section aims to show how this connective was used in the RC.

5.2.1.2. Position and functions of amâ in the RC

Considering that there were only three occurrences of *amâ* in the RC, no patterns could be ascertained regarding the position of *amâ* in the RC.

Therefore, the next step taken was to explore the possible substitutes for *amâ* in the RC. The following section describes how these possible substitutes were accounted for.

5.2.1.3. Substitutes for amâ in the RC

The first step was to look up synonyms for *amâ* in the keyword list to see if any synonyms for *amâ* were used in the RC to compensate for the lower frequency of *amâ*. One synonym for *amâ* in Persian is the contrastive connective *vali* (but). *vali* appeared 227 times (relative frequency of 33.38) and 113 times (relative frequency of 37.66) in the RC and LoPSC, respectively. Therefore, *vali* appears with a similar relative frequency in both corpora, and this indicates that the speakers of the RC are not using *vali* as a substitute for *amâ*.

Nonetheless, the functions of *vali* in the RC were explored to investigate the differences between the functions of *amâ* in the LoPSC and *vali* in the RC. This comparison was made in order to explore whether there was a stronger reliance on *amâ* in the LoPSC to perform a certain function(s) that was not accounted for or appeared with less frequency in the RC.

Table 5.4. shows the frequency of functions for the connective *vali* in the RC.

Table 5.4.

Frequency of functions for vali in the RC

Functions	Raw Frequency	Relative Frequency
Contrastive device	18	0.264
Mitigating device	19	0.279
Opener (introduction of new topic/ terminating a topic)	8	117
Change of focus (away or towards a topic and then back again)	120	1764
Change of mind (seen in narratives)	4	58
Attention getter	1	14
Showing /surprise	21	308
Signalling annoyance/ disappointment	11	161
Presenting a stance	10	147

Note: Three cases of the occurrences of *vali* were not considered since the speakers' turn was interrupted and consequently the function was not made evident.

As Table 5.4. shows, *vali* in the RC performs similar functions to *amâ in the LoPSC* albeit with relatively less frequency. This discrepancy in the relative frequency of the two forms is especially more conspicuous regarding the use of *amâ* as a contrastive and mitigating device. That is, whereas according to Table 5.4. the relative frequency of the occurrence of *vali* in the RC as a contrastive and mitigating device is 2.64 and 2.79, respectively, in the LoPSC, according to Table 5.5, *amâ* appears as a contrastive and mitigating device, respectively, 17.66 and 9 times.

The more frequent reliance of the speakers of the LoPSC on *amâ* to fulfil the function of a mitigating device is also reflected in the comparison between the collocations of *amâ* in the LoPSC and *vali* in the RC. As Table 5.5 which indicates the top five collocations for *amâ* in the LoPSC shows, *amâ* collocates frequently with the two affirmative markers *âreh* (yeah) and *baleh* (yes). As shown in Extract 13, the combination of *amâ* with these two words was frequently formed to show mitigation when contradicting the interlocutor's prior statement(s).

Extract 13

<Speaker 17> ***بله آره اما*** ساحلش خیلی زیباست

<Speaker 17> ***baleh âreh amâ*** sâhelash kheili zibâst

<Speaker 17> ***yes yeah but*** the beach is very beautiful

Extract 13 is Speaker 17's response to his interlocutor's previous statement about Dubia not being a good location for travel since there are only workers and tourists there. In response to this statement, Speaker 17 contradicts what his interlocutor has said. However, by initially using the two affirmative markers *baleh* and *âreh,* he also

mitigates the direct contradiction he is making. As Table 5.5 shows the combination of *amâ* with the two affirmative markers is a frequently reoccurring pattern in the LoPSC.

Table 5.5

Top 5 Collocations for amâ in the LoPSC

Position	Collocate	MI3 Score
L	keh (that)	57.12
L	kheili (very)	56.09
R	âreh (yeah)	45.24
R	baleh (yes)	43.07
L	fekr (think)	24.79

However, this does not hold true for *vali* in the RC. The speakers of the RC take a more direct and emphatic approach when contradicting their interlocutors' statements. This is attested by the use of the first-person singular pronoun *man (I)* as one of the collocates for *vali* as shown in Table 5.6.

Table 5.6

Top 5 Collocations for vali in the RC

Position	Collocate	MI3 Score
L	keh (that)	46
L	yeh (one)	37
L	khob (well)	33
L	**man (I)**	**32**

L	in (this) 31

Since Persian is a pro-drop language and speakers tend to use pronominal pronouns for the purpose of emphasis (Karimi, 2001), the strong collocation of *man* (I) with *vali* indicates that the speakers of the RC are more emphatic when presenting a contradiction. Extract 14 shows an example from the RC on the use of *man* with *vali* to show an emphatic contradiction.

Extract 14

<M1> [من نمیدونم چرا به هیچ رشته ای اون جور که باید علاقه ندارم] خنده

<M2> ***من ولی*** رشته خودمو دوست دارم

<M1> man nemidoonam cherâ be hich reshtei ounjor keh bâyad alâgheh nadâram (laughs)

<M2> ***man vali*** reshteye khodamo doost dâram

<M1> I don't know why I don't like any subject the way that I should (laughs)

<M2> ***But I*** like my subject

Speaker M1 had talked about how unhappy and disinterested he was with his subject of study in previous turns and, as shown in Extract 14, he generalises this disinterest to all subjects. However, M2 brings a contradiction indicating that he likes his subject of study. This unsolicited and contradictory opinion from M2 about his interest in his subject of study is further emphasised by using the pronoun *man*.

In conclusion, although *vali*, the near-synonym to *amâ* in the RC, displayed similar functions, these functions appeared with a lower relevant frequency. In addition, the comparison of the collocates of *amâ* in the LoPSC and *vali* in the RC also indicated that contradictions to interlocutor statements were made more directly and

emphatically compared to the LoPSC. Therefore, to explore what other possible forms the speakers of the RC may be using instead of *amâ,* the next step taken was to look at the negative keywords to see if any forms were used as counterparts to *amâ* in the RC. Therefore, a negative keyword list was generated. Table 5.7 shows the five top negative keywords of LoPSC (See Appendix VI for a list of the first 50 negative keywords of the LoPSC).

Table 5.7

Top 5 LoPSC Negative Keywords

Token	Frequency in LopSC	Frequency in LoPSC	Frequency in RC	Frequency in RC	Keyness Score
bebin (look)	0	0	148	21.76	0.05
âkheh (adversative discourse marker)	**0**	**0**	**107**	**15.73**	**0.06**
douneh (numeral classifier)	0	0	93	13.67	0.07
hey (again)	0	0	90	13.23	0.07
boro (go)	0	0	87	12.79	0.08

According to Table 5.7, *âkheh* is one of the negative keywords of the LoPSC with a keyness score of 0.06, a raw frequency of 107 in the RC and no occurrences in the LoPSC. According to Mohammadi (2018), *âkheh* is a discourse marker that is used to

express “denial, disapproval, objection, or refusal” with “extenuation” and with no equivalent in English (p. 157). This function of *âkheh* is shown in Extract 15 below.

Extract 15

<M1> اونجا چیزتره مثل این که قیمتاش بهتره تو همون بلواره بگیری بهتره

<M2> **آخه** بستگی داره اون ور اگه در حال رشد باشه نه

<M1> ounjâ chiztareh mesle in keh gheymathâsh behtareh to hamoun bolvâreh begiri behtareh

<M2> ***âkheh*** bastegi dâreh oun var ageh dar hâl roshd bâsheh, nah?

<M1> (it seems) like the prices are better there if you get it in that boulevard, it’s better

<M2> ***âkheh*** *(*it) depends if that area is in the process of development, no?

In Extract 15, M1 thinks that buying a flat in a certain boulevard would be better for M2. In his turn, M2 contradict M1 by using the discourse marker *âkheh* at the beginning of his turn followed by an explanation of why he thinks that buying a flat in another area would be better. To further mitigate the force of his contradiction, he uses a tag question at the end of his turn.

Therefore, one explanation for the use of lower frequency use of the contrastive conjunction *amâ* to mitigate the force of contradicting an interlocutor by the speakers of the RC is that they use the discourse marker *âkheh* instead. Whereas the speakers of the LoPSC rely heavily on the use of *amâ* to fulfil this function.

5.2.1.4. Summary for amâ

This section has shown, in the LoPSC, there is a preference to use forms associated with the standard written Persian instead of colloquial Persian. This is attested by the use of *ou* instead of *oun*, *barâye* instead of *vâseh*, *amâ* instead of *vali* and *âkheh.* In the specific case of the contrastive conjunction *amâ*, the speakers of the LoPSC have shown that they rely heavily on this form to perform several functions, including a

mitigating contrastive device. In contrast, the speakers of the RC, used contrastive conjunctions with a more direct and emphatic approach and preferred the use of the discourse marker *âkheh* to mitigate contradictions.

5.2.2. baleh (yes)

According to Appendix V, which shows the positive keywords in the Learner of Persian Spoken Corpus (LoPSC), the affirmative maker *baleh* (yes) has a keyness score of 28.8. *baleh* appears significantly higher in the LoPSC when compared to the Reference Corpus (RC). Therefore, this section aims to provide a description of the use of *baleh* in the LoPSC and the RC in order to make a comparison. However, before making this comparison, a short description of *baleh* in Persian is presented to describe which instances of *baleh* were considered for the purpose of analysis.

baleh in Persian: The Formal Affirmative Marker

Persian has two affirmative markers: baleh and *âreh.* Generally speaking, and for translation purposes, *baleh* is considered equivalent to *yes* in English, and *âreh* as an equivalent to the English *yeah*. Similar to *yes* and *yeah* in English, *baleh* is the preferred choice between the two affirmative markers in Persian formal contexts, and *âreh* is the preferred form used in informal contexts.

However, the restrictions on the use of *baleh* and *âreh* are not limited to the register in which they occur. One restriction for the choice between these two words is that while *baleh* can be used in response to vocatives, *âreh* cannot be used in such instances. However, since *baleh* was not used in either corpus as a response to a vocative, this use of baleh was not considered in the analysis of the data.

Another restriction on the use of baleh and *âreh* is that when answering phone calls, baleh can be used and not *âreh.* Extract 16 from the RC shows an example of *baleh*

being used when answering a sudden phone call while the two speakers were having a conversation.

Extract 16

بله ؟ سلام چه طوری ؟ چه طوری آناهیتا ؟

baleh? salâm che tori? che tori ânâhita?

Yes? Hi, how are you? How are you, Anahita?

(file.CCP24)

There were 5 instances, in the RC, in which *baleh* was used to respond to phone calls. However, since telephone conversations are not the focus of this study, these 5 instances were not taken into account when analysing the data. Therefore, the raw frequency of *baleh* in the RC was calculated as 64 instead of the initial 69 hits, as is shown in Appendix V.

Having provided a short description of the use of *baleh* in Persian and having determined which of its uses were (and were not) considered in the analysis of this study, the following sections set out to describe how the use of *baleh* differs in the LoPSC and RC. To do so, in addition to comparing the frequency and use of *baleh* in both corpora, the frequency and use of *âreh,* the informal counterpart of *baleh*, is also compared. *âreh* was used in this comparison to determine whether it was used (or could have been used) interchangeably with baleh, and if not used interchangeably, to help explain the preference of one form (i.e., *baleh* or *âreh)* over the other by the speakers in both corpora.

5.2.2.1. baleh in the LoPSC and the RC: frequency and distribution

Table 5.8 shows the frequency and dispersion measures for the two affirmative markers *baleh* and *âreh* in the LoPSC and the RC.

Table 5.8

Frequency and dispersion measure of baleh and âreh in the LoPSC

Affirmative marker	Raw Frequency	Relative Frequency per million words	Dispersion measure (Jullian's_d)
baleh	486	121.25	0.82
âreh	480	120	0.84

As shown in Table 5.8, in the LoPSC *baleh* occurs with a raw frequency of 486 (relative frequency of 12,125 per million words) and a Jullian's_d dispersion measure of 0.82. The informal counterpart of *baleh* (i.e., *âreh*) occurs with a similar raw frequency of 480 (relative frequency of 120) and a similar dispersion measure of 0.84. Therefore, both words appear with almost the same frequency and distribution within the LoPSC. However, in the RC, there are differences in terms of both frequency and dispersion of *baleh* and *âreh,* as shown in Table 5.9.

Table 5.9.

Frequency and dispersion measure of baleh and âreh in the RC

Affirmative marker	Raw Frequency	Relative Frequency per million words	Dispersion measure (Jullian's_d)
baleh	64	10.14	0.76
âreh	702	103.23	0.91

In contrast to the LoPSC, in the RC, *baleh* occurs with a raw frequency of 64 (relative frequency of 10.14) and a dispersion measure of 0.76, whereas *âreh* occurs 702 times (relative frequency of 103.23) with a distribution measure of 0.91. The frequency and

dispersion of both words show a strong preference for *âreh* by the speakers in the RC.

The high frequency of *âreh* in the RC asserts the inclination of Persian speakers to use *âreh* instead of *baleh* in informal colloquial speech. However, this still does not explain if *baleh* has similar functions in both corpora. That is, more specifically, it does not answer the two following questions:

1. Is the almost equal frequency and distribution of baleh and *âreh* in the LoPSC also an indicator that the two forms are used interchangeably in the LoPSC? And, if not, what are the differences between these two forms in terms of function, in the LoPSC?
2. If *âreh* is the preferred use in the RC, then how can the occurrences of *baleh* be explained? That is, is *baleh* used interchangeably with *âreh* in the RC? And if not, what are the differences between these two forms in terms of function, in the RC?

By addressing these two questions, the main question which is "what are the differences between the use of baleh in terms of function in the LoPSC and the RC?" can be addressed.

Therefore, the next section aims to provide a description of the functions of *baleh* in both corpora by first providing answers to the two questions mentioned above.

5.2.2.2. baleh in the LoPSC and the RC: turn position, collocations, and functions

To determine whether *baleh* and *âreh* were used interchangeably in both corpora, three steps were taken using corpus tools. First, the frequency of the turn positions (i.e., initial, mid, or final-turn position) for both forms was assessed. Next, the

collocations for both forms were examined in the corpora. Finally, using concordance lines, the functions of the forms were examined in their context of occurrence.

Turn position of baleh and *âreh*

Forms that are commonly referred to as affirmative markers can also demonstrate a number of other functions, such as forming tag questions or yielding the floor (Wouk, 2001). These functions tend to correspond to the position of the form within a turn. Therefore, examining the turn positions of *baleh* and *âreh* were considered of relevance to this study since, similar to affirmative markers in other languages, the two forms can also have different functions which correspond to their turn position in the interaction. Hence, similarity in turn positions could also indicate similarities in functions and overall usage of the form.

Table 5.10 shows the frequency of turn positions for both *baleh* and *âreh* in the LoPSC.

Table 5.10

Frequency of turn positions for baleh and âreh in the LoPSC

Turn Position	Turn-initial frequency	Mid-turn frequency	Turn-final (raw frequency/ frequency pm)
baleh	450 (11,250)	14 (350)	22 (550)
âreh	448 (11,150)	13 (315)	19 (495)

As shown in Table 5.10, most hits for both *baleh* and *âreh* occur in the turn-initial position. That is, from the 486 occurrences of baleh and the 480 cases of *âreh,* respectively, 450 and 448 hits are in the turn-initial position, in the LoPSC. All these instances occurred as an affirmative marker to a question or statement in the prior turn. Extract 17 shows an example of such an occurrence for the word baleh in the LoPSC.

Extract 17

<speaker 13> سالها هست که او را ندیدید؟

<speaker 12> **بله** فکر میکنم مثلاً چهار چهار سال پیش او را دیدم

<speaker 13> sâlhâst ke ou râ nadidid?

<speaker 12> **baleh** fekr mikonam masalan châhâr châhâr sâl pish ou râ didam.

<speaker 13> (You) haven't seen him for many years?

<speaker 12> **Yes**, I think for example four four years ago I saw him.

(LoPSC: Session 6)

In Extract 17, speaker 13 asks speaker 12 how long it has been since she last saw her grandfather (who, in previous turns, speaker 12 had mentioned she was not in contact with regularly). In this extract, *baleh* is used as an affirmative marker.

According to Table 5.10, in addition, to the preference for both *baleh* and *âreh* to occur in the turn-initial position and as an affirmative marker, in the LoPSC, both forms also appear with almost the same frequency for the mid-turn and turn-final position. That is, *baleh* and *âreh* occur 14 and 13 times, respectively, in the mid-turn position and 22 and 19 times, respectively, in the turn-final position.

In terms of the functions for the forms *baleh* and *âreh* occurring in the mid or final-turn position, in addition, to the affirmative marker function, both forms were also used as a continuer marker, topic changing and floor-yielding device. The following extracts from the LoPSC show examples of all of these functions as used by the speakers in the LoPSC

Extract 18 (*baleh* in mid-turn position as an affirmative marker)

<speaker10 قور .. فکر میکنم شکایت میکنن بهتر >

<speaker11 نه چی گفت؟ >

<speaker10 من داشتم میگم که قور قور >

<speaker11 قور قور **بله** و قور قور میکردن که زندگی خیلی سخته >

<speaker 10> ghor.. fek mikonam shekâyat mikonand behtareh

<speaker 11> na chi goft?

<speaker 10> man dâshtam migam keh ghor ghor

<speaker 11> ghor ghor **baleh** va ghor ghor mikardan keh zendegi kheili sakhteh

<speaker 10> na… I think (they) complain is better

<speaker 11> no ,what did (you) say?

<speaker 10> I was saying that nagging

<speaker 11> nagging **yes** and (they) were nagging that life is very difficult

In Extract 18 ,speaker 10 suggests using *ghor ghor kardan* (to nag) to speaker 11 .In the last line of this extract ,speaker 11 agrees with the use of *ghor ghor kardan* by using it and by adding baleh to her turn.

Similar to Extract 18 in Extract 19 ,speaker 5 also suggests a word speaker 4 could not remember and speaker 4 agrees with the use of this word by adding both a *baleh* and *âreh* to the end of his turn.

Extract 19 (*baleh* and *âreh* in the turn-final position as an affirmative)

<speaker5 احساسشون >

<speaker4 احساسشون **آره بله** >

<speaker 5> ehs*â*seshoun

<speaker 4> ehs*â*seshoun ***âreh baleh***

<speaker 5> their feelings

<speaker 4> their feelings **yeah yes**

Therefore ,as shown in Extracts 18 and 19 ,in the LoPSC ,both *baleh* and *âreh* were used as an affirmative marker in both the mid and final positions.

In Extract 20 ,*baleh* occurs in mid-position and is used as a continuer marker.

Extract 20 (*baleh* in mid-turn position as a continuer marker)

<speaker 10>

چی میگن نمیدونم سنتیین یا محافطه کار نه؟ **بله** احتمالاً خیلی دوست دارم که وقتی بچها داشته باشم مادرم یک مادربزرگ hmm باشد woked

<speaker 10> hmm chi migan nemidounam sonatiyan ya mohafezehkar nah? **baleh** ehtemalan kheili doust daram vaghti keh bacheha dashteh basham madaram yek madarbozorge woked bashad

<speaker 10> hmm what do they say I don't know their traditional or conservative, no? **Yes**, probably when I have kids I'd like my mother to be a woked grandmother.

In Extract 20, speaker 10 uses *baleh* to continue the conversation after he tries to think of an opposite for the word *woked* in Persian. However, in some cases, *baleh* and *âreh* were used as both affirmative and continuer markers by the speakers of the LoPSC, as in Extracts 21 and 22.

Extract 21 (*baleh* in mid-turn position as an affirmative and continuer marker)

مثل پیاز که میخریم آن وزن hmm و مثل گندم و از **بله** طبیعی نیست ولی باید> توسعه یافتن و طبیعی نیست <speaker 5>

<speaker 5> mesle piyâz keh mikharim ân vazne tabi nist hmm va mesle gandom va az **baleh** tabi nist

<speaker 5> like (the) onion that we buy that isn't the normal weight and like wheat and **yes** (it) isn't natural

In Extract 21, speaker 4 uses *baleh* as an affirmative marker to his interlocutor's claim in previous turns that produce that are labelled as natural are artificial in some way. However, baleh is also used in Extract 21 as a continuer marker. The same can be said about Extract 22 with the difference that, in addition to using *baleh* as a continuer marker, in this extract, speaker 14 uses *baleh* as an affirmation of what she had already mentioned in previous turns.

Extract 22 (*baleh* in mid-turn position as an affirmative and continuer marker)

<speaker 14 پس فکر میکنم **بله** باید پیوسته به اصولمون یا آن چیز مهم باشه >

<speaker 14> pas fekr mikonam **baleh** bâyad beh osolemoun ya ân chiz mohem bâsheh

<speaker 14> so I think **yes** (it) must (be related) to our principles or that important thing

Extract 23 below shows how *baleh* was used as a topic-changing device in the LoPSC. In this extract, speaker 16 begins the turn with *âreh* as an affirmative marker to agree with his interlocutor that he must make a certain phone call. Then, speaker 16 uses *baleh* to change the topic and ask the other speaker in the conversation what else was going on.

Extract 23 (*baleh* in mid-turn position as topic-changing device)

<speaker16 ؟چه خبر دیگه **بله** همین کار رو باید آره >

<speaker 16> *âreh* bâyad hamin kâr ro **baleh** digeh che khabar

<speaker 16> yeah) ,I) must do this thing **yes** what other news do you have?

In Extract 23 ,speaker 16 begins the turn with *âreh* as an affirmative marker to agree with his interlocutor that he must make a certain phone call .Then ,speaker 16 uses *baleh* to change the topic and ask the other speaker in the conversation what else was going on .The same function for baleh also happens in Extract 24.

Extract 24 (*baleh* in mid-turn position as topic changing device)

speaker4 به یک بیماری سوپر نیاز داریم >>

<speaker5> <laughs دیدی **بله** چون بیاید شاید شاید <hmm که فکر میکنم خیلی خوب hmm گوشت …

<speaker 4> beh yek bimâriye super niyâz darim

<speaker 5> <laughs> shâyad shâyad biyâyad chon **baleh** didi hmm keh fekr mikonam kheili khob hmm keh gosht…

<speaker 4> (we) need a super disease

<speaker 5> <laughs> maybe maybe (it) will come because **yes** did you see hmm that I think it's really good hmm (a) meat…

In Extract 24, speaker 4 makes a joke about probably needing a "super disease" that kills many people to solve the issue of climate change. In his turn, speaker 5 uses baleh to change the topic of the "super disease" and continues to talk about a different topic (vegan meat).

