Towards a Design Process for the Sociolinguistic Corpus of Spoken Saudi Arabic

Dr. Yahya Abdu A. Mobarki
Assistant Professor of Linguistics and Applied Linguistics
English Department, Faculty of Arts and Humanities, Jazan University

ymobarki@jazanu.edu.sa

Towards a Design Process for the Sociolinguistic Corpus of Spoken Saudi Arabic

Dr. Yahya Abdu A. Mobarki

(Received 06/04/2020; accepted 02/05/2020; Published 14/02/2021)

Abstract: This paper has two main goals. First, it discusses how the availability as well as the accessibility of spoken corpora of English, for example, or any other language (e.g., Spanish) have contributed to the sociolinguistic studies of discourse markers (DMs). Looking forward to being informed by the content from books and other materials adopting some useful methods in the literature, the paper, as an initial step, delineates the significance of building a sociolinguistic corpus of spoken Saudi Arabic. Second, the paper explores three basic steps and procedures for the construction of a spoken Saudi Arabic corpus: (1) the collection of spoken sources and structures of data (e.g., talk shows, radio, oral narratives, field work, and sociolinguistic interviews), (2) the development of transcription protocols, and (3) the incorporation of metadata for this kind of sociolinguistic spoken corpus. The paper discusses some of the expected methodological challenges associated with the three discussed steps and procedures of corpus building (e.g., data collection and adding and/or modifying sociolinguistic categories).

Keywords: Saudi Arabic; corpus linguistics; sociolinguistics; spoken discourse; discourse markers.

❊ ❊ ❊

نحو عملية تصميم لمدونة لغوية اجتماعية للعربية السعودية المنطوقة

د. يحيىٰ عبده مباركي

(قدم للنشر في 13/ 08/ 1441هـ؛ وقبل للنشر في 09/ 09/ 1441هـ؛ ونشر في 02/ 07/ 1442هـ)

المستخلص: تحتوي هذه الورقة علىٰ هدفين رئيسيين. أولاً، تناقش كيف أسهمت إتاحة مدونات اللغة الإنجليزية المنطوقة وكذلك إمكانية الوصول إليها، علىٰ سبيل المثال، أو أي لغة أخرىٰ (مثل الإسبانية) في الدراسات اللغوية الاجتماعية لأدوات الخطاب (DMs). وبالنظر إلىٰ المستقبل والاطلاع علىٰ محتوىٰ كتب ومواد أخرىٰ ذات الفائدة المنهجية من الأدبيات، فإن هذه الورقة، كخطوة مبدئية، تحدد أهمية بناء مدونة لغوية اجتماعية منطوقة من العربية السعودية. ثانيا، تستكشف الورقة ثلاث خطوات وإجراءات أساسية لبناء المدونة المنطوقة للعربية السعودية: (1) جمع مصادر المواد المنطوقة وتراكيب البيانات (علىٰ سبيل المثال، البرامج الحوارية، والإذاعة، والسرد الشفهي، والعمل الميداني، والمقابلات اللغوية الاجتماعية)، (2) تطوير بروتوكولات التفريغ الكتابي، و(3) تضمين البيانات الوصفية لهذا النوع من المدونة اللغوية الاجتماعية المنطوقة. تناقش الورقة بعض التحديات المنهجية المتوقعة والمرتبطة بالخطوات والإجراءات الثلاثة التي تمت مناقشتها لبناء المدونة (علىٰ سبيل المثال، جمع البيانات وإضافة و/ أو تعديل الفئات اللغوية الاجتماعية).

الكلمات المفتاحية: العربية السعودية؛ لسانيات المدونات؛ علم اللغة الاجتماعي؛ الخطاب المنطوق؛ الأدوات الخطابية.

***

1. Introduction and objectives

The availability and the accessibility of different spoken English corpora have greatly benefited sociolinguistic investigations, which have shown discourse variation in general, and variation upon using discourse markers (henceforth, DMs) in particular. Such corpora have facilitated sociolinguistic studies of DMs through which scholars can investigate, compare, and contrast DMs across different social variables including context, gender, age, region, style, and socio-economic status, to name a few. However, the case in Arabic is different with a relative to complete lack of spoken Arabic corpora.