Finally, Extracts 25 and 26 show how *baleh* and *âreh* were used as floor-yielding devices when they occurred in the turn-final position.

Extract 25 (*areh* in the turn-final position as a floor-yielding device)

<speaker3 **آره** و داری آبی چشم چشم یه نمیدونستیم جالب چقدر بابا نه گفتن <

<speaker 3> goftan na bâbâ cheghadr jâleb nemidoonestim yeh cheshm cheshm âbi dari va ***âreh***

<speaker 3> they said no way (it) is so interesting we didn't know you had a blue-eyed eyed (one) and **yeah**

In Extract 25 ,speaker 3 is recounting the reaction her distant Iranian relatives had when they saw her for the first time .She mentions that they were surprised that her mother (who was half-Iranian) had a daughter with blue eyes and then gives the floor to the speakers by using *âreh.*

In Extract 26 ,baleh can be regarded as being used as an affirmative marker ,continuer marker and/or a floor-yielding device.

Extract 26 (*baleh* in the turn-final position as an affirmative marker, continuer marker and floor-yielding device)

<speaker9 **بله** اینجوریه میکنم فکر نمیدونم <

<speaker9> nemidoonam fekr mikonam injooriyeh **baleh**

<speaker 9> I don't know I think it's like that **yes**

In Extract 26, speaker 9 is responding to her interlocutor's question on whether ,in Saudi Arabia ,citizens of Britain and the USA are treated differently to citizens from India and Pakistan .Speaker 9 responds to this question by stating that she does not by using *nemidoonam* (I don't know) then she states that this might be the case by using *fek mikonam* (I think) to still show her uncertainty ,since she has never been to Saudi Arabia .Therefore ,although *baleh* could be still seen as an affirmative marker , the hesitancy of speaker 9 at the beginning of the turn shows that *baleh* may have

been used as a continuer marker signalling to her interlocutor to continue with the conversation or as a floor-yielding device indicating that she has no certain answer to the question and is therefore giving the floor to the other speaker .This extract shows that overlaps of functions for *baleh* and *âreh* occurred in the LoPSC.

In summary, in the LoPSC, in addition to the similar frequency and dispersion of *baleh* and *âreh* , the two forms also show a similar preference for the turn position they take, within an interaction, and show similar functions corresponding to the turn position they have.

Having explained the turn-position and functions of *baleh* and *âreh* in the LoPSC, the frequency of turn positions for the two forms in the RC is now considered. Table 5.11 illustrates this frequency of occurrence.

Table 5.11

Frequency of turn positions for baleh and âreh in the RC

Turn Position	**Turn-initial frequency**	**Mid-turn frequency**	**Turn-final frequency**
baleh	64 (941)	0	0
âreh	668 (9676)	9 (132)	25 (367)

As shown in Table 5.11, *baleh* and *âreh* show different preferences for the turn they take in the RC. *baleh* only occurred in the turn-initial position whereas, similar to occurrences of *baleh* and *âreh* in the LoPSC as shown in Table 5.11. That is, in the RC, *âreh* also occurs in different turn positions albeit the majority of the cases for *âreh* occur in the turn-initial position. Therefore, unlike *baleh* and *âreh* in the LoPSC, the two forms do not show similar frequencies nor distribution of turn positions, in the RC.

In terms of functions, *âreh* is used as an affirmative marker in all its occurrences in the turn-initial position and with 5 of its occurrences in the turn-final position, such as in Extract 27.

Extract 27

<M1> > سخته/M1>

<M2> > سخته آره/M2>

<M1> sakhteh </M1>

<M2> sakhteh ***âreh*** </M2>

<M1> (it) is difficult </M1>

<M2> (it) is difficult ***yeah*** </M2>

Therefore, in this regard, *âreh* functions similarly in the RC and the LoPSC. However, the other 20 remaining instances of *âreh* in the RC were used to form questions and this was not observed in the LoPSC. Extract 28 shows one of the examples of *âreh* in the turn-final question being used as a tag to form questions.

Extract 28

> M2/> کوبیده ام داشت آره ؟ <M2>

<M2>koubideh ham dâsht ***âreh***/> ?M2>

<M2>(It) had skewed minced lamb as well ,**yeah**/> ?M2>

As for the occurrences of *âreh* in the mid-turn position, all instances of *âreh* in this position were used when mentioning direct quotes from someone outside of the conversation. For example, in Extract 29, the speaker (F1) is talking about two friends and the text messages these two friends (Mazdak and Samira) had sent each other.

Extract 29

/>F1> بعد این ، مزدک گفته که ، مزدک گفته موهات چه قدر بلنده ، سمیرا زده که آره خیلی<F1>

/>F1> bad az in ,Mazdak gofteh keh ,Mazdak gofteh mouhât cheghadr bolandeh ,Samira zadeh keh *âreh* kheili<F1>

/>F1> after this ,Mazdak said ,Mazdak said your hair is so long ,Samira texted yeah very <F1>

Therefore, although *âreh* in the RC occurs with similar turn positions to the form in the LoPSC, there are additional functions that occur in the RC that did not appear in the LoPSC.

In the case of *baleh*, as shown in Table 5.11, it does not share the same frequency of turn positions with *âreh* in the RC and its counterparts in the LoPSC. That is, baleh only occurs in the turn-initial position.

To set a background for the presentation of the functions of *baleh* in the RC and to strengthen the claim showing that *baleh* and *âreh* are used interchangeably in the LoPSC whereas the two forms cannot substitute each other in the RC, the collocation tool of #LancsBox 6.0 (Brezina et al., 2020) was used to show if the words had similar collocations.

5.2.2.3. Collocations for baleh and âreh in the LoPSC and the RC

In this section, Tables 5.12, 5.13, 5.14 and 5.15 show collocations for *baleh* and *âreh* in both corpora. The collocations were chosen based on the words that co-occurred with either of the two forms within a span of 5 words to the left and right of the node word (i.e., either *baleh* or *âreh*). The collocations with a collocation score higher than 15 were considered as the top collocates of *baleh* and *âreh.*

This collocation score, which indicates the strength of collocation between the forms, was calculated using the cubed version of the Mutual Information (MI) statistic, which is referred to as MI3. Both MI and its other version MI3 determine the collocational

strength by comparing the raw frequency of each collocational pairing against what would likely happen based on the relative frequency of each word and the size of the corpus. The difference between the raw and expected frequency of co-occurrence is then converted into the MI score, with higher MI scores indicating stronger collocations (Gablasova et al., 2017). However, the difference between MI and MI3 is in that MI tends to include low-frequency words when calculating collocation strength whereas MI3 focuses on higher-frequency collocations (Evert, 2008; Brezina, 2018). Therefore, MI3 has the advantage of showing collocations that are "more established" in the discourse (Brezina et al., 2015, p. 160) and was therefore chosen as the statistic for calculating collocation strength in the analysis of this study.

Table 5.12 shows the collocations for *baleh* in LoPSC including the MI3 score and the frequency of the collocations in the corpus.

Table 5.12

Collocations for baleh in the LoPSC

Position	Collocate	MI3 Score	Frequency of collocation
-	baleh (yes)	18.9	74
R	âreh (yeah)	16.9	51
R	amâ (but)	15.6	23
R	kheili (very)	15.3	24

**Notes*. In #Lancsbox, the position of the collocate is indicated by one of these signs or letters: the

letter “R” shows that the collocate tends to occur to the right of the node word (i.e., *baleh*, in the case of Table 5.12); the letter “L” shows that the collocate appears more to the left of the node word; the letter “M” (which stands for the word *Middle*, here) indicates that the collocate occurs equally within the right or left of the node; finally, the dash sign(-) is used when the collocate and the node are the same word, as is the case in Table 5.12 where *baleh* as the node word also co-occurs with itself.

As shown in Table 5.12, the top collocates for *baleh* in LoPSC are: *baleh*, *âreh, amâ (but),* and *kheili (very/ a lot).* Table 5.13 also shows similar collocates (excluding *amâ (but))* for *âreh* in the LoPSC.

Table 5.13

Collocations for âreh in the LoPSC

Position	Collocate	MI3 Score	Frequency of collocation
-	âreh (yeah)	17.34	98
L	baleh (yes)	15.90	51
M	kheili (very)	15.76	46

Therefore, based on Tables 5.12 and 5.13, in addition to appearing with similar frequencies and similar patterns in terms of turn positions and functions, *baleh* and *âreh* also have similar collocates. Most importantly *baleh* and *âreh* are collocates for each other; therefore, strengthening the claim for the interchangeability of the two forms in the LoPSC.

However, as shown in Tables 5.14 and 5.15, in the RC, *baleh* and *âreh* do not co-occur together.

Table 5.14

Collocations for baleh in the RC

Position	Collocate	MI3 Score	Frequency of collocation
-	baleh (yes)	18.6196	30
M	Khoubeh (it is good)	15.59242	22

Table 5.15

Collocations for âreh in the RC

Position	Collocate	MI3 Score	Frequency of collocation
-	âreh (yeah)	18.3	132
R	Digeh (other/discourse marker with no exact equivalent in English)	17.2	112
R	kheili (very)	16	66

Table 5.15 shows that in addition to co-occurring with itself, *âreh* also co-occurs with the word *kheili* (very). This strengthens the claim that *âreh* in the RC shows similar functions to *baleh* and *âreh* in the LoPSC.

As Table 5.15 shows, similar to *âreh* in the RC, *baleh* also only co-occurs with itself; therefore, showing that the two forms cannot be used together or as a substitute for one another as was the case in the LoPSC.

The other collocate for *baleh* in Table 5.14 is the phrase *khoubeh* (it is good). An example of the occurrence of this phrase with *baleh* is shown in Extract 30 which has been translated into *OK*.

Extract 30

</M1>>؟حالت خوبه/M1>

</F1> >**بله**/F1>

</M1>>؟پرند هم دفاع کنه میخواد بره/M1>

</F1> >آره/F1>

</M1>halet khoubeh? </M1>

</F1> **baleh** </F1>

</M1>Parand ham defa koneh mikhâd bereh ?</M1>

</F1> **âreh** </F1>

</M1>Are you OK? </M1>

</F1> **yes** </F1>

</M1>Parand will leave as well when she defends (her thesis) </M1>

</F1> **yeah**</F1>

Extract 30 shows both *baleh* and *âreh* occurring in the interaction between the two speakers, M1 and F1. However, whereas *âreh* occurs as an affirmative marker to M1's question in the third line of the interaction, the choice of *baleh* to M1's first question corresponds to the nature of M1's question. That is, M1's second question is a "genuine" question whereas his first question (Are you OK?) is not a genuine concern for F1's health but he uses this question to imply that F1 is not OK based on what she had said in previous turns. This example not only illustrates that the speakers of the RC do not use *baleh* and *âreh* interchangeably, but the speakers also use the two

forms with the intention to perform different functions, and, in the case of *baleh,* to show an attitude or create an effect instead of acting as an affirmative marker. Similar examples of *baleh* being used for conveying an attitude or creating an effect are shown in Extracts 31 to 33.

Extract 31

<M1> حالا همونی که شما میفرمایید /M1>

<M2> واژگان درست استفاده کن /M2>

<M1> **بله** /M1 >

<M2> ناراحت شدی ؟ /M2>

<M1> hâla hamouni keh shomâ mifarmâyid/> M1>

<M2 <vâzhegân dorest estefâdeh kon/> M2>

<M1> **baleh**/> M1>

<M2> narâhat shodi/> M2>

M1> OK ,that (word) that you commanded/> M1>

<M2> Use words correctly/> M2>

<M1> **Yes**/> M1>

<M2> Are you upset/> ?M2>

In Extract 31 ,M1 is annoyed that M2 has corrected his choice of words in previous turns and shows this in the first line of this interaction by using the word “command .” M2 teases him by saying that he should use words correctly in the second line of the conversation then M1 replies by using *baleh* .As in Extract 30 ,*baleh* in Extract 31 is not used solely as an affirmative marker otherwise *âreh* would have been chosen .M1 uses *baleh* to convey his annoyance to M2 .This is picked up by M2 when he asks if M1 is upset by his correction.

A similar conveyance of annoyance (or contempt) is created by using baleh in Extract 32.

Extract 32

<F2> > **بله** ، اون رفته از ما شکایت کرده که اینا تو زندگی من دخالت میکنن </F2>

<F2> **baleh**, oun rafteh az m*â* shekayat kardeh keh in*â* to zendegi man dekhalat mikonan </F2>

<F2> **Yes**, he has gone and formed a complaint against us (saying) that they interfere in my life </F2>

However, *baleh* was mainly used in the RC to create humour or for the purpose of teasing the other speakers in the conversation, as shown in Extract 18.

Extract 33

<F1> > به من گفتی چشات بد رنگه؟ </F1>

<M1> > **بله** </M1>

<F1> beh man gofti cheshm*â*t bad rangeh? </F1>

<M1> **baleh** </M1>

<F1> (Did) you tell me that my eye colour was ugly? </F1>

<M1> **Yes** </M1>

In Extract 32, M1 uses *baleh* instead of *âreh* to tease F1 and to create humour. Similar to Extract 30, *baleh* is not a "genuine" response to a question but is used to create an effect which in this case was to tease F1.

5.2.2.4. Summary

In summary, this section has shown that the speakers of the LoPSC used the two affirmative markers, *baleh* and *âreh,* with similar patterns, regarding frequency, turn positions, collocations and functions. However, as expected considering the differences in register, the two forms, *baleh* and *âreh,* did not show these similarities in the RC. And, whereas *âreh* showed similar turn positions, collocations and functions

to the use of *baleh* and *âreh* in the LoPSC, *baleh* was used in the RC to convey different attitudes and to create certain effects. Therefore, unlike the speakers of the LoPSC, the speakers of the RC made a distinction between these two forms and hence restricted the use of the formal *baleh* only to situations in which they intended it to be used for certain purposes.

5.3. Key N-grams

This section looks at the differences between the multi-word expressions, more specifically, the two-word expressions (hereafter, referred to as 2-grams), used in the LoPSC and RC. To this end, the keyness score of the 2-grams in both corpora was calculated using #LancsBox 6 (Brezina et al., 2020). Table 5.16 and Table 5.17 respectively show the list of the top 3 positive and negative 2-grams for the LoPSC compared to the RC.

Table 5.16

Top 3 Positive 2-grams for the LoPSC

Type	Raw Frequency LoPSC	Relative Frequency in LoPSC	Dispersion LoPSC	Frequency RC	Relative Frequency in RC	Dispersion RC	Keyness Score
Fekr mikonam (I think)	221	73.66	0.29	11	1.61	0	45.6
Mikonam keh (do that)	35	14.33	0.46	3	0.14	0	20.9
yek kam (a bit)	24	8	0.53	0	0	0	20.6

Based on Table 5.16, the three positive 2-grams for the LoPSC are *fekr mikonam* (I think), *yek kam* (a bit) and *mikonam keh* (do that). *Fek mikonam* occurs with a keyness score of 45.6, mikonam keh with a keyness score of 20.9 and *yek kam* with a keyness score of 20.6.

fekr mikonam and *mikonam keh* both occurred in the RC with a relative frequency of 1.61 and 0.14, respectively. Whereas based on Table 5.16, *yek kam* (a bit) does not occur in the RC. However, a manual search of the RC showed that yek kam occurred in the RC albeit with different orthographic forms (i.e. yeh kam (یه کم) and yekam (یکم)). Therefore, with the consideration of the other orthographic forms of *yek kam* in the RC, the actual raw frequency of this form was 17 and its relative frequency is 3.5. Although there is still a discrepancy between the use of *yek kam* in the two corpora, this discrepancy is not significant enough to be considered as a positive keyword.

In addition, from the 1433 occurrences of *mikonam ke,* 1400 hits were part of the phrase *fekr mikonam keh* (I think that). Therefore, *mikonam keh* and *fekr mikonam* are treated as the same entity and are further discussed in the section related to *fekr mikonam.*

Table 5.17 shows the top 3 negative 2-grams of the LoPSC.

Table 5.17

Top 3 Negative 2-grams for the LoPSC

Type	Raw Frequency LoPSC	Relative Frequency in LoPSC	Dispersion LoPSC	Frequency RC	Relative Frequency in RC	Dispersion RC	Keyness Score
Fekr konam (I think)	0	0	0	128	18.82	1.1	0.05
va inâ (and these)	0	0	0	110	16.02	0.73	0.06
Ye dooneh	0	0	0	81	14.83	0.56	0.08

According to Table 5.17, the negative 2-grams for LoPSC are *fekr konam* (I think), *va inâ* (and them/these), and *yeh dooneh* (one), with keyness scores of 0.05, 0.06 and 0.08, respectively. Since *va inâ* occurs in both corpora, this 2-gram will be further discussed in section 5.2.2. However, the two remaining 2-grams, i.e., *fekr konam* and *yeh dooneh* do not occur in the LoPSC. Since there were no cases of the phrase ye dooneh, this 2-gram will not be further discussed in this section. However, in the case of *fekr konam* (I think), since the equivalent of this phrase, albeit in a similar but different form (i.e., *fekr mikonam*) appeared in the LoPSC as a positive keyword, *fekr konam* is discussed in the following section.

5.3.1 fekr mikonam and fekr konam

The equivalent for the English phrase *I think* is *fekr mikonam* in Persian. As shown in Table 5.16, this is a highly frequent 2-gram in the LoPSC with a frequency of 73.66 but it only occurs 1.28 times in the RC. However, the phrase does occur as *fekr konam* in the RC. Although the form *fekr konam* does not occur in the LoPSC. Hence, although the 2-gram *I think* is highly frequent in both corpora, there is a clear preference for using two different forms for the expression of this 2-gram in either corpus. These occurrences are elaborated on below.

5.3.1.1. Fekr konam in the RC

The infinitive form of *fekr mikonam* is *fekr kardan*. *fekr kardan* is a compound verb which consists of the noun *fekr* (thought) and the verb *kardan* (to do). In the simple present tense, *fekr kardan* is conjugated as *fekr mikonam* in the first-person singular tense. However, in the RC, *mikonam* appears, as mentioned, mainly as *konam*.

The question that remains is whether there is a preference for the truncated form *konam* over the complete form *mikonam* in the RC. To answer this question, first, the frequency of the forms was checked in both corpora. Table 5.18 shows the frequency of these forms.

Table 5.18

Frequency of konam and mikonam in LoPSC and RC

Form	Raw Frequency in LoPSC	Relative Frequency in LoPSC	Raw Frequency in RC	Relative frequency in RC
mikonam	49	16.33	62	9.11
konam	33	11	99	14.55

*Note: The verbs *mikonam* and *konam* that were part of the phrases *fekr mikonam* and *fekr konam* were not included in this search.

Table 5.18 shows a higher frequency for *mikonam* in the LoPSC (16.33) compared to the RC (9.11). On the other hand, there is a higher frequency for *konam* in the RC (14.55) in comparison to the LoPSC (11). However, this higher frequency does not indicate if there is a preference for truncating the form *mikonam* to *konam* in the RC, since *konam* can replace either *mikonam* or the first-person singular in the subjunctive form, i.e., *bekonam*. Therefore, all the cases of *konam* in both corpora were checked to see whether *konam* was used as a truncated form for *mikonam* or *bekonam*. This check showed that all occurrences of *konam* were used to shorthen the form *bekonam* and not *mikonam.* Therefore, there is a clear indication that the form *mikonam* is only shortened in the case of *fekr mikonam* in the RC. However, this is not reflected in the LoPSC; that is, the speakers of the LoPSC only use *fekr mikonam*.

In addition, as shown in Table 5.19, the collocations for *fekr konam* in the RC also strengthened the argument for the preference of its speakers for a more shortened

version. That is, the standard form for presenting the epistemic marker *I think* in Persian is *man fekr mikonam keh* (I think that). However, as Table 5.19 shows, in the RC, *fekr konam* collocates with neither *man* (I) nor *keh* which are added to the standard form *fekr mikonam*.

Table 5.19

Top 5 Collocations for fek konam in the RC

Position	Collocate	Statistics
L	konam (I do)	11.99
R	fekr (think)	11.95
R	toman (toman)	11.41
L	âreh (yeah)	11.38
R	nemidoonam (I don't know)	9.90

This is in contrast to both *fekr mikonam* in the LoPSC and fekr mikonam in the RC which, as Table 5.20 and Table 5.21 show, have both *man* and *keh* as their strongest collocates.

Table 5.20

Top Collocations for fekr mikonam in the RC

Position	Collocate	Statistics
L	keh (that)	11.86
L	man (I)	10.01

Table 5.21

Top 5 Collocations for fekr mikonam in the LoPSC

Position	Collocate	Statistics
R	keh (that)	18.56
L	baleh (baleh)	15.64
L	amâ (but)	14.28
L	âreh (yeah)	13.98
L	man (I)	13.56

Nonetheless, since both *fekr konam* and *fekr mikonam* does occur in the RC, the following section looks at the functions of these 2-grams to see if the formal preference for the truncated form by the speakers of the RC also entails functional differences between the two forms.

5.3.1.2. Functions of fekr konam and fekr mikonam

For this analysis, instances of *fekr konam* and *fekr mikonam* were considered as epistemic markers with regard to their form. For instance, *fekr mikonam* occurs 11 times in the RC, but of these 11 times, 3 of the occurrences appear with the auxiliary verb *dâram* to form the present progressive form. Extract 34 shows an example of such an occurrence.

Extract 34

قرارداد رو به اسم تو ببندم ***دارم فکر میکنم***

dâram fekr mikonam gharârdad ro be esme to bebandam

I am thinking to use your name for the contract

Instances such as the example in Extract 34 were deleted, and 8 instances of *fekr mikonam* remained for the analysis of the functions.

In addition, 11 instances of *fekr konam* occurred in the RC that were the truncated form of *fekr bekonam* and not *fekr mikonam*, as shown in Extract 35.

Extract 35

باید در موردش ***فکر کنم***

bâyad dar moredesh ***fekr konam***

(I) have to ***think*** about it

Therefore, the remaining instances of *fekr mikonam* and *fekr konam* were considered for their functional analysis.

A manual coding of the functions of *fekr mikonam* in the LoPSC and *fekr konam* in the RC showed that both forms occurred as an epistemic marker or as a mitigating device. The following extracts show examples from both corpora showing the occurrence of these two functions.

Extract 36 (fekr mikonam as an epistemic marker in the LoPSC)

<Speaker 10> hmm نمیدونم هنوز شصت سال ندارد ***فکر میکنم*** شصت و هشت یا چیز نمیدونم

<Speaker 10> hmm nemidoonam hanooz shast sâl nadard ***fekr mikonam*** shast o hasht sâl yâ chiz nemidoonam

<Speaker 10> hmm I don't know he still isn't 60

years ***I think***

don't know

Extract 36 shows the uncertainty of Speaker 10 when providing an answer to the question about his father's age. This uncertainty is further pronounced by the hesitant marker hmm, the expression nemidoonam (I don't know) and the VCM (ya chiz).

The collocations for fekr mikonam in the LoPSC (as shown in Table 5.21) also show *baleh* (yes) and *nemidoonam* (I ,(that this 2-gram frequently occurs with â*reh* (yeah don't know) to express the speaker's uncertainty when answering a question or expressing a thought as was illustrated in Extract 36.

Similar to Extract 36, in Extract 37, the speaker is answering a question from her interlocutor regarding certain university regulations. The speaker marks her uncertainty by starting her answer with *fekr konam* and follows by changing the answer to her question from 5 to 6.

Extract 37 (fekr konam as an epistemic marker in the RC)

فکر کنم از سال پنجم به بعد ، سال شیشم به بعد

Fekr konam az sâl panjom be bad, sâl shishom be bad

I think from the fifth year onwards, from the sixth year onwards

I think from the fifth year onwards, (from the) sixth year onwards.

The high frequency of the function of uncertainty by the speakers of the RC is also further highlighted by collocates such as fekr *konam*, *nemidoonam* and *toman*, as shown in Table 5.20. The collocation of *fekr konam* with itself and the expression

nemidoonam further marks the hesitancy and uncertainty of the speakers. The collocation of fekr konam with the Iranian currency, *toman*, also shows that *fekr konam* occurs frequently when answering questions regarding the approximate prices of items.

Extract 38 (fekr mikonam as a mitigating contrasting device in the LoPSC)

بله بله تخفیف بود ***اما*** hmm ***فکر میکنم*** او گفت چون آن فروشگاه خیلی معروف در (name-of-city1) و بعضی

ازمردم که علاقه به این کالاها دارن فکر میکنن که آنجا بهترین جا در شهر هست قیمت بالا میره آره حتی با تخفیف

baleh baleh takhfif bood ***amâ*** ***fekr mikonam*** ou goft chon an foroshgâh kheili maroof dar

(name-of-city1) va bazi az mardom keh alâgheh beh in kâlâhâ daran fekr mikonan keh

ânjâ behtarin jâ dar shahr hast gheymat bâlâ mireh âreh hatâ bâ takhfif.

Yes yes there was a discount ***but*** ***I think*** that he said because that shop is very famous in

(name-of-city1) and some people that like these things think that place is the best place in

the city, (so) the prices go up yeah even with a discount.

In Extract 38, the speaker is contradicting what his interlocutor had said in a previous turn about the prices in a certain shop being reasonable due to the discounts available in the shop. The speaker in Extract 38 is contradicting his interlocutor's statement as he believes that the prices are still expensive; however, he uses the phrase *yes but I think* to mitigate the force of his contradiction. As Table 5.21 shows, *fekr mikonam*

collocates frequently with *âreh*, *baleh* and *amâ*, therefore indicating the frequent occurrence of these mitigating contradictory functions in the LoPSC.

However, the mitigating contradictory function did not occur for *fekr konam* in the RC. As it stated in section 5.1 the speakers of the RC used other means of mitigating contradictions, namely the discourse marker *âkheh*.

As for the functions of fekr mikonam, it appeared mainly to mean to *imagine*. This is shown below in Extract 39 from the RC.

Extract 39

فکر میکنم دارم آهنگ گوش میدم

fekr mikonam daram ahang goosh midam

I (will) imagine I'm listening to music

In Extract 39, the speaker uses *fekr mikonam* to say that he will be placing putting headphones on and imagining he is listening to music. Therefore, in Extract 39, *fekr mikonam* is not an epistemic stance nor a mitigating contrastive marker.

This was in contrast to the functions of *fekr mikonam* in the LoPSC and *fekr konam* in the RC; therefore, indicating that the speakers of the RC made a clear distinction regarding the functions of *fekr konam* and *fekr mikonam*. That is, *fekr konam* was used in the standard form to function as an epistemic marker or contrastive device whereas *fekr mikonam* was used to indicate to *imagine*.

5.3.1.3. Summary for the results of fekr mikonam

This section has shown that the speakers of the RC prefer the use of the truncated form of *fekr mikonam* (i.e., *fekr konam*). This finding is significant for two reasons. First, the speakers only use the truncated form of *konam* to replace *mikonam* when it appears in the phrase *fekr mikonam.* Second, the use of *fekr mikonam* in the RC

shows different functions to that of the LoPSC; that is, *fekr mikonam* loses its function as an epistemic marker and is used to mean imagine.

5.3.2. va inâ

According to Table 5.17, *va inâ* (and these) is one of the negative n-grams of the LoPSC with a keyness score of 0.06. *va inâ* appears in the RC with a raw frequency of 109 (16.02) and no frequency in the RC. However, *va inâ* did occur in the LoPSC, albeit in its standard orthographic form (i.e., va inh*â*), with a raw frequency of 5 (1.66). Nonetheless, there is still a significant discrepancy between the frequency of *va inâ* in the LoPSC and the RC.

In addition to the low frequency of *va inâ* in the LoPSC, there is a low dispersion of this form in the LoPSC. This low dispersion is due to it only being used by 2 speakers of the LoPSC.

The analysis of *va inâ,* in this study, showed that it has 2 functions in the RC. Table 5.24 shows these 2 functions with their corresponding frequency of occurrence.

Table 5.22

Functions of va inâ in the RC

Functions	**Raw Frequency**	**Relative Frequency**
Vague Category Marker	101	1485
Associative Plural	9	147

As shown in Table 5.22, *va inâ* has two functions in the RC. That is, this phrase functions as either a Vague Category Marker or an associative plural, which is in line with previous research on the functions of this phrase (Ghomeishi, 2018). On the other hand, although there were occurrences of *va inâ* as a Vague Category Marker in the LoPSC, there were no occurrences of *va inâ* as an associative plural. Therefore, due to the absence of *va inâ* as an associative plural in the LoPSC and its low frequency in the RC, the analysis of *va inâ* as an associative plural was not considered in this study, and only *va inâ* as a Vague Category Marker was taken into account.

5.3.2.1. va inâ as a Vague Category Marker

As shown in Table 5.22, *va inâ* (and these) occurred 101 times (14.85) in the RC as a Vague Category Marker. Vague Category Markers (VCMs) in this study refer to "a relatively homogeneous set of forms consisting of a conjunction plus a noun phrase" that are added to other phrases or clauses (Overstreet, 1999, p.12) for the purpose of creating an ad hoc category (O'Keeffe, 2014, p.19). The meanings of these ad hoc categories are "socio-culturally grounded and are co-constructed within a social group that has a shared socio-historic reality". (O'Keeffe, 2004, p.6). That is, the meanings of VCMs are context-dependent. For example, Extract 40 from the RC shows how M1 uses a Persian VCM, namely, *inâ*, and how his interlocutor deduces the meaning of this VCM through the shared social and cultural background, M2 understands what M1 is referring to when he says *prayers* because of their shared social and cultural *,(and) these*

Extract 40

سارا هم چیز میکنه ؟***نماز اینا***-ست ؟

سارا هم فکر کنم میخونه ، آره ، سارا هم مشروب

نمیخوره .. سیگارم نمیکشه ..... جفتشون خیلی سالم و چیزن

ham chiz mikhoneh? ***namâz inâ*-st? Sara**

Sara ham mashroub ,*â*reh ,Sara ham fekr konam mikhoneh

nemikhoreh…sigâr nemikesheh…jofteshoun kheili sâlem o

chizan

do as well? (Say) ***prayers (and) these***?

Sara ,yeah, I think Sara says (prayers) as well

doesn't drink alcohol as well.. doesn't

smoke…they're both very healthy and

thing

As mentioned in section 2.4, VCMs may appear with or without conjunctions. The same pattern also applies to Persian VCMs.

Extract 43 from the RC shows the occurrence of *inâ* as a VCM.

Extract 43

.. بعد فکر کن سال دیگه بزنن یهو بره از یک و پونصد **یک و هفتصد *اینا*** بشه سه

bad fekr kon sal digeh bezanan yeho bereh az yek o ponsad yek o haftsad ***inâ*** besheh seh

Then think next year suddenly it goes from one (million) and five hundred one million and seven hundred ***these*** to become three (million)

Regarding the frequency of occurrence, from the 186 (27.35) cases of *inâ* (these) that occurred in the RC, 49 (7.20) cases were VCMs.

inâ (*inhâ*) occurs in the LopSC with a frequency of 9. From these nine occurrences, 6 were considered as VCMs. Extract 44 shows an example of *inâ* (*inhâ*) as a VCM in the LoPSC.

Extract 44

اما بستگی داره به چه کار میکنند ***اینا***

amâ bastegi dâreh beh cheh kâr mikonand ***inâ***

To understand what other forms the learners were using instead of *va inâ (inâ)* as a VCM two steps were taken. First, previous literature on VCMs in Persian was consulted to see what forms were identified as VCMs in Persian and whether the learners were using similar forms. The next step taken was to consider the list of positive keywords and n-grams (See Appendix V) to find if there were any words or phrases that were used by the learners instead of *va inâ* (or *inâ*). These two steps are presented in the following.

5.3.2.2. VCMs used in the LoPSC Based on Previous Literature

As was shown in Table 2.1 and Table 2.2 in Section 2.4, VCMs in Persian can be formed using the lemma *chiz*. A search of the LoPSC showed that the speakers of the LoPSC also use VCMs mainly formed with the use of *chiz (thing)*. That is, a search of the corpus using the wildcard *chiz** showed that from the 250 hits (83.33) of the occurrence of the lemma *chiz*, 25 cases (8.33) were categorised as VCMs. Extract 45 shows an example of *chiz* occurring with conjunction to form a VCM.

Extract 45

<<Speaker 7 مامانت hmm به شما کمک میکنه با چمدانت و

همه چیز

<Speaker 7> mamanet hmm be shoma komak mikoneh ba

chamedanet va hame

chiz

<Speaker 7> your mom hmm will help you with the suitcase and all **thing**(s)

However, there were new forms used by the learners that were not stated in the RC or the previous literature. As shown in the following extracts.

Extract 46

<<Speaker12 سختر هست که یک جا hmm ترک میکنیم بعد از این که به یه جا و یه روتین و برنامه ***و همه چیزا*** رو عادت کردیم

<Speaker 12> sakhtar hast keh yeh ja hmm tark mikonim bad az in keh beh yeh ja va yeh rotin va barnameh ***va hameh chiza*** ro adat kardim

<Speaker 12> It's very difficult that when hmm you leave somewhere after you are used to a routine and schedule ***and all things***

In Extract 46, the speaker uses the form *va hameh chiza* (and all things) which is not a standard VCM form in Persian. Similarly, in Extract 47 and Extract 48, the VCMs *hame chiz* and *va chiza* also occur which are not Persian VCMs.

Extract 47

<<Speaker 6 پروازتون رو رزرو کردی ***همه چیز***؟

<Speaker 6> parvazetoon ro reserve kardi ***hame chiz***?

<Speaker6> Did you reserve your flight ***everything***?

Extract 48

<<Speaker 8 به پلاستیک ***و چیزا*** اونو تبدیل میکنند

<Speaker 8> beh pelastik ***va chiza*** ouno tabdil mikonand

<Speaker 8> they change it to plastic ***and things***

In addition to the new forms of VCMs, the speakers of the LoPSC would use adjunctive VCMs instead of disjunctive VCMs and vice versa. Extract 49 shows an example in which an adjunctive VCM is used by Speaker 6 when a disjunctive VCM would have been used.

Extract 49

<<Speaker 6 باید یه هتل ***و چیزی*** رو رزرو کنیم؟

<Speaker 6> bayad yeh hotel ***va chizi*** ro reserve konim?

<Speaker 6> (We) have to reserve a hotel ***and a thing***?

However, in the RC, *chiz* was used less frequently as a VCM compared to the LoPSC. A search for the wildcard *chiz** showed that the lemma *chiz* was used 608 times (89.41) in the RC, but it only occurred as part of a VCM 10 times (1.47). Nonetheless, the speakers of the RC used VCMs, especially the VCM *va inâ* (or *inâ*) significantly more frequently than the speakers of the LoPSC.

5.3.2.3. Summary of va inâ

This section on the comparison between VCMs in the LoPSC and the RC showed four main findings. First, there is a significantly relatively lower frequency of VCMs used in the LoPSC compared to the RC. Second, in the LoPSC, there is a preference for VCMs formed using the lemma *chiz* whereas, in the RC, VCMs formed with this lemma are

the least frequent forms of VCMs. Third, the speakers of the LoPSC used forms of VCMs that were not standard forms of VCMs in Persian, such as *hame chiz*, and *va chiza*. Finally, in the LoPSC, the speakers would use adjunctive VCMs in contexts in which a disjunctive VCM would have been used and vice versa.

5.4. Frequency Lists

Up to this point of the analysis, the main differences between the LoPSC and RC have been identified using a keywords' list analysis. However, there are features of Conversational Persian that are difficult to identify using automated corpus tools, such as the keywords' list. For example, as explained in section 3.3., in Conversational Persian and especially in more informal interactions, the verb ending for the simple present tense for the second person singular changes from *-id* to the suffix *-i.*

To identify such features, first, a frequency list containing the 50 most frequent words for the LoPSC and the RC was generated. (See Appendix VII and Appendix VIII). This was followed by a manual inspection of the two corpora to see if there were any significant differences between the two corpora that were left unnoticed by the keywords' analysis. This observation showed two main differences between the LoPSC and the RC. First, there was a lower frequency of the substitution of the phoneme /u:/ for the phoneme /a:/ in the LoPSC. This substitution is a common feature of Conversational Persian (Miller, 2011). As the frequency lists showed, this feature was not frequently used by the learners of the LoPSC.

It is to be noted that notice of this phonetic difference was made based on the orthographic script and not a phonetic transcription of the data. As was mentioned in Section 4.4. Transcription, in Conversational Persian, certain differences in the phonemes used in "Standard" and Conversational Persian are also represented in the orthographic script without the necessity of a phonetic transcription.

Second, in Conversational Persian, there is also the frequent substitution of the verb ending *-ad* to *–eh* in third person singular verbs in the simple present tense (Saffar Moghaddam, 2013). The comparison of the frequency lists of the LoPSC and the RC also showed that the speakers of the LoPSC used this feature significantly less compared to the speakers of the RC.

5.5. Results from the Interviews

As was mentioned in section 4.6 Interviews, the interviews conducted in this study served two purposes. The first was to identify the learner profile of the learners by asking about their first language, age, gender, any other second languages they spoke, whether they had travelled to Iran or not, and the average number of hours they spent practicing spoken Persian with a first language speaker of Persian. The answers to these questions have been illustrated in Table 4.2. as part of the description used of the LoPSC.

In addition to these questions, the students were also asked a set of questions designed at gathering further information that could potentially justify certain decisions that learners made regarding the use of Conversational Persian. These questions included the types of material the learners used for learning Persian, in general, and Conversational Persian, specifically. The learners all stated the use of the taught material in their university tutorials and conversations with first language speakers as their main input for Conversational Persian.

The were two other questions included in the interviews, namely, what the students thought were the most important aspects of Conversational Persian and under what circumstances (if any) they would use Conversational Persian. Regarding the former question, 16 out of 18 of the participants stated the non-use of the second person

plural to refer to the second person singular, the substitution of the phoneme /u:/ for the phoneme /a:/, and the change in the verb ending for the simple present tense in the second person singular from -id to -i, to be the main features of Conversational Persian. In addition, 2 out of 18 of the participants referred to the use of different vocabulary as the main feature of Conversational Persian when compared to the Standard Written Form.

As for the latter question regarding instances in which the participants would find themselves deciding to use Conversational Persian, 17 out of the 18 participants found speaking in Conversational Persian to be highly informal and something they would generally refrain from using or as a variety of the Persian Language that was specific to L1 users of Persian. One out of the 18 participants stated that they would entirely avoid the use of Conversational Persian since it was “too informal”.

To reiterate, the purpose of this set of interview questions was to provide a possible explanation for the differences between the LoPSC and RC that have been reported on in earlier sections of this chapter. These possible explanations are further elaborated on in Section 6.5 Factors influencing the differences.

5.6. Summary of Chapter

The overall results show three trends regarding both the formal and functional words and multi-word units used by the speakers of the LoPSC in comparison to the speakers of the RC. One is the overall preference in the LoPSC for forms associated more with the formal written language in Persian rather than colloquial Persian forms. This was attested by the use of *bâraye* instead of *vâse*, *amâ* instead of *vali* and the discourse marker âkheh, *baleh* instead of *âreh*, and *fekr mikonam* instead of *fekr konam*. In

addition, forms closely associated with spoken conversations were used with a significantly lower frequency in the LoPSC compared to the RC. These forms included VCMs and discourse markers.

Second, there were new forms in the LoPSC that were not used in the RC nor reported in previous literature. This was specifically the case for VCMs used in the LoPSC. In addition to using the novel forms, the speakers of the LoPSC would also confuse the appropriate use of adjunctive and disjunctive VCMs.

Finally, regarding the differences in the use of functions, first, there was an over-reliance on the use of certain forms to perform a large number of functions, such as the case of *amâ.* Second, there was a mismatch between the use of functions for certain forms in the two corpora. For example, the speakers of the LoPSC used contrastive connectives to mitigate contradictions made towards their interlocutor's prior statements. However, the speakers of the RC used contrastive connectives to show a more direct and emphatic approach in their contradictions and preferred the use of discourse markers to mitigate the force of the contradictions made. Finally, the use of the same form showed completely different functions for the two corpora, such as the case of *fekr mikonam* which was used as an epistemic marker in the RC but was used for a different purpose in the RC and the case of the affirmative marker *baleh*.

6. Discussion Chapter

6.1. Introduction

Based on the results of this study, the formal differences between the learners and the L1 speakers of Persian can be divided into two categories of differences: pronunciation and word choice. In addition, further analysis of the word choices of the two groups of speakers also showed functional differences. The results of this study are considered in light of the previous research on learners of Persian in the following chapter. Finally, this chapter ends by presenting a series of factors influencing the differences in the choice of Conversational Forms between learners and L1 speakers of Persian.

6.2. Formal differences between learners and L1 speakers' language use

6.2.1. Differences in pronunciation

The most significant difference regarding pronunciation was related to the pronunciation of the long vowel of */ɑ:/.* In similar contexts, the learners preferred using the long vowel whereas their L1-speaker counterparts substituted */ɑ:/* for */u:/.* For example, the learners pronounced the direct object marker */rɑ:/* with a significantly higher frequency compared to the speakers of the RC, who preferred pronouncing this marker as */ru:/.*

Although, in this study, this was the most significant finding regarding differences in pronunciation between speakers of the LoPSC and the RC, this difference was not reported in previous research on Persian language learners. The reason behind this may be due to the spoken variety of Persian in the RC. That is, the speakers of the RC were all speakers of Tehrani Persian, and the change of the long vowel */ɑ:/* to */u:/ is* a characteristic of spoken colloquial Persian in Tehran (Kahn and Bernstein, 1981). Therefore, although this difference in pronunciation is a significant finding in this study,

previous studies may have left out or even not encountered the change in the pronunciation of */u:/* due to exploring other varieties of Persian other than the Tehrani Persian or ignoring, what they may have considered being, idiosyncrasies of a certain variety.

However, since more speakers of Persian are using Tehrani Persian as the standard form (Sedighi and Shabani, 2018), in recent studies, the change of the vowel */ɑ:/* to */u:/* has been considered as the most important phonological feature of spoken colloquial Persian (Miller, 2011). In addition, this change in vowels is not only specific to the Tehrani variety of Persian and is seen in other varieties of Persian, such as Isfahani Persian (Miller et al., 2014). Moreover, although with a lower frequency compared to the speakers of the RC, the learners also changed the vowel */ɑ:/* to */u:/*.

In addition to the corpus data, the interviews with the learners also showed that they considered the change in the pronunciation of the vowel */ɑ:/* to */u:/* to be the most noticeable characteristic of spoken colloquial Persian. Therefore, although this change in pronunciation may have been left out by previous research due to its connection with a specific language variety, namely, the Tehrani variety, recent studies on colloquial Persian and data from the corpora and the interviews used in this study indicate that this change requires more attention when teaching and researching on colloquial Persian used by learners.

In addition to the difference in frequency of the change of the vowel */ɑ:/* to */u:/,* the other significant difference in pronunciation between the LoPSC and RC was related to morpho-syntactic changes. That is, there was a difference between the frequency of use of the suffix *–ad* and *–eh* to end third-person singular verbs in the simple present tense. That is, the learners preferred the use of the suffix *–ad* whereas the L1 speakers

used the suffix *–eh*. The difference between the frequency of the two forms was most significant for the form *dar-eh (she/he/they/it has)*.

This finding reflects findings from previous studies on colloquial Persian; in that, the change of the suffix *–ad* to *-e* at the end of present tense verbs for the third person singular is considered a feature of spoken colloquial Persian, especially in the case of the verb *dashtan* (to have) (Miller et al, 2014; Saffar Moghaddam, 2013). Therefore, based on the results of this study it can be concluded that the learners of this study use this feature of colloquial Persian with a significantly low frequency. This once again highlights the importance of using research on colloquial Persian in the context of teaching Persian.