Therefore, this paper has two main goals. First, it extensively reviews the literature available on discourse variation, digging deeply into the literature to show the methodologies that have significantly helped sociolinguistic investigations of DMs. Importantly, the paper shows how the availability as well as the accessibility of spoken corpora of English, for example, or any other language (e.g., Spanish) have contributed to the sociolinguistic studies of DMs. Second, the review and critique are taken as a point of departure to suggest an initial design for the sociolinguistic corpus of spoken Saudi Arabic, which serves Saudi Arabic linguistics studies in general, and Saudi Arabic sociolinguistics studies in particular, including sociolinguistic studies of DMs. Looking forward to being informed by the contents of methodology books and other materials from the literature, the paper, as an initial step, delineates the significance of building such a spoken corpus of Saudi Arabic. Then, the paper explores three basic steps and procedures for a corpus construction: the collection of spoken sources and structures of data (e.g., talk shows, radio, oral narratives, field work, and sociolinguistic interviews), the development of transcription protocols, and the incorporation of metadata for this kind of sociolinguistic spoken corpus. The paper discusses some of the expected methodological challenges associated with the three discussed steps and procedures of corpus building (e.g., data collection and adding and/or modifying sociolinguistic categories).

Proposing a spoken corpus for Arabic, however, is overambitious for several reasons: (1) the extreme variation between varieties of Arabic, (2) the fact that nowadays, Arabic is becoming more of a designated term for populations who even never speak the language, and (3) the fact that such a substantial effort is expected to meet challenges and obstacles (e.g., collaborative, sociopolitical, academic, and financial) that might obstruct or even stop initiatives for a corpus construction. Therefore, this proposal concerns steps for building a spoken sociolinguistic corpus for Saudi Arabic; which in this case is a cover term used to refer to the linguistic varieties spoken in Saudi Arabia.

This paper is structured as follows: Section 2 details the approach and procedures followed for originating this paper. Section 3 is a general background, which sets up corpus linguistics, sociolinguistics, and DMs as the major themes. Section 4 introduces the design process for the spoken corpus of Saudi Arabic. Subsection 4.1 discusses the rationale and the purpose of the design process. Section 4 also explores three procedures of great importance for constructing a corpus that fulfills sociolinguistic queries of DMs, namely: sources, collection and structures of data (subsection 4.2); transcription protocols (subsection 4.3); and metadata (subsection 4.4). Section 5 is the conclusion.

2. Approach and procedures

This section describes the approach for originating this paper leading to the proposal and design process for building and launching a spoken Saudi Arabic corpus that meets sociolinguistic queries about Arabic DMs. First, the literature was searched mainly looking for sociolinguistic treatments, which can help demonstrate the necessary steps and procedures for constructing corpora that meet sociolinguistic queries (Baker, 2010; Biber, 2009; Childs et al., 2011; Kendall, 2011; Kendall & Van Herk, 2011). Journal articles such as Davies (2009; 2010) informed this proposal of a spoken corpus, so did materials from methodology books whose interest revolves around corpora and corpora construction (Davies, 2013). Scholarly treatments mentioned previously inspired significantly this paper and the spoken corpus proposal. However, since the purpose of this project is to propose a corpus that fulfills sociolinguistic queries of DMs, the treatments mentioned previously were general in nature and lacking essential information on how to specifically build corpora for sociolinguistic investigations of DMs. As such, the literature was searched again to locate empirical scholarly work that addressed DMs from sociolinguistic perspectives. The approach, at this stage, was purely methodological, which enabled the categorization of the found sociolinguistic investigations of DMs into two categories, namely: (1) non corpus-based sociolinguistic treatments of DMs; or sociolinguistic treatments based on 'unconventional' corpora— privately constructed corpora by scholars interested in these linguistic elements (Babel, 2014; Dailey O'Cain, 2000; Erman, 1992; Fuller, 2003; Holmes, 1986; 1988; 1990; Huspek, 1989; Macaulay, 2002; Meyerhoff, 1994; Sankoff et al., 1997; Tagliamonte, 2005; Wouk, 1999), and (2) corpus-based sociolinguistic treatments of DMs; or sociolinguistic treatments of DMs based on 'conventional' corpora— large, publicly available corpora or corpora where access was granted (Andersen, 2001; Barbieri, 2008; Grant, 2010; Erman, 2001; Levey, 2006; Miller, 2009; Pichler, 2009; Stubbe & Holems, 1995; Torgesrsen et al., 2011; Tottie, 2011). Overall, the sociolinguistic treatments of DMs compensated for the shortage in the views of the earlier work concerning the construction of sociolinguistic corpora. These empirical studies explicitly helped by providing important information, for example in terms of the sources and data collection procedures and the social variables regularly influencing the frequency and use of DMs. These studies also gave essential views about the metadata required. Most importantly, these empirical treatments contained views on transcription that were important for the sociolinguistics of DMs.

3. Background

3.1 Corpora and corpus linguistics

Baker (2010) observes that the word *corpora* is the plural form of the word corpus, originally a Latin word meaning *body*. Based on this etymological relationship, Baker (2010, p. 6) defined corpora "as a 'body' of language, or more specifically, a (usually) very large collection of naturally occurring language, stored as computer files.” Along similar lines, Davies (2013, p.210) defined corpora as “searchable collections of spoken and written language (nearly always in electronic format) which can be used for linguistic analysis.” What these two definitions have in common is the view that corpora feature “true balancedness, representativity, machine-readability” (Kendall & Van Herk, 2011, p. 1), and proper characteristics maintained by corpus linguists (Kendall, 2011). Recently, sociolinguists have found corpora and corpus-based methods advantageous. The next subsection discusses the connections between corpus linguistics and sociolinguistics.