6.2.2. Differences in word choices

Another significant finding from this study was the reported differences between the word choices of the two groups of speakers. Since this study is the first to compare the differences between the word choices between learners and L1 speakers of Persian, the pattern of categorisation chosen reflects the results of the current study. The differences found can be divided into two categories. The first category includes forms that were used interchangeably in the two corpora. The second category deals specifically with the significantly lower occurrence of discourse markers and vague Category Markers in the LoPSC compared to the RC. Each of these categories will be described in the following sections.

Words Used Interchangeably

Table 6.1 presents a list of the most significant words that were used interchangeably in the two corpora.

Table 6.1.

Interchangeable words across the two corpora

LoPSC	RC
ou (he/she/they)	*oun (he/she/they/it)*
baleh (yes)	*areh (yeah)*
ama (but)	*vali/akheh (but/discourse marker)*
fekr mikonam (I think)	*fekr konam (I think)*
kami (a bit)	*zareh (a bit)*
barâyeh	vâseh

Similar to the previous section on pronunciation differences, the word choices of the two groups were not reported on in previous literature on learners of Persian (Ghaffari, 2020). Therefore, this study has provided a list of significant differences between the word choices of the two speakers.

The results of this study also add to previous research in colloquial Persian. First, although previous literature on colloquial Persian has reported the preference for *oun* (Ghomeishi, 2018), *areh* and *vali* (Alami, 2012) in colloquial Persian, new forms common to this variety of language, namely, *akheh* instead of *ama*, *fekr konam* instead

of *fekr mikonam* and *zare* instead of *kami* were findings from this study that were not reported in previous research on colloquial Persian. Since the learners used these forms with a significantly lower frequency or with no frequency in their conversations, the teaching and learning context of colloquial Persian would benefit from including these forms. The inclusion of these forms in teaching colloquial Persian is especially of importance, since the different forms listed in Table 6.1 demonstrated different pragmatic functions in the two corpora.

In addition to the preference for different words in each corpus, discourse markers appeared with a significantly higher frequency in the RC compared to LoPSC. Therefore, the following section looks at the use of discourse markers in the two corpora specifically.

6.3. Discourse Markers

The low frequency of discourse markers used by the learners was evidenced regarding both occurrence and types. That is, based on the discourse markers that were considered in this study, the speakers of the LoPSC used significantly fewer discourse markers with less variation in forms. The following section first deals with the frequency of occurrence and the next section deals with the frequency of type.

6.3.1. Frequency of Occurrence

Since this study is the first to observe the use of discourse markers by learners of Persian, there is no previous research to compare the findings of this study in the context of Persian Interlanguage; however, the lower frequency of occurrence of discourse markers in the LoPSC reflects the findings from previous literature on the interlanguage of learners from other first and target languages. For example, Fung

and Carter (2007) found that Cantonese Speakers of English used significantly fewer discourse markers compared to their English-speaking counterparts (See Section 2.6. for more similar studies).

6.3.2. Frequency of types

Considering the type of discourse markers used, the most frequent discourse markers used in the RC also correspond with Alami's (2016) study of discourse markers in colloquial Persian in the informal register. Alami found *bebin* and *aslan* to be two of the highest-frequency discourse markers in a corpus of spoken interactions in informal colloquial Persian. This highlighted the high frequency of discourse markers in spoken interactions as evidenced by previous research (See Section 2.4.3.). This also underscores the lack of use of discourse markers closely associated with informal colloquial Persian by the learners.

In addition, the discourse markers used in the RC were more interactive rather than textual regarding their function (namely, in the case of *bebin* (look), *akheh* (adversial discourse marker with *no* equivalent in English), *hey* (again), and *boro* (go)). This echoes findings from previous research on the use of discourse markers by speakers in that the discourse markers used are more interactional rather than textual. For example, in English, this is reflected by the use of discourse markers, such as, *you know, I mean,* and *like* (See Section 2.6).

However, the types of discourse markers used in the LoPSC were less interactive compared to the RC. The use of less interactive discourse markers by learners is also reflected in previous research on other target languages. Fung and Carter (2007) found that whereas the learners of their study used discourse markers such as *and*, *but*, *I think*, *yes*, *OK*, *so* frequently, the discourse markers that were frequently used

by the English speakers were less frequent in the learner speech, especially discourse markers functioning in the interpersonal category. This also echoes the findings of this study in which discourse markers that appeared as negative keywords were used in the interpersonal category whereas the discourse markers that appeared as positive keywords were used for referential and structural purposes (i.e., in the textual category compared to the interpersonal category). Similar findings were reported by De Cock (2004) with French learners of English.

In the same line, Vyatkina and Cunningham (2015) conclude that based on the word-based research on interlanguage pragmatics using a learner corpus and Contrastive Language Analysis, learners tend to "overuse discourse markers intensifiers and modality expressions associated with certainty, directness and authority, but underuse interpersonal hedges and expressions of doubt and uncertainty" (p. 287). Vyatkina and Cunningham also concluded that with the increase in their level of proficiency in the L2, an increase in the proficiency of their interlanguage pragmatics is generally observed. As mentioned earlier, since this study is the first to look at the use of discourse markers by learners of Persian, there is no base for comparison regarding the different levels of proficiency of the learners. Nonetheless, from the results based on this group of specific learners, it can be concluded that learners of Persian demonstrated a low frequency of using the types of discourse markers, especially regarding discourse markers that served an interactional function.

The difference between the frequency of occurrence and types was also particularly salient regarding the occurrences of Vague Category Markers (VCMs). Therefore, the next section deals with this form, specifically.

6.3.3. Vague Category Markers

Regarding the differences between how speakers of the two corpora used VCMs, the findings of this study can be summarised as follows:

1. There is a relatively lower frequency of VCMs used in the LoPSC compared to the RC.

2. In the LoPSC, there is a preference for VCMs formed using the lemma *chiz (thing)* whereas, in the RC, VCMs formed with this lemma are the least frequent forms of VCMs.

3. The speakers of the LoPSC used forms of VCMs that did not occur in Persian, such as *va chiza (and things)* and *ya chiz* (*or something*).

4. Disjunctive VCMs appeared with a significantly higher occurrence in the LoPSC compared to the RC whereas adjunctive VCMs occurred with a higher frequency in the RC compared to the LoPSC.

Each of the following findings is explored further in the following sections.

Frequency of VCMs

The results of this study showed that VCMs were used with a significantly lower frequency in the LoPSC compared to the RC. Regarding the frequency of use of VCMs in studies that have compared learner language with L1 speakers, there is no single consensus. That is, some studies report a higher frequency of VCMs used by learners whereas others report less or a similar frequency of use.

Regarding studies that have reported a lower frequency of VCM use by learners, for example, DeCock (2004) looked into the differences between the above 2-grams used by French advanced learners and L1 speakers of English. Her study used a corpus-driven approach by using two corpora, namely, the French component of the Louvain International Database of Spoken English Interlanguage (LINDSEI) as the learner

corpus and the Louvain Corpus of Native English Conversation (LOCNEC) as the L1 speaker corpus. DeCock found that vague category markers were significantly underused by the French speakers.

In the same line and using the same corpora as DeCock's (2004) study, when exploring the differences between the use of N-grams by Czech and L1 speakers of English, Zvěřinová (2016) found that VCMs were used with less frequency and with different types compared to the L1 speakers. This also resonates with Larsson Aas' (2011) study on Swedish compared to L1 speakers of English. Similarly, from their study of learners from various backgrounds (namely, Arabic, Chinese, German, Korean, Spanish, Polish, Russian, Turkish, and Uzbek), Fernandez and Yuldashev (2011) found that learners used fewer VCMs, especially disjunctive VCMs. Metsä-Ketelä (2012) found similar results in a corpus-based study of learners of English from 40 different language backgrounds.

On the other hand, some studies have reported a higher frequency of the use of VCMs by learners compared to L1 speakers. In a corpus-based comparative study, Ruzaite (2018a) studied the use of general extenders by Lithuanian advanced learners of English and L1 speakers of English. The results of her study showed that the learners significantly used VCMs more in written argumentative essays when compared with their L1 counterparts.

To account for the difference between the frequency of the use of VCMs by learners and L1s speakers, there are two reflections from Ruzaite's study on the use of VCMs that are pertinent to the current study. First, Ruzaite's (2018a) study looked into written argumentative essays whereas this study and former studies which found a lower frequency of the use of VCMs by learners considered the spoken informal conversations. As indicated by Ruzaite, the frequency of the use of VCMs was

relatively low in both the L1 and learner corpora, which reflects the characteristic of the genre used in both corpora. That is, in general, texts associated with the formal written genre use fewer VCMs, especially when compared to the informal spoken genre (Crystal, 1975; Channell, 1994; Jucker et al., 2003; Martínez, 2011). Therefore, the current study follows the trend with previous studies on the use of VCMs in informal spoken conversations; that is, there is a lower frequency of the use of VCMs in learner corpora. In addition, the lower frequency of use of VCMs is in line with the other findings from the learners of this study; in that, there is a general tendency to use forms less associated with the spoken informal form of Persian.

However, the results from the current study diverge from other studies on the use of VCMs by learners in two ways. First, previous studies have reported a higher frequency of VCM forms more closely associated with the formal genres whereas this was not the case in the current study. For example, DeCock (2004) found that the French learners of her study used *and so on* and *et cetera* which are associated with formal genres more than L1 English speakers. She believed that this "add[ed] to the impression of detachment and formality they [(the learners)] may well give in informal situations" (DeCock, 2004, p.236). Similarly, DeCock also found that Italian learners of English used *and so on* significantly higher than L1 speakers of English. However, the learners of the current study did not use any of the formal VCM forms in Persian.

In addition, to the high frequency of *and so on* and *etc*., other studies have shown the low frequency of the fewer variation and frequency of VCMs more closely associated with spoken informal contexts, such as, *and stuff* in English. For example, in a study of the asynchronous computer-mediated conversations between L1 and L2 speakers of English, found that *and stuff* was used significantly less by L2 speakers compared

with L1 speakers. In the same line, significantly lower frequencies of the use of *and stuff* by Cantonese speakers of English have been reported (Drave, 2001).

However, in the current study, although the learners did not use the most frequent form of VCM in informal colloquial Persian, namely, *va ina*, they also did not use formal forms of VCMs, similar to learners from other studies. In fact, the learners used no VCM forms associated with Persian VCMs. Alternatively, the learners of the LoPSC resorted to using new forms instead of forms previously identified as Persian VCMs.

The second point of diversion of this study from other studies is that although the frequency of occurrence of the VCMs in this study was significantly less, the frequency in the variation of types of VCMs used by the speakers of the LoPSC was significantly higher than the RC. This finding contradicts previous studies in that although there is no single consensus on whether learners use less or more VCMs, findings from previous studies all indicate that learners use fewer variants of VCMs, compared to their L1 counterparts.

The second finding from Ruzaite's studies on VCMs which is relevant to this study is her observations from comparing the frequency of use of VCMs used by L1 speakers of English and L1 speakers of Lithuanian (Ruzaite, 2018b). She observed that contrary to the high frequency of use of English VCMs, the use of VCMs in Lithuanian is far less frequent. In addition, unlike English VCMs which have been reported to have a higher frequency in spoken interactions compared to their written counterparts of the language, (Ruzaite, 2018b) found this not to be the case in the spoken and written mode of her corpus-based study in Lithuanian. Therefore, a reason behind the use of VCMs with a significantly lower frequency by learners could be the result of a lower frequency of the use of VCMs in the L1 of the learners. However, in the case of this study, the learners of this study all shared English as their L1, and the use of VCMs in

the spoken form of English is highly frequent. Nonetheless, whether VCMs are used more frequently in Persian compared to English has not been studied.

In conclusion, although based on previous studies and the patterns observed in this study regarding the use of forms less associated with the spoken form by learners, it can be concluded that the use of a lower frequency of VCMs is more related to other factors such as exposure to spoken forms and perceptions about the spoken form rather than L1 interference. (These factors will be further explored in section 6.5).

Nevertheless, based on the findings of this study, the possibility of the interference of the L1 although not regarding the frequency but the form of the VCMs is still a possibility to be considered for future studies for two main reasons. First, there was a non-occurrence of different types of VCMs that occur in Persian but do not have an equivalent form in English. Based on their study of the vague language structure and pragmatic use in informal colloquial Persian, Parvaresh and Tayebi (2014) reported the occurrence of three vague forms that were found in Persian but did not have an equivalent in Persian, namely, 'rhyming words', 'replacing expressions', and 'affective completers'. However, none of these forms appear in the LoPSC. This further strengthens the hypothesis for the lack of exposure or lack of noticing of VCMs by the learners in this study. This will be further discussed in section 6.5.

In addition to the non-occurrence of the three vague forms unique to Persian in the LoPSC, there are strong indications that L1 transfer has also influenced the choice of VCMs used by the learners of this study, especially regarding the higher frequency of the lemma *chiz* (*thing*) and the appearance of new forms in the LoPSC. These two findings will be further explored in the following section.

VCM Forms: Lemma chiz and New VCM forms in the LoPSC

As the results showed, there was a higher frequency of the use of the VCMs formed by the lemma *chiz* in the LoPSC. Whereas the most frequent VCM in the RC; i.e., *va ina,* occurred with a significantly lower frequency in the LoPSC. In addition to this discrepancy in the frequency of the different forms used by speakers of the LoPSC and RC, the learners also used forms that were not reported in previous studies on Persian VCMs. Since both findings strongly point to the interference of the learners' L1 (i.e., English), they are both considered in this section.

Generally, studies have shown the effect of the learners' L1 in the formation of VCMs. Aijmer (2015) showed that Swedish, Dutch, German and French learners of English used fewer variants of VCMs when speaking English. Aijmer also found that certain forms were preferred more than other forms; therefore, leading to the use of a higher frequency of certain forms of VCMs by the learners compared to L1 English speakers. As an example, she found that *or something* was the preferred choice of VCMs for the Dutch, German and Swedish learners. Aijmer attributed this finding to the influence of the learners' L1, since the equivalent of *or something* was found in Dutch, German and Swedish, and with a wider variety of functions compared to the VCM *or something* in English. This hypothesis of the influence of the learners' L1 was strengthened even more since Aijmer also included data from French learners of English in her study and did not find similar findings in the French sub-corpus; that is, an equivalent for *or something* in French is not used as frequently as in Dutch, German and Swedish. This finding was also reflected in the speech of Dutch speakers of English.

Similarly, in a corpus of conversations between German and New Zealand speakers of English, Terraschke (2007) found that there was a significantly higher frequency of the use of VCMs by German speakers of English, which was directly influenced by the

significantly higher occurrence of one form of VCM by German speakers, namely, *or so*. To explore the reason behind the higher frequency of *or so*, in a consequent study, Terraschke (2010) used three corpora to compare the use of the English VCM, *or so* and its German equivalent *oder so*. The three corpora compromised of a total of 224,000 words. The first set of data was from New Zealand speakers of English, the second was set from German speakers of German and the final set was from conversations between German and New Zealand speakers of English. The results of Terraschke's study showed that the use of *or so* was different in terms of both the frequency and the functions of use in the New Zealand L1 corpus and the German learner corpus. This difference was attributed to the "wider semantic scope" of the German equivalent of *or so*, i.e., *oder so*, as evidenced by the German corpus (p. 449); thus leading to a higher frequency of the use of *or so* in the learner corpus.

In the same line, L1 transfer of VCM forms has been reported with learners with different L1 backgrounds. For example, Zvěřinová (2016) found that the Czech speakers of English used the vague category marker *and so on so* which has not been reported to be used by L1 speakers of English, but its equivalent was found in the Czech language. Similarly, in a study of the use of VCMs by Persian learners of English (Parvaresh, Islami, and Rasekh, 2012) found that the learners used the VCM form *and and and*, which was not reported as an English VCM, but its equivalent in Persian, namely, *va va va*, functioned as a VCM in Persian (p. 277).

In the case of this study, the speakers of the LoPSC used VCMs formed with the lemma *chiz* more than the speakers of the RC. However, the VCM form with the highest frequency in the RC was *va ina (and these*). *va ina* was also found to be the most frequent in previous studies on Persian VCMs in colloquial Persian (Parvaresh and Tayebi, 2014). In addition, the lemma *chiz* was also reported as the most frequent

lemma used to form VCMs in Persian (Parvaresh et al., 2012; Parvaresh and Tayebi, 2014). This finding resonates with studies on English VCMs; that is, the nearest equivalent to *chiz* in English is the semantically empty lemma *thing* (Ghavamnia and Eslami-Rasekh, 2013) which is used as the main lemma in forming VCMs in English (Overstreet, 1999). In addition, *chiz* can also be considered as an equivalent for *something* which also forms VCMs in English, especially disjunctive VCMs such as *or something,* and another semantically empty noun*, stuff*, which is especially frequent in forming VCMs in informal spoken conversations in English (Jucker et al., 2003; Carter and McCarthy, 2006).

However, as the findings of this study and previous studies on Persian VCMs show, Persian differs from English in that the most frequent of VCMs in spoken Persian are not formed using a semantically empty lemma but the pronoun *ina* (these) is used instead. Nevertheless, since the speakers of the LoPSC all shared English as their L1, based on the frequency of occurrence, there was a clear preference for using the lemma *chiz* to form VCMs by the learners.

In addition to the preference for using the lemma *chiz* by the speakers of the LoPSC, the new forms that were used by the learners also pointed to a strong tendency for transfer from English to Persian. The new forms used by the learners of this study were *va chiz-ha (and things)* and *ya chiz (or something).* These forms have not been reported in previous research on Persian VCMs, and they were not found in the RC, but they appeared with a significantly high frequency in the LoPSC. The English equivalents for *va chiz-ha* and *ya chiz* are *and things* and *or thing/something*. Therefore, although these forms are not considered as VCMs in Persian, they are frequently occurring VCMs in English (Overstreet, 1999). This again strengthens the

hypothesis for considering L1 transfer as the reason behind the higher frequency of occurrence for the lemma *chiz* and the appearance of new forms in the LoPSC.

Adjunctive and Disjunctive VCMs

Disjunctive VCMs appeared with a significantly higher frequency of occurrence in the LoPSC compared to the RC whereas adjunctive VCMs occurred with a higher frequency in the RC compared to the LoPSC. The higher occurrence of adjunctive VCMs in the RC asserts findings from previous research on informal colloquial Persian (Parvaresh et al, 2012). However, the findings from this study diverge from previous findings in that there are only 4 occurrences (67 occurrences per million words) of disjunctive VCMs in the RC. In addition, these four occurrences all follow the same structure (i.e., *ye* (*one*) + *hamchin* (roughly translated as *in this way*) + a form of the lemma *chiz* (*thing*)). This is in contrast to Parvaresh et al.'s (2012) findings from Persian VCMs in which *ya chiz-i* (*or a thing*) occurred with the highest frequency for disjunctive forms. In the current study, there were no instances of *ya chiz-i* in the RC.

Regarding the use of disjunctive VCMs in the LoPSC, as mentioned there was a significantly higher frequency of these VCMs in comparison to adjunctive VCMs. However, similar to the RC, the form that appeared with the highest frequency was not *ya chiz-i,* but the form *ya chiz* (*or thing*). Due to the similarity between *ya chiz-i* and *ya chiz,* the initial impression may be that the learners' choice of *ya chiz* is due to an error. However, as mentioned in the previous section on the forms of VCMs, *ya chiz* is also translated roughly into *or something*, since the pronoun *something* does not have an equivalent in Persian, the noun *chiz* is used in translations from English to Persian. Therefore, there is a higher probability of L1 transfer for two reasons. First, *or something* is a highly frequently used VCM in English (Overstreet, 1999). Second, the strongest collocation for *ya chiz* in the LoPSC was the word *dige*, which when

occurring with the phrase, *ya chiz*, translates roughly into *or something else*. Since, *or something else* is also a VCM in English, this strengthens the hypothesis for L1 transfer instead of learner error.

Another finding from this study that diverges from previous studies on Persian VCMs, is the significantly higher frequency of VCMs without a conjunction. Previous studies on Persian VCMs have only reported the occurrence of VCMs with conjunctions (Parvaresh et al., 2012; Parvaresh and Tayebi, 2014) whereas the current study shows that the speakers of the RC preferred dropping conjunctions for the formation of both adjunctive and disjunctive VCMs. However, in contrast, the speakers of the LoPSC did not drop the initial conjunctions when forming VCMs. The main reason behind the learners' choice of keeping the conjunctions could be related to the learners' L1. That is, although, in English, VCMs can occur with and without the initial conjunctions, English speakers tend to keep the initial conjunctions (Overstreet, 1999).

Although the structure of the adjunctive and disjunctive VCMs in the learner data strongly points to L1 transfer, this factor does not seem to play a definite role in the higher frequency of disjunctive VCMs. That is different studies on the frequency of the use of adjunctive and disjunctive VCMs by L1 speakers of English, for example, report contrasting results. As an example, Cheshire's (2007) finding on L1 British English speakers used adjunctive VCMs more than disjunctive VCMs. However, Overstreet's (2005) findings from L1 American English speakers showed that the speakers used disjunctive VCMs with a higher frequency. Since 17 speakers of the LoPSC had British English as their L1, and there was only one American English speaker, the influence of the L1 in the case of the frequency of the VCMs is difficult to justify.

In the case of the influence of the interlanguage on the frequency of the types of VCMs, findings from the study of the use of VCMs by Dutch learners of English were similar to this study; i.e., the learners used disjunctive VCMs more frequently than adjunctive VCMs, and the L1 speakers of English used adjunctive VCMs more frequently than the learners (Aijmer, 2015).

Therefore, considering that L1 and learner language are not definite factors in influencing the higher frequency of the disjunctive VCMs, since adjunctive and disjunctive VCMs have different functions, the effect of functions on the frequency of the forms is also considered in studies on the use of VCMs by learners. Such studies mainly report the use of similar functions to L1 speakers by L2 speakers. Aijmer (2015) found that similar to L1 speakers of English, Swedish speakers of English used VCMs as referring expressions, to establish rapport, and as hedging devices. Cheng and Martin (2001) and Cheng (2007) also found that contrary to previously held beliefs reflected in the design of language resources, not only did VL use in intercultural conversations between Hong Kongese and L1 speakers of English not impede the natural flow of conversation, but it also led to cooperation between the interlocutors. For example, VL was found to be used for accommodating purposes, such as simplifying words that were thought to be too technical or difficult for the addressee to understand (See Section 2.5.3).

However, deficiencies of functions were also reported. Aijmer (2015) found that the Swedish English speakers in her study needed more instruction on how to use vague forms to perform functions, such as continuing to hold the floor in cases where there were gaps in their lexical knowledge. Drave (2001) also found that in the case of the general noun “stuff”, Hong Kongese L2 speakers used fewer varieties of functions of this vague form compared to L1 speakers in intercultural conversations.

In addition, examining VCMs in Czech, observed that certain functions of English VCMs, such as soliciting agreement, were not a function for VCMs in Czech (Zvěřinová, 2016). They found that Czech speakers used other means to perform this function. Other studies have shown functions that were not observed for certain English VCM forms. As an example, the equivalent of “and everything” in certain dialects of Arabic, namely the Iraqi dialect, was seen to function as an adverbial intensifier preceding adjectives due to the grammaticalisation of this VCM form (Fargel and Haggan, 2005).

In the case of this study, the use of fewer adjunctive VCMs by the learners points to distancing rather than creating a sense of solidarity, which is the main function of adjunctive VCMs (O’Keeffe, 2004). In addition, adjunctive VCMs have also been shown to establish a sense of rapport and add to the informality of the situation (Jucker et al., 2001; O’Keeffe, 2004). This is linked to other pragmatic differences that were shown in the data and are further discussed below.

6.4. Pragmatic differences between learners and L1 speakers’ language use

These pragmatic differences are reflected in this section based on the individual forms analysed in the previous chapter.

6.4.1. baleh and âreh

The results showed that the speakers of the LoPSC used the two affirmative markers, *baleh* and *âreh*, with similar patterns, regarding frequency, turn positions, collocations and functions. However, as expected considering the differences in register, the two forms, *baleh* and *âreh*, did not show these similarities in the RC. That is, whereas *âreh* showed similar turn positions, collocations and functions to the use of *baleh* and *âreh* in the LoPSC, *baleh* was used in the RC to convey different attitudes and to create

certain effects. Therefore, unlike the speakers of the LoPSC, the speakers of the RC made a distinction between these two forms and hence restricted the use of the formal *baleh* only to situations in which they intended it to be used for certain purposes.

On the differences between *baleh* and *areh* in Persian, Sharifi (2012) found that *areh* is significantly more frequent in informal colloquial Persian whereas *baleh* is significantly more frequent in formal colloquial Persian. However, Sharifi (2012) made no account of the reasons behind this difference in frequencies and whether the difference in frequency was related to differences in the pragmatic functions of the two terms.

Regarding the pragmatic differences in affirmative markers in languages other than Persian, Wouk (2001) looks at the two allomorphs used in Indonesian equivalent to the English *yes*. They conclude that the two allomorphs, *ya* and *iya*, in addition to being used as affirmative markers are used as a continuer marker, this was also similar to the functions that *baleh* and *areh* showed in both corpora of this study. Wouk (2001) also found that the form *ya* was used in tag questions as a discourse marker for the purpose of creating a sense of solidarity between the interlocutors. Whereas the form *iya* did not show to have this function. In the current study, *areh* was also used to form tag questions in the RC, but it was not used for this specific function in the LoPSC.

In the case of the use of *yes* and *yeah*, in English, in a study of the differences between the discourse markers used by Cantonese speakers of English and British speakers of English, Fung and Carter (2007) also found that *yes* was one of the highly frequent words whereas *yeah* was not in the learner corpus and this was the opposite in the L1 speaker corpus. However, in this study, the Persian equivalent of *yes* did occur with a significantly higher frequency in the learner corpus compared to the RC placing it as a positive keyword. Contrary to Fung and Carters' study, the Persian equivalent of

yeah was also a high-frequency word occurring as the seventh most frequent word in the LoPSC (raw frequency of 480) compared to the ninth most frequent word in the RC (raw frequency of 702).

Another noticeable difference between the use of the Persian equivalent of *yeah* in this study and that of Fung and Carter's study is that although Fung and Carter state that *yeah* is used to "acknowledge, agree, affirm, and mark continuation" (p. 431), in the data of this study it was used alongside *but* as a marker of disagreement and therefore reflecting the *yeah but* structure used in speakers' L1, that is English.

6.4.2. fekr mikonam and fekr konam

In the case of *fekr konam* and *fekr mikonam*, there are two issues to be considered. First, the issue of the case of formality in pronunciation. This can be compared to using and not using contradictions in English. For example, in her study of the N-grams used by Czech speakers of English, Zvěřinová (2016) found that the expression *I would like to* appears as one of the most frequent n-grams of the Czech speakers' corpus whereas it does not appear to be the case for L1 speakers of English from the LOCNEC who preferred the use of the contracted form *I'd like to* which appears as one of the high frequency n-grams. However, in the case of *fekr mikonam* and *fekr konam*, there are two points that separate this from the case of contractions in English. First, *fekr konam* does not appear at all in the LoPSC. That is, unlike the contractions which occur in the learner corpora of Zvěřinová's study, albeit with a lower frequency, there are no instances of *fekr konam* in the LoPSC. Second, the speakers of the RC indicate a strong preference for the use of this form which further indicated a difference in functions between the two forms *fekr konam* and *fekr mikonam* in the RC. Therefore, the case of *fekr konam* and *fekr mikonam* goes beyond the case of sounding more or less formal and displays functional differences in informal colloquial Persian.

In addition, the equivalent of *fekr mikonam* in English *I think* has appeared as a significantly high-frequency multi-word expression in other studies of second language learners in various other languages. For example, Fung and Carter (2007) found that *I think* was one of the multi-word expressions that occurred frequently as a discourse marker in the corpus of Cantonese English speakers. Fung and Carter also reported that its high frequency was due to the pragmatic fossilisation that had occurred in the speakers' interlanguage. That is, the "inappropriate" systematic use of certain forms at the pragmatic level" (Romero Trillo, 2002, p. 770). This indicates that *I think* was used in both corpora with a high frequency but with different orthographic forms for the verb part of the multi-word expression. Regarding *I think*, Fung and Carter (2007) also point to the highly "automatic" and "routinised" use of *I think* which they believe is also the result of pragmatic fossilisation (p.431).

However, the functions of *fekr mikonam* in the LoPSC showed that there were no instances of pragmatic fossilisation. In addition, the equivalent for *I think* in Persian appeared as a high-frequency 2-gram in both corpora. The reasons for this high frequency will be further explained in section 6.5.

Other similar findings of this study and Fung and Carter's (2007) study point out the high frequency of the use of the discourse markers "*but*, *because*, and *I think* as a kind of pragmatic fossilization" (p. 429). Therefore, there is a similarity between Fung and Carter's study and this study in that there is a significantly higher frequency of *but.* In their study, they found a high frequency of occurrences of the word *but* is "consecutively used across turns, primarily at a turn-initial position to present contrastive viewpoints and counterarguments".

6.5. Factors influencing the difference

In previous research on learners of Persian, the reason for the differences between learners and L1 speakers spoken colloquial Persian has been mainly attributed to having non-native speakers of Persian as teachers and lack of attention to the colloquial form in teaching material (Saffar Moghaddam, 2011). With regard to this study, in the former case, the participants had L1 speakers of Persian as their language instructors. In the latter case, as shown by previous research, the difference in pronunciation is regarded as the main difference between learners and L1 speakers of Persian (Ghaffari, 2020; Miller, 2011), and this has led to its inclusion in textbooks for teaching colloquial Persian. However, as this study has shown, this was not the main difference. The reason behind this discrepancy could be the result of two factors. First, since the difference in pronunciation has been overemphasised, this has led to the inclusion of explicit mention of the differences between colloquial and the literary/formal written form in teaching colloquial Persian textbooks (Sedighi, 2018.). Second, since the pronunciation is at such a surface level, picking up this would be easier for learners in contrast to pragmatic features which have been reported to benefit from more explicit instruction as opposed to implicit instruction (Bardovi-Harlig and Vellenga, 2012).

In the following, factors that may have led to the differences between learner and L1 language are expanded on.

6.5.1. L1 Transfer

Transfer from the learners' L1 has been considered as the most important factor in the cause behind discrepancies between learners' and L1 speakers' language production, especially in studies using learner corpora (Granger, 2021). For example,

in the case of English VCMs, specifically, studies have reported that L1 transfer has led to the production of VCM forms that do not exist in English, such as the production of the form *or so* by German Speakers of English (Terraschke, 2010) and the use of *and and and* by Persian speakers of English (Parvaresh, Tavangar, Rasekh, & Izadi, 2012).

Others have suggested that differences in Vague Language use, for example, may be within the use of different styles for different languages. For example, Ruzaite (2018) states that the higher frequency of use VCMs and vague quantifiers (such as, *a bit, a lot*, etc,) by Lithuanian speakers of English in argumentative essays may be due to different styles that Lithuanians have in writing argumentative essays when compared to L1 speakers of English.

In the case of this study, and specifically regarding learners' VCM use, although the appearance of some forms may reflect L1 transfer, the overall significant low frequency of VCMs points to reasons other than the L1 of the speakers. This hypothesis is further asserted by the highly frequent use of VCMs in the learners' first language, namely English.

Drave (2001) believes that discrepancies between the use of Vague Language forms and functions by Hong Kongese and L1 speakers of English are not due to the influence of the L2 speakers' language background, since Vague Language in Cantonese has shown similar properties to English Vague Language. He believes these differences are due to the lack of exposure of L2 speakers, especially their lack of exposure to vague forms in language textbooks. Similarly, Overstreet et al. (2006) highlight the importance of VCMs in everyday communication in German; however, they also acknowledge the lack of attention paid to German VCMs in language resources for L2 learners of German.

Therefore, with the acknowledgement that the learners' L1 may influence certain aspects of the differences found in this study, further possible factors are considered in the following.

6.5.2. Language Input

Fung and Carter (2007) attribute the lack of diversity in discourse markers used by the learners to the "un-naturalistic input" and the "grammar-centred pedagogic focus" (p. 433) that ignores the pragmatic meaning of words in favour of their "literal" meanings. This is reiterated in other second language studies that have found significant differences between learner and L1 language production, especially regarding the use of discourse markers (Romero Trillo, 2002; Müller, 2005).

In the case of Persian textbooks, Foutouhi, Khatami and Mirdehghan (2019, p.97) found that the majority of the texts books for teaching Persian to speakers of other languages consisted of literary texts instead of samples of authentic colloquial spoken language. Foutouhi et al. (2019) present a comprehensive list of the differences between literary texts and the colloquial spoken form. Some of these differences are the use of human pronouns to refer to objects, the use of singular pronouns to refer to plural noun references, and the use of adjectives such as *hame* and *har* interchangeably in literary texts. These differences were found in the LoPSC.

Regarding the frequency and word choice for second language teaching textbooks, Miton (2009), Jimmins and Mansebo (2008), and Keraydo and Sanchez (2012) have also shown that material prepared for the purpose of teaching vocabulary does not reflect the frequency of occurrence of words. In the teaching Persian context, Jahangiri (2016) looked at the lexicon of three textbooks that he believed to be representative of the textbooks of teaching Persian, namely, Zolfaghari et al. (2008), Saffar-Moghadam (2007), and Rasuli (2012). Jahangiri (2016) compared these

textbooks to a corpus of spoken and written Persian, namely the Persian Language Data Base (PLDB) (Assi, 1997). Based on a comparison between the frequency list of the most frequent 5000 words of the three textbooks and the latter corpus, Jahangiri found that the textbooks covered less than 50% of the most frequent words in spoken and written Persian. Jahangiri states that the vocabulary chosen for teaching learners of Persian is chosen based on experience from previous years. However, this *experience* is not triangulated by any other means.

In the case of the learners of this study, the overemphasis on phonological differences between colloquial and the literary form was also evidenced in the learners' course book, "Persian: A comprehensive grammar" (Yousef and Tayebi, 2018). For example, when explaining the changes made for the 3rd person singular pronoun in colloquial Persian, the following is presented:

"There are changes in 3rd person:

ا و (ou) (he / she) and آ ن (an) (it / that) both change to او ن (oun, he / she / it / that)

آنها (anha) (they / those) changes to اونا (ouna)

ایشان (ishan) (she/he /they in polite language) changes to ایشون (ishoun)" (Yousef and Tayebi, 2018, p. 328).

However, whereas the pronunciation of *an* has changed due to the change of the pronunciation of the sound *a* to *ou*, the change in *ou* is not due to change in pronunciation. As Ghomeishi (2018) explains, there is a different word choice for the 3rd person singular pronoun with *ou* changing to *oun*. Therefore, this is a change in word choice and not a reflection on phonological changes in colloquial Persian as described in the learners' course book.

To overcome this deficiency in textbooks of Persian, Vakilifard (2009) believes that the best way to approach the design of textbooks is to research the spoken form of

Persian. However, as current textbooks on teaching colloquial Persian have shown, there is an overemphasis on pronunciation and little or no attention to other aspects of the language, especially pragmatics. This also holds true in other second language teaching contexts. As O'Keeffe (2020, p.176) states, we should be aware of features of naturally occurring speech such as discourse markers, and they should be added to classroom vocabulary lists due to both their frequency of occurrence and their importance to successful interaction.

In addition to including research on the spoken colloquial form by L1 speakers, this study also shows the importance of looking into learner language, since it emphasizes the significance of word choice and the discrepancy in the use of discourse markers as the issues that need to be addressed instead of differences in pronunciation and other surface features of the language.

A final observation from the participants in this study regarding language input relates to the types of speech activities the learners were expected to conduct during their Persian classes, which were referred to as *eraeh* (performance). In these *performance speeches*, the participants would be asked to read about a topic and present it to their peers in the form of lectures. These performance-type speeches correspond to the performance function of speech, first introduced by Richards (2008). In addition to the functional and interactional functions of speech presented by Yule and Brown (1983), Richards (2008) presents another category, namely, the performance function. Although this function is present in speech, it is more closely associated with written rather than spoken language. That is, it is pre-prepared and does not include the characteristics of spoken speech such as pauses, false starts, and fillers (Nushi, 2020), and in turn, does not have the interactivity and real-time language processing features attributed to the conversation register. In addition, speech that is associated

with performance-type spoken language is impersonal in that the speaker is not in direct contact with the speakers in the way that the listeners are left to form their meanings. As Nushi (2020) states one thing that can be used in the context of teaching Persian is to focus more on conversational and interactional functions of spoken language instead of the performance function which closely mirrors the written language and draws attention away from the spoken language. As Nunan (2015) states since transactional and interactional functions permeate our everyday conversational exchanges, they should be included in the context of the language classroom. Therefore, in addition to the language input received from their textbooks, there is a strong possibility that the types of speech activities expected from learners in language classrooms affects their production of the conversational register.

6.5.3. Defining the scope of Conversational Persian

In addition to probable L1 transfer and the focus on a limited set of phonological differences between colloquial and literary Persian in textbooks, the scope of Conversational Persian has also not been well-defined in research on Conversational Persian nor in the textbooks used for teaching it.

More specifically in the case of textbooks used by the learners of this study, Yousef and Tourabi (2018), for example, present the following categorisation for the Persian affirmative marker as illustrated in Table 6.2.

Table 6.2.

Representations of the Persian affirmative marker in different genres and registers (adapted from Yousef and Tourabi, 2018)

Genres and registers	Yes
literary	ari
casual colloquial	areh
respectful colloquial written form	baleh
written form	bali

As can be seen from Table 6.2, Yousef and Tourabi (2018) claim that *baleh* is used for "respectful colloquial" Persian; thus, implying that the use of *areh* would be disrespectful in a way. In addition, as the results of this study have shown, *baleh* was used in casual/ informal situations but with different functions compared to its function as an affirmative marker.

However, Yousef and Tourabi (2018) writers of the book state the following when introducing Colloquial Persian:

"Colloquial Persian and polite Persian should not be seen as opposites. What is meant here by *colloquial* Persian is *spoken* Persian, which can have its degrees of formality and politeness – or lack thereof." (p. 327)

Nonetheless, they present a chapter in their book titled "Colloquial Persian and Polite Persian" which may mislead judgement about the nature of Colloquial Persian. This misconception is also reflected in the learners' conceptions of what constitutes the scope of colloquial spoken Persian.

This misjudgement about the nature of informal colloquial Persian also presented itself in the interviews with the learners of this study, when in answer to the question “In what situations and with whom would they use colloquial Persian?”, the answer to this question from 16 out of the 18 participants was that they would not use features of Conversational Language, because they did not consider these features to be “polite”.

This misconception about the nature of Conversational Persian by English L1 learners of Persian could also be the result of Persian being a near “diglossic” language (Jeremias, 1984). That is in Persian, there are two varieties of the language (often referred to as the high and low variety) that are used by the speakers in different contexts (Mahmood-Bakhtiari, 2018). The contexts in which these varieties occur differ according to different references. This difference is usually the result of changes in time. For example, Jeremias (1984) was the first to introduce the Persian language as a diglossic language. In their categorisation of the differences between the high and low variety of the language, they state the following:

“These differences correspond to the traditional stylistic differences of the language varieties used in formal situations (official occasions, radio, newspapers, etc. and in informal ones (e.g. everyday conversations).”

However, in the Persian spoken in Iran today, the “low variety” of the language that corresponds to colloquial Persian, is also used in newspapers, TV and radio programs (Shabani,2018; Alami, 2012). However, these changes have not been reflected in the context of teaching speaking. For example, Nushi (2020) presents the following for the description of Persian as a diglossic language:

> “In diglossic contexts, two distinct varieties of a language are spoken within the same speech community; one is the variety that is prestigious and used for

formal and literacy purposes and is called the high variety (H-variety). This contrasts with the low variety (L-variety), which is used for informal, mostly spoken purposes" (p. 258).

Nushi also goes on to state that the differences in the high and low variety of the language manifest themselves mainly at the level of phonology instead of morphology, syntax and semantics. This resonates with the differences attributed to colloquial Persian presented in the first part of this section. Therefore, this all calls for a clearer definition of what constitutes Conversational Persian.

6.5.4. Cross-cultural pragmatic differences

Wouk (2001) attributed the use of affirmative markers as tag questions, such as the case of *areh* in the RC, to the speakers' attempt at involving the listener in the conversation in order to create a sense of solidarity. Wouk also believed that this reflected the Indonesian culture where creating solidarity is more prevalent compared to Western cultures. The sense of creating solidarity by the speakers of the RC was also reflected in the use of discourse markers such as *bebin* (look) and the high-frequency VCMs, which were absent or appeared with significantly lower frequency in the LoPSC.

Similarly, in a study comparing the use of "I think" and its collocations by Persian learners, Chinese learners of English, and American L1 speakers of English, Sabet and Zhang (2015) reported that the Persian speakers used the clusters *I think we* and *you know I think*, with a significantly higher frequency compared to their Chinese and American counterparts. Sabet and Zhang believed that this helped the Persian learners create a sense of intimacy and cooperation (p. 191). In contrast, these patterns were not found in the RC data of this study, as shown in Table 6.3. However, the cluster *ma, fekr mikonam* (we, I think) appeared in the LoPSC. This cluster was

not reported in Sabet and Zhang's study nor did it occur in the RC and is therefore unique to the learner data of this study.

Sabet and Zhang (2015) make another observation which is relevant to this study; that is, they believe that based on their data, Persian speakers were less authoritative compared to Chinese and American speakers of English. This less authoritative attitude of the Persian speakers was evidenced by the use of *but I think* to "softly express disagreement and indicate contrast" in their data (p.191). However, the results of this study differ from that of Sabet and Zhang in that the learners of this study used the *but I think* cluster (*vali/ama fekr mikonam/konam*) with a significantly higher frequency compared to the speakers of the RC, as shown in Table 6.3. The difference between the results of this study and that of Sabet and Zhang's study could be the result of two factors. First, Sabet and Zhang's data was taken from a corpus compiled from classroom data. This could indicate that the Persian learners opt for a less authoritative attitude in the classroom setting compared to colloquial Persian. Another reason could be the result of the use of other forms, namely, the discourse marker *akheh* in the RC instead of the Persian equivalent of *but I think*. In either case, the learners of this study showed a less authoritative attitude compared to the speakers of the RC.

Table 6.3

Clusters for fekr mikonam

LoPSC (raw/relative frequency)	RC (raw/relative frequency)	Forms
16	30	vali fekr konam/ mikonam (but I think)
27	0	ama fekr mikonam (but I think)
13	0	ma, fekr mikonam (we, I think)
0	0	midooni fekr konam/fekr mikonam (you know I think)

This less authoritative attitude of the learners in itself may be the result of two factors. First, as Fung and Carter (2011) and Sabet and Zhang (2015) themselves acknowledge, learners adhere to forms such as *I think* since they have a limited amount of resources compared to the L1 speakers. This was demonstrated by the lack of use of discourse markers such as *akheh* in the LoPSC. Therefore, in conclusion, there can not be any firm claims on the influence of cultural differences for the two groups of speakers of this study.

6.5.5. Learners stay abroad in Iran

Students' stay abroad has been shown as an influential variable in the frequency and variety of the use of different forms, especially regarding the use of discourse markers. (See Section 2.6).

In this study, learners were asked whether they had stayed in Iran and if so the length of their stay abroad (See Table 4.1). All of the participants had experience with living in Iran except for two students. From these two students, one of the students had a three-month stay in Tajikistan and the other student had no experience at all in staying in a country where Persian was considered the official language.

Nonetheless, the results of this study did not correspond to previous research in the case of the frequency of discourse markers in the sense that the learners with no experience of staying in a country where Persian was the official language, was the most frequent user of VCMs. Although the number of participants in this study is very limited and therefore generalisations cannot be drawn upon the results of this study regarding the relationship between the frequency of VCMs and the length of stay abroad, this study does indicate in this specific context, the length of stay abroad had no effect on the use of colloquial Persian forms, especially the use of VCMs.

6.6. Chapter Summary

In this discussion chapter, I started by providing a summary of the findings from the previous chapter by categorising the differences found between the two groups of speakers. The two categories of differences found were differences in the phonemic and word choices of the learners and L1 speakers of Persian. These differences were especially pronounced regarding the use of Vague Category Markers between the two groups of speakers. I then proceeded to consider the findings of this study in the light of previous studies. I concluded this chapter by exploring the possible factors that may have influenced the differences between the use of Conversational Persian between the learners and L1 speakers of this study. These factors not only pointed to the possible transfer from the learners' L1, but also highlighted other possible factors such as the influence of language input and the learners' perceived attitude towards using the Conversational register of Persian.