3.1.1 Corpus linguistics and sociolinguistics

The use of corpora to aid sociolinguistic analysis is not a new trajectory of research interest in sociolinguistics; rather, corpora, more specifically described as conventional corpora, have always enhanced sociolinguistic investigations from the earliest origins of the sociolinguistics field. However, it was not until the past decade that sociolinguistics as a field found connections with the field of corpus linguistics such as sharing the nature of quantitative and qualitative analysis, and the empirical analysis of actual language use (Baker, 2010; Gregersen & Barner-Rasmussen, 2011; Kendall, 2011; Kendall & Van Herk, 2011). With these connections in mind, Baker (2010) and Kendall (2011) described the mainstream sociolinguists as still uninterested in enhancing such relationships because of issues related to variability, comparability, and representativeness.

Baker’s (2010) recent book, *Sociolinguistics and Corpus Linguistics*, highlighted the previous sociolinguistic literature that utilized corpora and corpus-based methodologies in their conventional sense. For instance, Baker (2010) presented corpus-based sociolinguistic research on sex-related differences, age-related differences, register-related differences, genre-related differences, and a wide range of sociopolitical differences. Additionally, Baker (2010) reviewed sociolinguistic corpus-based research that shed light on phonological patterns, morphological patterns, syntactic patterns, and discourse patterns influenced by complex sociolinguistic factors. In summary, although "[s]ociolinguists have been slower to adopt conventional corpora for research" (Kendall, 2011, p. 368), the importance of corpora and corpus-based methods is growing in sociolinguistic research (Baker, 2010; Gregersen & Barner-Rasmussen, 2011; Kendall, 2011, Kendall & Van Herk, 2011).

3.2 Discourse markers:

Discourse markers are of special interest in this paper for a number of reasons. First, for a concrete discussion, this paper needs concrete sociolinguistic examples. Second, DMs still present persistent and challenging theoretical and methodological questions that need: (a) scholarly attention and researchable formulations (e.g., language variation and change), (b) development of both innovative quantitative and qualitative methodologies (e.g., functional paradigm/analysis and language variation), and (c) large collections of data (e.g., corpora). Third, the sociolinguistic investigations of DMs are still blossoming. Fourth and most importantly, based on regular research and an extensive review of the literature on Arabic linguistics, DMs were selected to highlight the fact that these linguistic elements are completely dismissed from Arabic perspectives. Fifth, the absence of reliable spoken Arabic corpora, along with other factors, might be one significant factor contributing to such an inexcusable dismissal.

3.2.1 What are discourse markers?

The literature shows that DMs have risen to become one of the most studied topics by Western linguists, largely in English. DMs have been researched carefully and discussed extensively over the past two decades from different perspectives and approaches, for example discourse coherence models (Lenk, 1998; Schiffrin, 1987; 2001), grammatical and semantico-pragmatic models (Fraser, 1999; 2009), and relevance-based approaches (Schourup, 1999; 2011).

The literature also shows that these lexical items challenge any attempt to accurately define them, or to group them under one lexical or syntactic category or appropriately delineate their functions for a number of reasons. First, DMs are seen to belong to different lexical categories including conjunctions (e.g., *and, but, or, because*) and adverbs (e.g., *now, then*). Some DMs are considered clauses (e.g., *y'know*) and some are not (e.g., *I mean; mean* is a transitive verb requiring the presence of a completion [Fraser, 1999; Schiffrin, 1987]). Schourup (1999) adds interjections (e.g., *oh* and *gosh*) and verbs (e.g., *look*, *say*, and *see*). Second, DMs do not have restricted syntactic positions (Fraser, 1999; Gonzalez, 2004; Muller, 2005; Schiffrin, 1987). To support that, Schiffrin (1987, p.31) claimed that:

> Although markers often precede sentences, […] they are independent of sentential structure. Removal of a marker from its sentence initial position, in other words, leaves the sentence structure intact. Furthermore, several markers — *y'know, I mean, oh, like* — can occur quite freely within a sentence at locations which are very difficult to define.

Namely, some DMs appear in clause-initial position, clause-medial position, or even clause-final position (Fraser, 1999; Gonzalez, 2004; Muller, 2005; Schiffrin, 1987) as in the following instances of *y'know* (adapted from Schiffrin, 1987, p. 275, 276, 282):

- *Y'know* they say an apple a day keeps the doctor away?
- I'm not a—… we're all no perfect, *y'know*.
- and she said, *y'know*, 'I got a problem Zelda.'