7.Conclusion Chapter

7.1. Introduction

In this concluding chapter of the thesis, I first present the answers to this study's research questions. I then present the contributions of this study. The contributions of the study are followed by illustrating a set of implications. As with any other study, this study had a set of limitations, which I also precede to highlight. I conclude this chapter by suggesting directions for future research.

7.2. Summary of Findings and Answers to the Research Questions

In this section, I set out to present a summary of the findings of this study. I outline these findings in the form of answers to the study's research questions, which were, in Conversational Persian:

1. What are the significant differences between the forms used by learners and L1 speakers of Persian?
2. What are the differences in the pragmatic functions of the most significant forms used by learners and L1 speakers?
3. What are the differences between the use of discourse markers by learners and L1 speakers of Persian?
4. What are the differences between the use of Vague Category Markers by learners and L1 speakers of Persian?

To answer these research questions, I used a mixed methods approach. This approach consisted of using corpus analytical tools, employing Contrastive Language Analysis and conducting semi-structured interviews. Before conducting the analysis of this study, first, I compiled LoPSC. The LoPSC is a 40,000-word spoken learner corpus consisting of conversations from 18 English-speaking learners of Persian. To

gain a better perspective of their use of Conversational Persian, I also interviewed the learners after recording each session.

For the analysis of this study, first, I drew on results from a corpus-based CLA analysis. I compared the LoPSC with a 60,000-word spoken corpus of conversational Persian from 30 speakers of Tehrani Persian. This analysis consisted of generating frequency wordlists from the LoPSC (See Appendix VII) and the RC (See Appendix VIII). In addition, positive and negative keyword lists were generated from the two corpora of this study (See Table 5.1 and Table 5.7). This initial analysis of the data provided answers to the first research question of this study. That is, based on the results from these analyses, I divided the differences between the forms used by learners and L1 speakers into two main categories, namely, differences in the use of phonemes and word choices. Each category of differences and their relation to the existing literature on Conversational Persian is summarised below.

Phonemic Differences

The main phonemic difference between the two corpora was the significantly lower use of the vowel /u:/. In the phonological system of Persian, the substitution of /u:/ for the vowel /a:/ has been the most significant change in the last two decades (Miller, 2015). As Miller (2011) states, although considered a feature of Conversational Persian, this substitution is now widely used in other registers of Persian. This substitution illustrates and highlights two factors. First, it highlights the permanence of features of Conversational Language in other registers. Second, it shows the sociolinguistic factors influencing the changes in the Persian Language used in Iran.

In addition, data from the interviews conducted with the learners of this study showed that they identified this substitution as the main noticeable difference between Conversational Persian and other registers. However, despite this awareness, as

illustrated by the results of this study, learners of Persian chose this frequently-used feature of Conversational Persian significantly less compared to L1 speakers.

Finally, when compared with findings from previous studies on learners of Persian, the lower frequency of substituting /u:/ for /a:/ had not been reported (See section 3.2.3). In addition, the findings from this study diverged from findings from other studies in that the learners did not show significant difficulties in pronouncing phonemes that their L1 (English) lacks, such as /kh/ and /gh/ (See section 3.2.3).

Differences in Choices of Words

In this study, the choice of words and phrases were the main difference between the forms chosen by learners and Iranian speakers of Persian. The most significant differences were the higher frequency of the words *amâ* (but), *ou* (him/her/it), *baleh* (yes), *barâyeh* (for), *kami* (a bit), and *fekr mikonam* (I think) in the LoPSC (See Table 5.1). All these forms corresponded to forms less associated with Conversational Persian and more commonly used in other registers of the language, especially, the Standard written form. This was observable by a comparison of the positive keyword list with the negative keyword list of LoPSC. That is, each positive keyword had a corresponding form in the RC as displayed in Table 6.1. This table shows that each form from the LoPSC has a corresponding form in the RC. Previous studies on Conversational Persian have recognised only one of these preferences, namely, the preference for using *âreh* (yeah) in Conversational Persian (Alami, 2016). Therefore, this study provides a more comprehensive list. With the list of differences between the forms used by learners and L1 speakers, further analysis of each form using corpus tools, namely, Collocations and Concordance Lines were used to answer the second research question of this study regarding the pragmatic differences between the two groups of speakers.

Pragmatic differences

From further analysis of the forms of the positive and negative keyword lists, three of the forms showed pragmatic differences, namely, *amâ* (but), *baleh* (yes) and *fekr mikonam* (I think). As shown in Table 6.1, the positive keyword *amâ* (but) was predominately replaced by the conversational form *vali* (but) in the RC. A comparison between the functions of *amâ* (but) and *vali* (but) showed the two forms performed similar pragmatic functions across the two corpora. However, *amâ* (but) was used in the LoPSC more frequently as a contrastive and mitigating device compared to *vali* in the RC. The negative keyword list showed this to be the result of the preference for the speakers of the RC to use the discourse marker *âkheh* as a contrastive and mitigating device, instead of the conjunction vali or its less conversational form, *amâ* (but). In her study on the most frequent discourse markers in Conversational Persian, Mohammadi (2018) found that *âkheh* was used as a discourse marker in Conversational Persian to extenuate expressions of disagreement. Mohammadi also stated that *âkheh* does not have an equivalent in English. However, in this study, âkheh was not used by the learners. Alternatively, the learners used the less conversational form, *amâ* (but), to express contrast.

Similar to *amâ* (but), *vali* and *âkheh*, the two forms *baleh* (yes) and *âreh* (yeah) were also used in the two corpora. *baleh* (yes) was used significantly more frequently in the LoPSC compared to the RC. However, there were two points of divergence between *amâ* (but), and its substitute forms *baleh* (yes) and *âreh* (yeah). First, whereas *vali* and *âkheh* did not occur in the LoPSC, the two forms *baleh* (yes) and *âreh* (yeah) occurred with similar frequencies in the LoPSC. Further analysis of *baleh* (yes) and *âreh* (yeah) in the LoPSC also showed that both forms are used by learners to express affirmation. Therefore, *baleh* (yes) and *âreh* (yeah) are used interchangeably in the LoPSC.

In addition, in contrast to *amâ* (but) that did not occur in the RC, *baleh* (yes) did occur in the RC, albeit with a significantly lower frequency than *âreh* (yeah). However, *baleh* (yes) and *âreh* (yeah) were not used interchangeably in the RC, and an analysis of the two forms showed that although *âreh* (yeah) was used as an affirmation marker in the RC, *baleh* (yes) had lost its function as an affirmation marker and was used for other functions such as displaying annoyance or sarcasm. Therefore, in contrast to the learners, the L1 speakers used *baleh* (yes) and *âreh* (yeah) for different purposes.

Finally, the analysis showed a significant preference for the phrase *fekr mikonam* (I think) in the LoPSC whereas the speakers of the RC preferred to use the shortened form *fekr konam*. More specifically, the speakers of the LoPSC used the progressive form of the verb kardan (to do) to form the expression *I think,* (i.e., **mi**konam), and the speakers of the RC used the contracted form *konam*. However, the concordance lines containing the forms of *fekr mikonam* (I think) and *fekr konam* (I think) in both corpora showed different functions. That is, in the LoPSC, *fekr mikonam* functioned as the equivalent to *I think* in English whereas, in the RC, the small number of cases of *fekr mikonam* functioned as the (near) equivalent to *I imagine*. Therefore, *fekr mikonam* had not only lost its original form, but its pragmatic function had also changed in the Conversational Persian used by the L1 speakers.

Regarding the third and fourth questions of this study, the results from the Negative Keyword List (See Table 5.7), first, showed the prevalence of discourse markers and VCMs in the Conversational Persian used by L1 speakers. Second, they also illustrated significant differences between how learners and L1 speakers use both these forms. This is summarised in the following two sections.

Discourse Marker Differences

The discourse markers chosen for the purpose of analysis were based on the positive and negative keywords analysis. Two of these discourse markers, namely, *fekr mikonam* and *amâ* (but), have been briefly discussed in the previous section related. However, one final point concerned with *fekr mikonam* relates to the frequency of the form in both corpora. Although *fekr mikonam* has not been previously studied in the Conversational Persian register, its equivalent *I think*, has been widely studied in English and with learners of English (See section 6.4.2). These studies have pointed to a higher frequency of *I think* in learner language, with some studies reporting on the use of *I think* as performing different functions for learners compared to L1 speakers. However, this was not the case in this study. That is, the learners of the LoPSC used the Persian equivalent of *I think* with similar frequencies and functions. Therefore, *fekr mikonam* (I think) is a high-frequency occurring expression for both leaners and L1 speakers of Persian. This indicates the preference for L1 speakers of Persian to indicate deference to their interlocutors (Sabet, 2013; Sabet and Zhang, 2015).

Similar to *fekr mikonam*, Persian speakers showed signs of deference by using the "extenuating" contrastive discourse marker *âkheh*. In contrast, there were no occurrences of this form in the LoPSC, and the learners relied on using the formal contrastive marker *âma*.

In terms of the overall differences between the discourse markers used by learners and L1 speakers of Persian, the findings of this study echoed two themes from previous studies on the comparison between the use of discourse markers by learners and L1 speakers. First, the learners used fewer discourse marker forms. That is, whereas the speakers of the RC used discourse markers in a variety of forms, the learners only used a limited number of discourse markers to perform various functions.

One of the clear examples of a single discourse marker performing a variety of functions in the LoPSC was the case of *amâ* (but), as explained in the previous section.

Another finding, which resonated with the findings of previous studies on learners' use of discourse markers, was the significantly lower frequency of interactional discourse markers. This was reflected in the top ten negative keywords, which included discourse markers, such as *bebin* (look), *hey* (again) and *boro* (go) (See Table 5.7).

A discussion of the factors leading to these differences was addressed in section 6.5. To reiterate, the lack of use of discourse markers and interactional discourse markers, specifically, reflects several factors. One of these factors includes the learners' lack of awareness of the importance of these interactional discourse markers in Conversational Persian. This lack of awareness in itself is a reflection of the input the learners receive through language teaching material such as language learning textbooks and course books. More specifically, whereas discourse markers in all registers of Persian are the highest occurring forms (Alami, 2016), this is not reflected in the frequency of words included in Persian teaching course books. In addition, discourse markers occur with higher frequency regarding both occurrence and variety of forms in Conversational Persian (Mohammadi, 2018); therefore, frequency lists for textbooks or course books aiming to teach Conversational Persian should also reflect the high frequency of these forms.

Vague Category Marker Differences

Similar to the use of discourse markers, the use of VCMs showed significant differences between the two groups of speakers. These differences involved both the frequency of occurrences and the variation of the structural forms of the VCMs. In the former case, VCMs occurred significantly higher in the RC when compared to the LoPSC. However, in the case of frequency of forms, the learners used various

structures to create VCMs whereas the speakers of the RC predominately used one form of the Persian VCMs, namely, *va inâ* (and these). Instances within the LoPSC also showed the use of adjunctive VCMs instead of disjunctive VCMs and vice versa by the learners.

The findings regarding the use of VCMs by learners were significant in that, not only did this finding show the features of VCM use by learners of Persian, which have not been previously studied before, but it also showed two important factors influencing the choice of forms in the Conversational Persian used by learners. The first factor was the influence of the learners' L1. That is, instead of using forms of Persian VCMs, the learners used word-to-word equivalents of VCMs in English, such as *and everything (va chizha)*. In addition, it further showed the lack of exposure to VCMs by learners; thus, leading to a significantly low frequency of use and the resort to creating VCM forms based on their L1 (i.e., English.

As shown in the use of discourse markers and VCMs, the lack of exposure to features of the conversational register played a significant role in the choice of forms by learners. However, as data from the interviews with the learners showed, the learners of this study expressed knowledge of certain forms related to Conversational Persian but refrained from using them in their conversations. For example, the learners expressed awareness of the substitution of the phoneme /a:/ for /u:/ in Conversational Persian and stated this as a main feature of Conversational Persian. However, as the data shows, the learners used this feature less frequently, especially when compared to L1 speakers. When asked about their perceptions towards using features of Conversational Persian and Conversational Persian in general, the learners stated that they would avoid using certain features such as the substitution of /a:/ since it was

not polite or that they considered this feature to be specific of certain dialects of Persian, such as the Isfahani dialect.

7.3. Contributions

This study has empirical and methodological contributions, which I outline in the following section.

Empirical Contributions

This study provides empirical evidence on the use of a lesser-studied variety of language, namely, the Conversational Persian used by learners. This empirical contribution of the study was mainly highlighted by comparing the language use of the learners with L1 speakers of Persian. This comparison showed word choice to be the main difference between the Conversational Persian used by learners and L1 speakers. The differences in word choice also implied different pragmatic functions for the same linguistic forms. This study also provided further empirical evidence on the use of two salient features of Conversational language by learners of Persian, namely, discourse markers and VCMs.

Although the focus of the study was on learners of Persian, this study also provided further insight into the Conversational Persian used by L1 speakers, such as the pragmaticalisation of *baleh* and *fekr mikonam*.

Methodological Contributions

In this study, I used a novel methodological approach, namely a combination of Corpus Linguistics and Contrastive Language Analysis to explore the use of Conversational Persian by learners of Persian. To use this approach, a new spoken learner corpus, LoPSC, was compiled. As the first spoken learner corpus for Conversational Persian,

LoPSC contributes to the field by providing insight into the compilation of such a corpus. This study also provided insight into the methodological nuances of comparing two corpora of Conversational Persian.

7.4. Implications of Findings

Bardovi-Harling (2012, p. 6) states that studies that compare the pragmatic functions used by learners and L1 speakers' speech

> "serve a number of functions: as primary research into L2 pragmatics; as a type of pragmatic error analysis; as a needs assessment for the development of pedagogical methods and materials for teaching pragmatics; and as models for pedagogical materials. They may also serve to define research areas for acquisitional ILP studies."

The implications of the findings from this study are in line with the above statement in that this study highlights the importance of using discourse markers and VCMs in Conversational Language and the necessity of their inclusion in language teaching, especially in textbooks on Conversational Persian. The empirical findings provided in this study can serve as a benchmark for showing the forms that need to be addressed more when teaching Conversational Persian to language learners.

The learner corpus compiled in this study can also be used for Data Driven Learning (DDL) (Johns, 1986). That is using, corpora and corpus tools in the context of the language classroom. The success of using DDL has been shown in other languages and especially in teaching English to language learners (Viana, 2022). Therefore, a corpus analysis of LoPSC in itself or the comparison of LoPSC with the reference corpus used in this study can also be integrated into the context of the classroom for language learners of Persian. The language learners of this study did express interest

in looking into LoPSC to check if they had any “errors”. Therefore, introducing DDL in the context of teaching Conversational Persian can provide learners with a potential new method of learning Persian.

7.5. Limitations of the Study

In this study, I used a hearer-based approach for the interpretation of the pragmatic functions, since as explained in section 2.5 there are obstacles to accessing the speakers’ intentions. Therefore, to specify the functions, I drew upon my intuitions as an L1 speaker of Persian. I suggest some potential steps that can be taken to further validate these intuitions. One step is using post-data collection interviews with the speakers to tap into their intentions. However, there are several practical and to an extent ethical concerns with this type of approach. For example, the speakers may not remember their intention at the time. Another obstacle may be the speaker’s unwillingness or discomfort in expressing their true intentions to the interview. Nonetheless, such measures will help in providing further insight into the intended pragmatic meanings of forms.

Another limitation of this study is related to the nature of corpus linguistics’ studies in general. This limitation concerns the non-occurrence of forms in the learner corpus. That is, the non-occurrence of forms does not necessarily reflect the learners’ unawareness or difficulty in comprehending these forms. Therefore, to provide a better illustration of the use of certain discourse markers or VCMs, other methods such as discourse completion tasks or other methods of explicit elicitation of these forms can be used in future studies. This study has provided a starting point for such further studies.

7.6. Directions and Suggestions for Future Research

Some suggestions for further research were mentioned in the previous section when discussing the limitations of this study, such as using other methods for triangulation, such as conducting studies with DCTs or using interviews with speakers to gain further insight into their intended meanings.

In addition to these suggestions, as a freely available corpus for research purposes, LoPSC provides a data set for further research on the use of Conversational Persian by learners. In addition, the size of LoPSC and the inclusion of learners from other first language backgrounds can be further expanded on to allow for more generalisation of the data especially in the case of determining the effects of transfer from the learners' first language.

REFERENCES

Aas, H. L. (2011). *Recurrent Word-Combinations in Spoken Learner English: A study of corpus data from Swedish and Norwegian advanced learners* [Doctoral dissertation, University of Oslo]. Retrieved from http://oatd.org/oatd/record?record=oai:www.duo.uio.no:10852%2F25311&q=learner%20corpora

Ädel, A., Granger, S., Gilquin, G., & Meunier, F. (2015). *Variability in learner corpora*. In A. Ädel, S. Granger, G. Gilquin, & F. Meunier (Eds.), *Learner corpora in language testing and assessment* (pp. 401-422). Cambridge University Press. https://doi.org/10.1017/cbo9781139649414.018

Adolphs, S., & Carter, R. (2003). *And she's like, it's terrible, like: Spoken Discourse, Grammar and Corpus Analysis*. *International Journal of English Studies*, 3(1), 25-58.

Adolphs, S., & Carter, R. (2007). *Beyond the word: Challenges in designing and analyzing a spoken corpus*. *European Journal of English Studies*, 11(2), 133-146. https://doi.org/10.1080/13825570701452698

Adolphs, S., & Knight, D. (2010). *Building a spoken corpus*. In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. Routledge.

Afghari, A. (2007). *A sociopragmatic study of apology speech act realization patterns in Persian*. *Speech Communication*, 49(3), 177-185. https://doi.org/10.1016/j.specom.2007.01.003

Afghari, A., & Karimnia, A. (2007). *A contrastive study of four cultural differences in everyday conversation between English and Persian*. *Intercultural Communication Studies*, 16(1), 243.

Ahmadian, A. (2004). *Barrasi-ye Xatāhā-ye Zabāni dar Neveʃtār-e Dāneʃāmuzān-e Fārsiāmuz-e Kord Zabān-e Guyeʃ-e Mahābād dar Sath-e Motevaset-e Zabānāmuzi [The Study of Linguistic Errors of the Written Texts of Kurdish (Mahabadi Dialect)-speaking Learners of Persian at Intermediate Level]* (Master's thesis). Allameh Tabatabaei University.

Aijmer, K. (2002). *English Discourse Particles: Evidence from a corpus*. John Benjamins Publishing Company. Retrieved from http://ebookcentral.proquest.com/lib/ed/detail.action?docID=622516

Aijmer, K. (2004). *Pragmatic Markers in Spoken Interlanguage*. *Nordic Journal of English Studies*, 3, 173-190.

Aijmer, K. (2013). *Understanding pragmatic markers: A variational pragmatic approach*. Edinburgh University Press.

Aijmer, K. (2015). *General extenders in learner language*. In (pp. 211-233). https://doi.org/10.1075/scl.73.10aij

Aijmer, K., & Altenberg, B. (Eds.). (2002). *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*.

Aijmer, K., & Elgemark, A. (2013). *The pragmatic markers look and listen in a cross-linguistic perspective*. In D. Allerton, J. Dendale, & T. Mortelmans (Eds.), *Of butterflies and birds, of dialects and genres: Essays in honour of Philip Shaw* (pp. 333-348).

Aijmer, K., & Rühlemann, C. (Eds.). (2015). *Corpus Pragmatics: A Handbook*. Cambridge University Press.

Aijmer, K., & Simon-Vandenbergen, A. M. (2006). *Pragmatic markers in contrast* (1st ed.). Elsevier.

Alami, M. (2016). *An investigation of pragmatic functions and position of prevalent Persian discourse markers used in casual conversations among Tehrani speakers*. *International Journal of Applied Linguistics and English Literature*, 5(1), 250-263.

Alayiaboozar, E. (2019). *A Corpus-based Study of Persian Noun and Adjective Homographs to help Correct POS Tagging*. *Iranian Journal of Information Processing Management*, 34, 897-922.

Allami, H., & Naeimi, A. (2011). *A cross-linguistic study of refusals: An analysis of pragmatic competence development in Iranian EFL learners*. *Journal of Pragmatics*, 43(1), 385-406.

Altenberg, B. (1998). *Connectors and sentence openings in English and*. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *Corpora and cross-linguistic research: Theory, method and case studies* (pp. 115). Rodopi.

Allwood, J., Nivre, J., & Ahlsén, E. (1990). Speech Management—on the Non-written Life of Speech. *Nordic Journal of Linguistics, 13*(1), 3-48. doi:10.1017/S0332586500002092

Amiridze, N., Davis, B. H., & Maclagan, M. (2010). *Fillers, Pauses and Placeholders*. John Benjamins Publishing Company.

Assi, S. M. (1997). *Farsi linguistic database (FLDB)*. *International Journal of Lexicography*, 10(3), 5.

Aston, G., Bernardini, S., & Stewart, D. (Eds.). (2004). *Corpora for Learners*. John Benjamins Publishing Company.

Azarbad, E., & Ghahraman, V. (2018). *A comparative study on the English to Persian translation of hedges in the abstracts of MA Theses in English translation studies*. *Journal of Language and Translation*, 8(3), 57-67.

Bābāyi, E. (2014). *"Barresi-ye Moqābelehi-ye Nezām- e Āvāyi-ye Zabān- e Fārsi va Zabān- e Rusi [A Contrastive Study of the Sound Systems in Persian and Russian]"*. *Pāyān Nāmeh- ye Kārshenāsi-ye Arshad, Dāneshgāh-e Allāmeh Tabātabāi*.

Baker, P., & Egbert, J. (2016). *Triangulating Methodological Approaches in Corpus Linguistic Research* (Vol. 17). Routledge. https://doi.org/10.4324/9781315724812

Baker, P. H., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh University Press.

Ball, N. C., & Ariel, M. (1978). *Or something etc. Penn Review of Linguistics*, 3(1), 35-45.

Bardovi-Harlig, K. (2012). *Pragmatics and second language acquisition*. na.

Bardovi-Harlig, K., & Vellenga, H. E. (2012). *The effect of instruction on conventional expressions in L2 pragmatics*. *System*, 40(1), 77-89. https://doi.org/10.1016/j.system.2012.01.004

Barsalou, L. W. (1983). *Ad hoc categories*. *Memory & Cognition*, 11(3), 211-227. https://doi.org/10.3758/BF03196968

Biber, D. (1993). *Representativeness in Corpus Design*. *Literary and Linguistic Computing*, 8.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). Longman.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511814358

Bijankhan, M., Sheikhzadegan, J., Roohani, M. R., Samareh, Y., Lucas, C., & Tebiani, M. (1994). *The speech database of Farsi spoken language*. *Proceedings of Speech Science and Technology Conference*, 826-831.

Bijankhan, M., Sheykhzadegan, J., Bahrani, M., & Ghayoomi, M. (2010). *Lessons from building a Persian written corpus: Peykare*. *Language Resources and Evaluation*, 45(2), 143-164. https://doi.org/10.1007/s10579-010-9132-x

Bijankhan, M., Sheykhzadegan, J., Roohani, M. R., Zarrintare, R., Ghasemi, S. Z., & Ghasedi, M. E. (2003). *Tfarsdat-the telephone Farsi speech database*. *Eighth European Conference on Speech Communication and Technology*.

Blakemore, D. (1987). Semantic Constraints on Relevance. Oxford: Blackwell.

Bordería, S. P. (2008). *Do discourse markers exist? On the treatment of discourse markers in Relevance Theory*. *Journal of Pragmatics*, 40, 1411-1434.

Bousfield, D. (2008). *Impoliteness in interaction.* John Benjamins Publishing Company.

Brezina, V. (2018). *Statistics in Corpus Linguistics: A practical guide*. Cambridge University Press. https://doi.org/10.1017/9781316410899.006

Brezina, V., McEnery, T., & Wattam, S. (2015). *Collocations in context: A new perspective on collocation networks*. *International Journal of Corpus Linguistics*, 20(2), 139-173. https://doi.org/10.1075/ijcl.20.2.01bre

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Bruyn, B. L., & Paquot, M. (2021). *Learner corpus research meets second language acquisition*. Cambridge University Press.

Buttery, P., McCarthy, M., & Carter, R. (2015). Chatting in the academy. *Corpora, Grammar and Discourse: In honour of Susan Hunston*, *73*, 183.

Buysse, L. (2014). *'We Went to the Restroom or Something'. General Extenders and Stuff in the Speech of Dutch Learners of English*. 2, 213-237. https://doi.org/10.1007/978-3-319-06007-1_10

Buysse, L. (2017). *The pragmatic marker you know in learner Englishes*. *Journal of Pragmatics*, 121, 40-57. https://doi.org/10.1016/j.pragma.2017.09.010

Callies, M., Granger, S., Gilquin, G., & Meunier, F. (2015). *Learner corpus methodology*. 35-56. https://doi.org/10.1017/cbo9781139649414.003

Carter, R. (1998). *Orders of reality: CANCODE, communication, and culture*.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English*. Cambridge University Press.

Carter, R., & McCarthy, M. (2017). *Spoken Grammar: Where Are We and Where Are We Going? Applied Linguistics*, 38(1), 1-20. https://doi.org/10.1093/applin/amu080

Channell, J. (1994). *Vague Language*. Oxford University Press.

Cheng, W., & Martin, W. (2001). *The Use of Vague Language in Intercultural Conversations in Hong Kong*. *English World-Wide*, 22(1), 81-104.

Cheng, W., & Warren, M. (2007). *Checking Understandings: Comparing Textbooks and a Corpus of Spoken English in Hong Kong*. *Language Awareness*, 16(3), 190-207. https://doi.org/10.2167/la455.0

Cheng, W. G., Chris Warren, Martin (2006). *From n-gram to skipgram to concgram*. *International Journal of Corpus Linguistics*, 11(4).

Cheshire, J. (2007). *Discourse variation, grammaticalisation and stuff like that 1*. *Journal of Sociolinguistics*, 11(2), 155-193. https://doi.org/10.1111/j.1467-9841.2007.00317.x

Crystal, D. (2008). A Dictionary of Linguistics and Phonetics (6th ed.). Oxford: Blackwell. http://dx.doi.org/10.1002/9781444302776

Crystal, D., & Davy, D. (1975). *Advanced Conversational English*. Longman.

Culpeper, J. (1996). Towards an anatomy of impoliteness. *Journal of pragmatics*, *25*(3), 349-367.

Culpeper, J. (2011). *Impoliteness: Using Language to Cause Offence* (Studies in Interactional Sociolinguistics). Cambridge University Press.

Culpeper, J., Mackey, A., & Taguchi, N. (2018). *Second Language Pragmatics: From Theory to Research* (1st ed.). Florence: Routledge. https://doi.org/10.4324/9781315692388

Curry, N., Love, R., & Goodman, O. (2020). *Adverbs on the move: Investigating publisher application of corpus research on recent language change to ELT coursebook development*. *Corpora*.

De Cock, S. (2004). *Preferred sequences of words in NS and NNS speech*. *Belgian Journal of English Language and Literatures (BELL)*, 2(1), 225-246.

Eslami, M. (2013). *Tahlil-e Xatāhā-ye Neveʃtāri-ye Rusi Zabānān-e Fārsiāmuz dar Sath- e Miyāni [Error Analysis of Written Text of Russian-speaking Learners of*

Persian at Intermediate Level]. MA Dissertation, Ferdowsi University of Mashhad, Mashhad.

Eilam, T. (2019). *A typological sketch of the Jewish Iranian dialects*. In *Essays on Typology of Iranian Languages*, 328, 167.

Dārābi, Kh. (2001). *"Barresi-ye Moshkelāt-e Āvāyi-ye Guyeshvarān-e Kord Zabān-e Kermānshāhi dar Yādgiri- ye Zabān- e Fārsi [An Investigation of Pronunciation Problems of Kermanshahi Kurdish Speakers in Learning Persian]"*. Pāyān Nāmeh-ye Kārshenāsi-ye Arshad, Dāneshgāh-e Shahid Beheshti-ye Tehrān, Iran.

Denis, D. (2017). *The Development of And Stuff in Canadian English: A Longitudinal Study of Apparent Grammaticalization*. *Journal of English Linguistics*, 45(2), 157-185. https://doi.org/10.1177/0075424217701182

Deshors, S., & Gries, S. (2020). *Gries, Stefan Th. & Sandra C. Deshors. There's more to alternations than the main diagonal of a 2×2 confusion matrix: improvements of MuPDAR and other classificatory alternation studies*. *ICAME Journal*, 44. https://doi.org/10.1111/ijal.12275

Dines, E. (1980). *Variation in discourse - "and stuff like that"*. *Language in Society*, 9, 13-31.

Drave, N. (2001). *Vaguely speaking: a corpus approach to vague language in intercultural conversations*.

Dubois, B. L. (1987). *"Something on the order of around forty to forty-four" - imprecise numerical expressions in biomedical slide talks*. *Language and Society*, 16, 527-541.

Dubois, S. (1992). *Extension particles, etc*. *Language Variation and Change*, 4(2), 179-203. https://doi.org/10.1017/S0954394500000740

Eghbalzadeh, H., Hosseini, B., Khadivi, S., & Khodabakhsh, A. (2012). *Persica: A Persian corpus for multi-purpose text mining and Natural language processing*. *6th International Symposium on Telecommunications (IST)*.

Falahati, R. (2020). *The acquisition of segmental and suprasegmental features in second language Persian: A focus on prosodic parameters of politeness*. In P. Shabani-Jadidi (Ed.), *The Routledge Handbook of Second Language Acquisition and Pedagogy of Persian* (pp. 9-35). Routledge.

Fischer, K. (2006). Frames, constructions, and invariant meanings: the functional polysemy of discourse particles. In *Approaches to discourse particles* (pp. 427-447). Brill.

Flowerdew, L. (2004). *The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings*. In U. Connor, & T. Upton (Eds.), *Discourse in the Professions: Perspectives from Corpus Linguistic* (pp. 11-33). John Benjamins.

Flowerdew, L. (2012). *Corpora and Language Education* (1st ed.). London: Palgrave Macmillan UK.

Flowerdew, L. (2017). *Learner Corpus Research: New Perspectives and Applications*. London: Bloomsbury Publishing Plc.

Fraser, B. (1996). *Pragmatic markers*. *Pragmatics*, 6(2), 167-190.

Fung, L., & Carter, R. (2007). *Discourse markers and spoken English: Native and learner use in pedagogic settings*. *Applied Linguistics*, 28(3), 410-439.

Gablasova, D., Brezina, V., & McEnery, T. (2019). *The Trinity Lancaster Corpus*. *International Journal of Learner Corpus Research*, 5(2), 126-158. https://doi.org/10.1075/ijlcr.19001.gab

Gassner, D. (2012). *Vague Language That Is Rarely VagueP: A Case Study of "Thing" in L1 and L2 Discourse*. *International Review of Pragmatics*, 4(1), 3-28. https://doi.org/10.1163/187731012x632045

Gebhardt, L. (2018). *Trends in Iranian and Persian Linguistics*. In *Accounting for yek ta in Persian* (pp. 213). De Gruyter Mouton. https://doi.org/https://doi.org/10.1515/9783110455793-012

Ghaffari, M. (2020). *Persian as an Interlanguage*. In P. Shabani-Jadidi (Ed.), *The Routledge Handbook of Second Language Acquisition and Pedagogy of Persian*. Routledge.

Ghaleno, E. T., & Moghaddam, M. D. (2019). The Comparison of Discourse Markers in the Narrative Discourse of 7 and 10–year-old Persian-Speaking Children with Adults.

Ghayoomi, M., & Momtazi, S. (2009). Challenges in developing Persian corpora from online resources. In Proceedings of IEEE international conference on Asian language processing, Singapore.

Ghomeshi, J. (2018). Trends in Iranian and Persian Linguistics. In The associative plural and related constructions in Persian (pp. 233). De Gruyter Mouton. https://doi.org/https://doi.org/10.1515/9783110455793-013

Gilquin, G. (2005). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The Cambridge Handbook of Learner Corpus Research. Cambridge University Press.

Gilquin, G. (2021). Learner Corpora. In M. Paquot & S. Gries (Eds.), A practical handbook of corpus linguistics. Springer International Publishing.

Goffman, E. (1955). On Face-Work. Psychiatry, 18(3), 213-231. https://doi.org/10.1080/00332747.1955.11023008

Götz, S. (2013). Fluency in native and nonnative English speech (Vol. 53). John Benjamins Publishing Company Amsterdam.

Götz, S., & Mukherjee, J. (2019). Learner corpora and language teaching. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2110210

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & T. Upton (Eds.), Applied corpus linguistics: A multidimensional perspective (pp. 123–145). Rodopi.

Granger, S. (2012). How to use Foreign and Second Language Learner Corpora. 5-29. https://doi.org/10.1002/9781444347340.ch2

Granger, S. (2017). Learner Corpora in Foreign Language Education. In S. L. Thorne & S. May (Eds.), Language, Education and Technology (pp. 427-440). Springer International Publishing. https://doi.org/10.1007/978-3-319-02237-6_33

Granger, S. (2021). Have Learner Corpus Research and Second Language Acquisition finally met? In B. Le Bruyn & M. Paquot (Eds.), Learner Corpus Research Meets Second Language Acquisition (pp. 243-257). Cambridge University Press.

Green, S. (2006). The management of projects in the construction industry: context, discourse and self-identity. *Making projects critical*, *13*, 232-235.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), Studies in Syntax and Semantics III: Speech Acts (pp. 183-198). Academic Press.

Gu, Y. (1990). Politeness phenomena in modern Chinese. *Journal of pragmatics*, *14*(2), 237-257.

Gumperz, J. J. (1977). The Sociolinguistic Significance of Conversational Code-Switching. RELC Journal, 8(2).

Hansen, M. B. M. (2006). A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of French toujours). In *Approaches to discourse particles* (pp. 21-41).

Hays, P.R., (1992). Discourse markers and L2 acquisition. Papers in Applied Linguistics-Michigan, 7, 24–34.

Hosseyni, S. A., M. Bijan Khān, and R. Moqadamkiyā. (2009). Barresi-ye Moqābelei-ye Nezām-e Āhang- e Fārsi va Jāponi bā negāhi Be Tekyeh va Zir va Bami-ye Hastei Dar Do Zabān [A Contrastive Analysis of Persian and Japanese Intonation Systems]. Nashriyye-ye Pazhuhesh- e Zabān- hā- ye Khāreji, 54, 5–26.

Ide, S. (1989). Formal forms and discernment: Two neglected aspects of universals of linguistic politeness.

Jahangardi, K., Assi, M., Afrashi, A., & Vakilifard, A. (2017). Vocabulary in the Textbooks of Teaching Persian to Non-Persian Speakers: A Corpus-Based Study. Journal of Teaching Persian to Speakers of Other Languages, 5(12), 3-. https://www.magiran.com/paper/1797180

Jahangiri, N. (1980). A Sociolinguistic Study of Tehrani Persian [University of London]. London.

Jahani, C. (2018). Trends in Iranian and Persian Linguistics. In To bring the distant near: On deixis in Iranian oral literature (pp. 309). De Gruyter Mouton. https://doi.org/https://doi.org/10.1515/9783110455793-017

Johns, T. (1986). Micro-Concord: A language learner's research tool. System, 14(2), 151–162.

Jucker, A. H. (1993). The discourse marker well: A relevance-theoretical account. Journal of Pragmatics, 19(5), 435-452. https://doi.org/https://doi.org/10.1016/0378-2166(93)90004-9

Jucker, A. H., Smith, S. W., & Lüdge, T. (2003). Interactive aspects of vagueness in conversation. Journal of Pragmatics, 35(12), 1737-1769. https://doi.org/10.1016/s0378-2166(02)00188-1

Kamju, M. (2011). Tahlil-e Xatāhā-ye Vaʒegāni-ye Dāneʃāmuzān-e Āmoli Zabān Maqta'-e Dabirestān dar Neveʃtār-e Zabān- e Fārsi [The Analysis of Morphological Errors of Amoli-speaking Learners of Persian in Their Persian Writing at High School]. MA. Dissertation, Allameh Tabatabaei University, Tehran.

Karimi, S., Samiian, V., Stilo, D., & Samiian, V. (2008). Aspects of Iranian Linguistics. Newcastle-upon-Tyne: Cambridge Scholars Publisher.

Khanbabazadeh, K. (2016). Xatāhā-ye Nahvi-ye Tāleʃi Zabānhā dar Fārsi-ye Me'yār, [Syntactic Errors of Taleshi Speakers in Standard Persian]. Tehran: Andishmandan Kasra Publications.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1(1), 7-36. https://doi.org/10.1007/s40607-014-0009-9

Koutlaki, S. (2009). Two sides of the same coin: how the notion of 'face' is encoded in Persian communication. In F. H. Bargiela-Chiappini, M. (Ed.), *Face, Communication, and Social Interaction* (pp. 115-133). Equinox.

Koutlaki, S. A. (2002). Offers and expressions of thanks as face enhancing acts: *tæ'arof* in Persian. *Journal of Pragmatics, 34*, 1733–1756.

Lado, R. (1957). *Linguistics Across Cultures: Applied Linguistics for Language Teacher*. Ann Arbor, MI: University of Michigan Press.

Lakoff, G. (1973). Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic, 2*(4), 458-508.

Lakoff, R. (1975). Linguistic theory and the real world 1. *Language Learning*, *25*(2), 309-338.

Leech, G. (2000). Grammars of Spoken English: New Outcomes of Corpus-Oriented Research. *Language Learning*, *50*(4), 675-724. https://doi.org/10.1111/0023-8333.00143

Lewis, D. M. (2006). Discourse markers in English: a discourse-pragmatic view. In *Approaches to discourse particles* (pp. 43-59).

Love, R. (2020). *Overcoming Challenges in Corpus Construction*. Routledge. https://doi.org/https://doi.org/10.4324/9780429429811

Macaulay, R. K. (1991). " Coz It Izny Spelt When They Say It": Displaying Dialect in Writing. *American Speech*, 280-291.

Mahmoodi-Bakhtiari, B. (2018). Trends in Iranian and Persian Linguistics. In *Spoken vs. written Persian: Is Persian diglossic?* (pp. 183). De Gruyter Mouton. https://doi.org/https://doi.org/10.1515/9783110455793-011

Majd, N. (2002). “Barresi-ye Moqābelehi-ye Dastgāh- e Vākehi-ye Engelisi va Fārsi [A Contrastive Analysis of English and Persian Vowels].” *Pāyān Nāmeh-ye Kārshenāsi- ye Arshad, Dāneshgāh- e Tehrān.*

Matsumoto, Y. (1988). Reexamination of the universality of face: Politeness phenomena in Japanese. *Journal of pragmatics*, *12*(4), 403-426.

Mehdizadeh, N. (2007). Tahlil-e Xatāhā-ye Goftāri-ye Zabānāmuzān- e Kord Zabān- e Eilāmi dar Yādgiri va Kārbord- e Zabān- e Fārsi [The Analysis of the Spoken Errors of Kurdish- speaking Learners of Persian in Eilam]. MA. Dissertation, Allameh Tabatabaei University, Tehran.

McCarthy, M., & Carter, R. (1995). Spoken grammar: what is it and how can we teach it? *ELT Journal, 49*(3), 207-218.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge.

Mehdi Zādeh, N. O. (2008). “Tahlil-e Khtā-hā- ye Goftāri-ye Zabān Āmuzān- e Kord Zabān- e Ilāmi- ye Shahrestān-e Shirvān-e Chardāvol dar Yādgiri va Kārbord- e Zabān-e Fārsi [An Analysis of Pronunciation Errors of Ilami Kurdish Speakers

Learning Persian].” *Pāyān Nāmeh-ye Kārshenāsi- ye Arshad, Dāneshgāh- e ‘Allāmeh Tabātabā’i.*

Metsä-Ketelä, M. (2012). Frequencies of vague expressions in English as an academic lingua franca. *Journal of English as a Lingua Franca, 1*(2), 263-285.

Metsä-Ketelä, M. (2016). Pragmatic vagueness: Exploring general extenders in English as a lingua franca. *Intercultural Pragmatics, 13*(3). https://doi.org/10.1515/ip-2016-0014

Miller, C. (2011). A Holistic Treatment of/ān/to [un] in Persian. *ICPhS.*

Miller, C. (2012). Variation in Persian Vowel Systems. *Orientalia Suecana, LXI*, 156-169.

Miller, C., Livingston, C., Vinson, M., & Prado, T. T. (2014). Persian Dialects as Spoken in Iran. In. United States of America: University of Maryland.

Mirdehghan, M., et al. (2014). “Xatāhā-ye Neveʃtāri-ye Fārsiāmuzān-e Ālmāni Zabān dar Sath-e Moqaddamāti: Xatāhā-ye Emlāyi-vāji’ [Written Errors of German-speaking Learners of Persian at Elementary Level: An Orthophonemic Analysis].” *Journal of Teaching Persian to Non-native Speakers of Persian, 6*(3–1), 91–116.

Mohammadi, A. N. (2018). ‘Discourse markers in colloquial and formal Persian: a corpus-based discourse analysis approach’, unpublished PhD thesis, University of Florida.

Mohammadi, A. N. (2019). *Corpus of Conversational Persian: Introduction*. DOI: 10.13140/RG.2.2.20630.09286/1.

Mohammadi, A. N. (2019). *Meaning potentials and discourse markers: The case of focus management markers in Persian*. *Lingua*, 229, 102706.

Morādkhāni, M. (2008). "Barresi-ye Khatā-hā-ye Āvāyi-ye Fārsi Āmuzān-e Tork Zabān Hengām-e Sokhan Goftan beh Zabān-e Fārsi Az Didgāh-e Āvāshenāsi-ye Ākowstik" [An Investigation of Pronunciation Errors Made by Turkish Speakers Learning Persian]. *Pāyān Nāmeh-ye Kārshenāsi-ye Arshad, Dāneshgāh- e'Allāmeh Tabātabā'i*.

Motevalian Naeini, R., & R. Malekian. (2014). "Tahlil-e 'Xatāhā-ye Nahvi-ye Fārsiāmuzān-e Ordu Zabān" [The Analysis of the Syntactic Errors of Urdu-speaking Learners of Persian]. *Journal of Teaching Persian to Non-native Speakers of Persian*, 6(3–1), 31–64.

Motevalian Naeini, R., & A. Ostovar Abarghuyi. (2013). "Xatāhā-ye Nahvi-ye 'arab Zabānān dar Yādgiri-ye Zabān-e Fārsi be 'onvān-e Zabān-e Dovvom" [Syntactic Errors of Arabic-speaking Learners of Persian as a Second Language]. *Journal of Teaching Persian to Non-native Speakers of Persian*, 4(2–2), 57–86.

Müller, S. (2005). Discourse markers in native and non-native English discourse. *Discourse Markers in Native and Non-native English Discourse*, 1-310.

Najafi Eskandari, E. (2016). "Tahlil-e Khtā-hā-ye Āvāyi-ye Fārsi Āmuzān-e Ordu Zabān-e Pākestāni jā me'ye Almostafā" [Analysis of Pronunciation Errors of Urdu-Speaking Learners of Persian at AlMustafa In].

Nanbakhsh, G. (2011). *Persian address pronouns and politeness in interaction The University of Edinburgh*.

Norrick, N. R. (2009). Interjections as pragmatic markers. *Journal of pragmatics*, *41*(5), 866-891.

Novotný, T., & Malá, M. (2018). *General extenders in Czech*. *Casopis pro Moderni Filologii*, 100, 42-59.

O'Keeffe, A. (2004). "'Like the wise virgins and all that jazz' – using a corpus to examine vague categorisation and shared knowledge". *Language and Computers*, 52, 1-20.

O'Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge Handbook of Corpus Linguistics*. Routledge.

Osati, S. (2015). *Moqāyeseh-ye Nezām-e Āvāyi-ye Zabān-e Fārsi va Zabān-hā-ye Itāliyāyi/ Farānseh/ Espāniyāyi va Rāhkār-hā-ye Āmuzeshi [The Comparison of Persian Sound System with Italian/French/Spanish Sound System and Pedagogical Strategies]*. Pāyān Nāmeh-ye Kārshenāsi-ye Arshad, Dāneshgāh-e 'Allāmeh Tabātabā'i.

Östman, Jan-Ola. 1982. The symbiotic relationship between pragmatic particles and impromptu speech. In Nils Erik Enkvist (ed.), Impromtu Speech: A Symposium, 147–177. Åbo: Åbo Akademi.

Overstreet, M., & Yule, G. (1997). On being inexplicit and stuff in contemporary American English. *Journal of English Linguistics*, *25*(3), 250-258.

Overstreet, M. (1999). *Whales, Candlelight, and Stuff Like That: General extenders in English Discourse*. Oxford University Press.

Overstreet, M. (2005). *And stuff und so: Investigating pragmatic expressions in English and German*. *Journal of Pragmatics*, 37(11), 1845-1864. https://doi.org/https://doi.org/10.1016/j.pragma.2005.02.015

Overstreet, M. (2019). *The English general extender*. *English Today*, 1-6. https://doi.org/10.1017/s0266078419000312

Parvaresh, V., & Sheikhan, S. A. (2019). *Pragmatic functions of 'sort of' in Persian: A vague language perspective*. *Journal of Asian Pacific Communication*, 29(1), 86-110. https://doi.org/https://doi.org/10.1075/japc.00022.par

Parvaresh, V., Tavangar, M., Rasekh, A. E., & Izadi, D. (2012). *About his friend, how good is she, and this and that: General extenders in native Persian and non-native English discourse*. *Journal of Pragmatics*, 44(3), 261-279. https://doi.org/10.1016/j.pragma.2011.12.003

Parvaresh, V., & Tayebi, T. (2014). *Vaguely Speaking in Persian*. *Discourse Processes*, 51(7), 565-600. https://doi.org/10.1080/0163853X.2013.874545

Prince, E. F. F., Joel Bosk, Charles. (1982). *On hedging in physician -hysician discourse*. In R. J. di Pietro (Ed.), *Linguistics and the Professions*. Ablex.

Qarehbāqi, N. (2005). *Barresi-ye Moqābelehi-ye Āvā-hā-ye Zabān-e Fārsi va Torkamani va Ta'sir-e Ān dar Mahārat-e Shenidan [A Contrastive Study of Persian and Turkmen Sounds and Its Influence on Listening Comprehension]*. Pāyān Nāmeh-ye Kārshenāsi-ye Arshad, Dāneshgāh-e Shahid Beheshtiye Tehrān.

Reppen, R. (2010). Building a corpus: What are the key considerations? In M. J. McCarthy & O'Keeffe (Eds.), *The Routledge Handbook of Corpus Linguistics*. Routledge.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology*, *104*, 192-233.

Rühlemann, C. (2006). Coming to terms with conversational grammar: 'Dislocation' and 'dysfluency'. *International Journal of Corpus Linguistics*, *11*(4), 385-409.

Rühlemann, C. (2019). *Corpus Linguistics for Pragmatics: A guide for research* (1st ed.). Milton: Routledge. https://doi.org/10.4324/9780429451072

Ruzaite, J. (2018). Vague language in English L2: A focus on Lithuanian learner English.

Ruzaitė, J. (2018a). Discourse variation of vague language. Vague quantifiers in spoken and written Lithuanian. *Taikomoji kalbotyra*, *10*, 45-67.

Ruzaitė, J. (2018b). General extenders and discourse variation. *International Journal of Corpus Linguistics*, *23*(4), 467-493. https://doi.org/10.1075/ijcl.17019.ruz

Sabet, P. G., & Zhang, G. (2015). Communicating through vague language: A comparative study of L1 and L2 speakers. In R. M. Kempson, E. Gregoromichelaki, & C. Howes (Eds.), Language in Action: New Studies of Language in Society (pp. 259-275). Palgrave Macmillan. https://doi.org/10.1057/9781137486387

Sabet, P. G. P. (2013). Interaction through vague language: L1 and L2 perspectives [Curtin University]. USA.

Sabet, P. G. P., & Zhang, G. Q. (2018). The pragmatic functions of 'I Don't Think' and 'I Think + Not'. *Australian Journal of Linguistics*, *38*(3), 421-441. https://doi.org/10.1080/07268602.2018.1470459

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, *50*(4), 696–735. https://doi.org/10.2307/412243

Sadeghi, V., & Mansoory Hararehdasht, N. (2016). Persian sentence stress production by Mandarin Chinese speakers. *Journal of Teaching Persian to Speakers of Other Languages*, *5*(1), 95–119.

Safari, S. (2016). From corpus linguistics to learner corpora. *National Corpus Linguistics*, Tehran.

Safari, S. (2017). Constructing and analysing an error-tagged learner corpus of Persian [University of Belgrade]. Belgrade.

Saffar Moghadam, A. (2013). Spoken and written variants in teaching Persian language to non-Persian speakers. *Language Studies*, *3*(6), 45-68. http://languagestudy.ihcs.ac.ir/article_669_a3df3a155f1deede5ebb2062068ef39c.pdf

Sām, N. (2011). *Moshkelāt va Rāhkār-hā- ye Āmuzesh- e Talaffoz- e Zabān- e Fārsi beh Gheyr – e Fārsi Zabānān- e Khāreji* [Teaching Persian Pronunciation to Foreign Non-Native Speakers: Challenges and Approaches]. *Pāyān Nāmeh-ye Kārshenāsi-ye Arshad, Dāneshgāh- e 'allāmeh Tabātabā'i*.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society, 29*(1), 1–63. https://doi.org/10.1017/S0047404500001019

Schiffrin, D. (1987). *Discourse Markers* (Studies in Interactional Sociolinguistics). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511611841

Schmidt, R. W. (1990). The role of consciousness in second language learning1. *Applied linguistics*, *11*(2), 129-158.

Schourup, L.C. (1985). Common Discourse Particles in English Conversation (1st ed.). Routledge. https://doi.org/10.4324/9781315401584

Secova, M. (2017). Discourse-pragmatic variation in Paris French and London English: Insights from general extenders. *Journal of Pragmatics*, *114*, 1-15.

Sedighi, A. (2010a). *Agreement Restrictions in Persian*. Amsterdam: Leiden University Press.

Sedighi, A. (2010b). *Teaching Persian to Heritage Speakers*. *Iranian Studies*, 43(5), 683-697. https://doi.org/10.1080/00210862.2010.518033

Sedighi, A. (2015). *Persian in Use: An Elementary Textbook of Language and Culture*. Leiden University Press.

Sedighi, A., & Shabani-Jadidi, P. (2018). *The Oxford Handbook of Persian Linguistics*. Oxford: Oxford University Press USA - OSO.

Seraji, M. (2015). *Morphosyntactic Corpora and Tools for Persian* (PhD Thesis). *Studia Linguistica Upsaliensia*, 16.

Shabani-Jadidi, P. (2018). Heritage Learners' versus Second Language Learners' Source of Errors in Advanced-Level Writing: Case of a Persian Media Course. Iranian Studies, 51(5), 747-778. https://doi.org/10.1080/00210862.2018.1496323

Shabani-Jadidi, P. (2020). The Routledge handbook of second language acquisition and pedagogy of Persian. New York, NY: Routledge.

Svartvik, J. (1980). Computer-aided grammatical tagging of spoken English. In COLING 1980 Volume 1: The 8th International Conference on Computational Linguistics.

Terraschke, A. (2010). Or so, oder so, and stuff like that - General extenders in New Zealand English, German and in learner language. *Intercultural Pragmatics*, *7*(3),

449-469. https://doi.org/10.1515/IPRG.2010.020Pragmatic connectives. Journal of Pragmatics, 3(5), 447-456. https://doi.org/10.1016/0378-2166(79)90019-5

Viana, V. (Ed.). (2022). Teaching English with Corpora: A Resource Book (1st ed.). Routledge. https://doi.org/10.4324/b22833

Waltereit, R. (2006). The rise of discourse markers in Italian: a specific type of language change. In *Approaches to discourse particles* (pp. 61-76).

Watts, R. J. (2003). Politeness. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511615184

Wilson, D., & Sperber, D. (1986). On defining relevance. *Philosophical grounds of rationality: Intentions, categories, ends*, 243-258.

Wouk, F. (2001). Solidarity in Indonesian conversation: The discourse marker ya. Journal of Pragmatics, 33(2), 171-191. https://doi.org/10.1016/S0378-2166(99)00139-3

Yousef, S., & Torabi, H. (2018). Persian: A comprehensive grammar. Routledge.

Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8, 338-353.

Zhang, G. (2011). Elasticity of vague language. Intercultural Pragmatics, 8(4), 415-442. https://doi.org/10.1515/iprg.2011.026

Zhang, G. (2020). Vague language challenged. International Review of Pragmatics, 12, 107-134.

Zhang, G., & Parvaresh, V. (2019). Elastic Language in Persuasion and Comforting. In A. Capone, F. Kiefer, & F. Lo Piparo (Eds.), Perspectives on Linguistic Pragmatics

(pp. 307-330). Springer International Publishing. https://doi.org/10.1007/978-3-030-28460-2

Zhang, G. Q., & Le, N. N. (2018). Vague Language, Elasticity Theory and the Use Of 'Some': A Comparative Study of L1 and L2 Speakers in Educational Settings. In D. Lasagabaster, A. Doiz, & J. M. Sierra (Eds.), Motivation and Foreign Language Learning: From Theory to Practice (pp. 57-74). Bloomsbury Publishing Plc. Retrieved from http://ebookcentral.proquest.com/lib/ed/detail.action?docID=5358469

Zowghdar Moghadam, R., & Dabir Moghadam, M. (2002). Discourse Markers. Pazhuhesh-e Zabanhaye Khareji, -(12), 61-78. Retrieved from https://www.sid.ir/en/Journal/ViewPaper.aspx?ID=35591

Zvěřinová, S. (2016). N-gramy v mluveném projevu českých a rodilých mluvčích angličtiny [N-grams in the Spoken Discourse of Czech and Native English Speakers]. Charles University of Prague. Retrieved from http://oatd.org/oatd/record?record=handle:20.500.11956%2F79338&q=learner%20corpora

APPENDICES

APPENDIX I: Participant Information Sheet for Pilot Study

Research Project: Informal Persian spoken by learners of Persian in the UK
Information Sheet for Participants

Researcher
Sepideh Daghbandan, PhD Research Student

Participation and Purpose of the Pilot Study

You are being invited to take part in a pilot study which is aimed at improving the design and future compilation of a spoken corpus. This corpus will be used to conduct a study exploring informal Persian spoken by learners of Persian in the UK. The results and findings of this study could be used to inform Persian teachers and course/materials designers.

Before you decide to participate in this pilot study, please take the time to read the following information sheet carefully. If there is anything that is not clear or if you need more information on a particular section, please feel free to ask the researcher.

Participation Procedure

If you agree to participate in this pilot study, you will be asked to:

- Have your normal conversation with a friend or classmate in Persian, on whatever topic comes naturally, and allow it to be audio recorded
- Participate in an interview regarding your experience with the recorded conversation

Confidentiality

All the information you provide will be kept strictly confidential. The researcher will be the only individual to view and maintain your contact information and provided data. As soon as the recording is completed, all identifying information will be anonymised by assigning a unique number. The researcher will not use your information for any purpose outside of this research project. Once the project is completed, all data will be securely destroyed.

Benefits of Participating

By participating in this study, you will benefit from practicing in your second language, Persian. In addition, you will be contributing to the purpose of this study which is to inform Persian teachers and course/materials designers about the needs of learners of Persian.

Voluntary Participation

Your participation in this pilot study is voluntary. This means that, if you decide to join the pilot study now, you are at liberty to leave it if you change your mind about taking part. You may skip any questions that you may feel uncomfortable answering.

Future Correspondence

Sepideh Daghbandan
Address: Room 3.22, St. Leonard's Land Moray House School of Education, University of Edinburgh, EH8 8AQ
Telephone: +44 (0)131 631 4109
Email:

APPENDIX II: Participant Consent Form for Pilot Study

Participation Consent form for Pilot Study

Research Project: Informal Persian spoken by learners of Persian in the UK

Researcher: Sepideh Daghbandan

Participant Consent

I confirm that I have read and understood the Information Sheet for the mentioned research project. I understand that my participation is voluntary. I also understand that I am free to withdraw at any time from participating in this pilot study. I understand that any information that is obtained in connection with this pilot study will remain confidential and will be used only for the purpose of research.

I confirm that I give my permission for my spoken conversation, and question and answer session with the researcher to be audio recorded.

I confirm that I have had the opportunity to ask any question on all the sections of the information sheet.

I confirm that I have received a copy of the information sheet and consent form.

Participant's Name _____
Participant's Email Address _____
Participant's Signature_____
Date_____

Researcher's Name_____
Researcher's Signature_____ Date _____

APPENDIX III: Participant Information Sheet for Corpus Compilation

Research Project: Informal Persian spoken by learners of Persian in the UK
Information Sheet for Participants

Researcher

Sepideh Daghbandan, PhD Research Student

Participation and Purpose of the Study

You are being invited to take part in the compilation of a spoken corpus. This corpus will be used to conduct a study exploring informal Persian spoken by learners of Persian in the UK. The results and findings of this study could be used to inform Persian teachers and course/materials designers.

Before you decide to participate in this study, please take the time to read the following information sheet carefully. If there is anything that is not clear or if you need more information on a particular section, please feel free to ask the researcher.

Participation Procedure

If you agree to participate in this study, you will be asked to:

- Have your normal conversation with a friend or classmate in Persian, on whatever topic comes naturally, and allow it to be audio recorded
- Provide the researcher with audio recorded answers to questions regarding certain biographical information.

Confidentiality

All the information you provide will be kept strictly confidential. The researcher will be the only individual to view and maintain your contact information and provided data. As soon as the recording is completed, all identifying information will be anonymised by assigning a unique number. The researcher will not use your information for any purpose outside of this research project. Once the project is completed, all data will be securely destroyed.

Benefits of Participating

By participating in this study, you will benefit from practicing in your second language, Persian. In addition, you will be contributing to the purpose of this study which is to inform Persian teachers and course/materials designers about the needs of learners of Persian.

Voluntary Participation

Your participation in this study is voluntary. This means that, if you decide to join the study now, you are at liberty to leave it if you change your mind about taking part. You may skip any questions that you may feel uncomfortable answering.

Future Correspondence

Sepideh Daghbandan

Address: Room 3.22, St. Leonard's Land Moray House School of Education, University of Edinburgh, EH8 8AQ

Telephone: +44 (0)131 631 4109

Email:

APPENDIX IV: Participant Consent Form for Corpus Compilation

Participation Consent Form for Study

Research Project: Informal Persian spoken by learners of Persian in the UK

Researcher: Sepideh Daghbandan

Participant Consent

I confirm that I have read and understood the Information Sheet for the mentioned research project. I understand that my participation is voluntary. I also understand that I am free to withdraw at any time from participating in this study. I understand that any information that is obtained in connection with this study will remain confidential and will be used only for the purpose of research.

I confirm that I give my permission for my spoken conversation, and question and answer session with the researcher to be audio recorded.

I confirm that I have had the opportunity to ask any question on all the sections of the information sheet.

I confirm that I have received a copy of the information sheet and consent form.

Participant's Name _____
Participant's Email Address _____
Participant's Signature_____
Date_____

Researcher's Name_____
Researcher's Signature_____ Date _____

APPENDIX V: Top 50 Positive Keywords of LoPSC

Type	Frequency LOPSC	Dispersion LOPSC	Frequency RC	Dispersion RC	Keyness Value (simple maths)
amâ (but)	293	0.37	3	3.25	74.128
ou (he/her/him/her	192	0.62	5	2.54	53
baleh (yes)	485	0.49	69	1.012	28.8
barâyeh (for)	270	0.33	43	1.53	28.75
kami (a little/ a bit)	47	0.6	1	4.35	27.84
mikonam (I do)	272	0.31	73	0.79	8.99
mardom (people)	79	0.72	18	1.19	8.53
kasâni (people)	19	0.54	0	0	8
niyâz (need)	19	0.80	0	0	8
ziyâdi (too much)	26	0.48	3	2.69	7.44
dar (in)	498	0.28	170	0.62	7.41
say (try)	23	0.55	2	4.35	7.4
motefâvet (different)	17	1.45	0	0	7.27
hastan	51	0.52	13	1.64	7

(they are)					
chon (because)	204	0.3	75	0.6	6.6
cheghadr (how much)	14	0.36	0	0	6.16
mânand (like/similar)	14	1.39	0	0	6.16
Farz (imagine)	18	0.89	2	3.56	5.96
Hamchenin (as well)	13	0.65	0	0	5.79
tor (way)	25	0.44	6	1.64	5.54
mitouni (you can)	35	0.63	11	1.87	5.46
bazi (some)	24	0.91	6	2.33	5.34

APPENDIX VI: Top 50 LopSC Negative Keywords

Type	Frequency LOPSC	Dispersion LOPSC	Frequency RF	Dispersion RF	Keyness Value

					(simple maths)
kardeh (has done)	2	1.41	78	0.85	0.14
yarou (person anonymous)	0	0	42	1.23	0.14
inam (he/she as well)	0	0	43	1.02	0.14
gofteh (has said)	1	2.23	62	1.16	0.14
pounsad (five hundred)	0	0	44	1.95	0.13
hashtâd (eighty)	0	0	44	1.62	0.13
jouriyeh (it is in a way)	0	0	45	0.98	0.13
mikhoreh (eats)	0	0	45	1.46	0.13
kon (do)	5	0.830011	144	0.457428	0.133831119

chizeh (the thing is)	2	1.41	88	0.76	0.12
sar (head)	3	1	112	0.89	0.12
zareh (bit)	2	1.41	92	0.81	0.12
rafteh (gone)	1	2.23	75	0.76	0.11
bia (come)	1	2.23	77	0.97	0.11
oumad (came)	0	0	55	1.01	0.11
mageh (if)	0	0	55	1.03	0.11
zang (call)	2	1.42	101	1.5	0.11
zadeh (has hit)	0	0	56	0.96	0.11
sho (become)	1	2.23	80	0.99	0.11
var	0	0	57	1.04	0.11
biyad (come)	0	0	58	1.17	0.1
halet (your health)		0	61	1.11	0.1

manam (me as well)	0	0	62	0.81	0.1
barâ (for)	0	0	62	1.05	0.1
payin (down)	0	0	63	0.83	0.10
oumadeh (came)	0	0	65	0.94	0.09
âghâ (sir)	0	0	65	1.25	0.09
miyâd (comes)	2	1.41	121	0.75	0.096
asan (not at all)	1	2.23	95	1.17	0.09
vâseh (for)	0	0	82	0.95	0.07
kolan (Completely)	0	0	84	0.94	0.07
boro (go)	0	0	87	0.69	0.07
hey (again)	0	0	90	0.69	0.073
douneh (classifier)	0	0	93	0.92	0.07

âkheh (but)	0	0	107	0.57	0.06
bebin (look)	0	0	148	0.71	0.04

APPENDIX VII: LOPSC Frequency List

Item	Translation	Frequency
که (keh)	that	555
و (va)	and	385
آره (areh)	yeah	309
در (dar)	in	265
این (in)	this	260
به (beh)	to	245
بله (baleh	yes	235
خیلی (kheili)	very/ a lot	190
برای (baraye)	for	158
بود (boud)	was	156
از (az)	from	154
نه (nah)	no	144
من (man)	me/I	142
اما (ama)	but	133
فک (fek)	think	133
خوب (khoub)	good	129

اره (areh)	yeah	125
یه (yeh)	one	124
رو (ru)	"direct object indicator"	120
یعنی (yani)	it means	118
نمیدونم (nemidoonam)	I don't know	114
میکنم (mikonam)	I do	111
یک (yek)	one	106
چون (chon)	because	106
ولی (vali)	but	94
چیز (chiz)	thing	86
هست (hast)	is	86
با (ba)	with	81
نیست (nist)	isn't	77
اگه (ageh)	if	76
است (ast)	is	75
یا (ya)	or	74
واقعاً (vaghean)	really	72
هم (ham)	as well	71
کار (kar)	thing/work	70
ما (ma)	we	67
مردم (mardom)	people	67
او (ou)	s/he	63
چی (chi)	what	62
آن (an)	that/it/he/she	62

باید (bayad)	must		62
میدونی (midooni)	you know		54
دیگه (digeh)	?		54
خب (khob)	well		54
همه (hameh)	all		51
گفت (goft)	said		49
فقط (feghat)	only		49
چه (cheh)	what		47
صحبت (sohbat)	speak/speech		42
همین (hamin)	this		42
چیزی (chizi)	thing		41
تو (tu)	you		40
مثلاً (masalan)	for example		39
مثل (mesl)	similar to		38
چرا (chera)	why		38
صد (sad)		100	36
درست (dorost)	right		35
اصلاً (aslan)	not at all		35
میکنن (mikonanan)	they do		33
پس (pas)	then		32
داره (dareh)	has		31
را (ra)	direct object indicator		31
آنها (anha)	they/them		31
کردم (kardam)	I did		30

APPENDIX VIII: Reference Corpus Frequency List

Item	Translation	Frequency
که (keh)	that	1269
و (va)	and	1112
این (in)	this	1091
یه (yeh)	a	998
دیگه (digeh)	discourse marker with no Englsih equivalent	874
تو (tu)	you (informal)	829
من (man)	I	782
به (beh)	to	708
آره (areh)	yeah	702
رو (ru)	"direct object indicator"	676
از (az)	from	673
بعد (bad)	after	607
نه (nah)	no	604
هم (ham)	as well	597
بود (boud)	was	539
اون (oun)	that/it/him/her	522
خب (khob)	well	461

با (ba)	with	452
خیلی (kheili)	very	422
چی (chi)	what	419
مثلاً (masalan)	for example	373
چه (cheh)	what	338
تا (ta)	"direct of a number"/ to	337
چیز (chiz)	thing	310
داره (dareh)	has	309
اینا (ina)	them	295
دو (du)	two	287
الان (alan)	now	282
باید (bayad)	have to	272
گفتم (goftam)	said	245
فکر (fekr)	think	241
کنم (konam)	I do	237
حالا (hala)	now	230
ولی (vali)	but	227
چرا (chera)	why	218