Third, DMs are seen to be multifunctional lexical items (Gonzalez, 2004; Muller, 2005; Schiffrin, 1987) (e.g., *like* signaling approximation, lexical focus, and quotative *like*) "that predicate changes in the speaker's cognition, attitudes, and beliefs and facilitate the transmission of illocutionary force and intentions" (Gonzalez, 2004, p. 1).

The ambiguities and disagreements around the semantic, structural, and pragmatic features of DMs are also reflected through the terms and definitions proposed for such a class of linguistic items. These terms result from the fact that scholars, interested in such linguistic phenomena, investigate DMs from different perspectives and approaches. These linguistic items have been called (among other things), discourse markers, discourse connectives, discourse particles, discourse signals, discourse operators, cue phrases, pragmatic markers, pragmatic connectives, pragmatic particles, formulaic expressions, particles, text organizers, and pop-markers (Fraser, 1999; Gonzalez, 2004; Muller, 2005; Schourup, 1999).

3.2.2 Discourse markers and sociolinguistics

DMs have received comparatively scholarly attention from several sociolinguistic aspects (Aijmer & Simon-Vandenbergen, 2011). Sociolinguistic work on DMs has shown that there are links between DMs and a number of social factors, including: gender (Erman, 1992; Holmes, 1986; 1988; 1990; Wouk, 1999), age (Andersen, 2001; Barbieri, 2008; Dailey-O'Cain, 2000; Erman, 2001; Tagliamonte, 2005), social class (Huspek,1989; Stubbe & Holmes, 1995), ethnicity (Meyerhoff, 1994), language contact situations (Sankoff et al., 1997), and geographical region (Huddlestone & Fairhurst, 2013), across language varieties, for instance British English compared to New Zealand English (Grant, 2010), and in specific speech contexts, for instance the use of *well* in court (Innes, 2010), and the use of *like* in interviews (Fuller, 2003).

Also, there are studies which investigated the impact of multiple social variables on the frequency and use of DMs, for instance gender, social class, and age in Macaulay (2002), sex, ethnicity, and geographical location in Torgersen et al. (2011), gender, age, and socio-economic status in Tottie (2011), and gender and age in Levey (2006) and Pichlar (2009). Interestingly enough, DMs have been approached from sociolinguistic indexical perspectives (Babel, 2014). In this recently published sociolinguistic investigation, Babel (2014) examined how the Spanish DM *pues* can be a highly salient index of regional as well as stereotypical identities in Bolivian Valley Spanish.

A close look at the sociolinguistic treatments of DMs has also shown that DMs have been sociolinguistically investigated in ‘unconventional (spoken) corpora’ (Babel, 2014; Dailey O’Cain, 2000; Erman 1992; Fuller, 2003; Holmes, 1986; 1988; 1990; Huspek, 1989; Macaulay, 2002; Meyerhoff, 1994; Sankoff et al., 1997; Tagliamonte, 2005; Wouk, 1999) reaching to the point where these linguistic elements (i.e., DMs) are being investigated in larger and comprehensive ‘conventional (spoken) corpora’ (Andersen, 2001; Barbieri, 2008; Grant, 2010; Erman, 2001; Levey, 2006; Miller, 2009; Pichler, 2009; Stubbe & Holmes; Torgesrsen et al., 2011; Tottie, 2011). These sociolinguistic empirical investigations of DMs significantly informed the steps and procedures of the proposal and the design process of the sociolinguistic corpus of spoken Saudi Arabic in the pages that follow. The next sections and subsections introduce and discuss this proposal and the design process.

4. The design process for the sociolinguistic corpus of spoken Saudi Arabic

This section mainly introduces the proposal and the design process of establishing and launching a sociolinguistic corpus of spoken Saudi Arabic. Taking together all the surveyed sociolinguistic studies of DMs in combination with the surveyed scholarly treatments of constructing sociolinguistic corpora, this section considers the following: what are the kinds of aspects that need consideration and incorporation (and also how) while building a corpus of spoken Saudi Arabic intended to fulfill sociolinguistics queries of DMs? Note that this paper does not consider technical models and/or perspectives of building corpora. The author does not have enough technical background in addition to lacking experience dealing with corpora technicalities.

‘Corpus sociolinguistic’ (Baker, 2010) principles guide this proposal and the design process for the sociolinguistic corpus of spoken Saudi Arabic, and therefore inform every step and stage of its construction and establishment. Corpus sociolinguistic principles include: (1) sampling and representativeness, (2) relatedness, (3) reliability, (4) transparency and consistency (5) well-balancedness, (6) comparability, (7) contextualization, (8) naturalistic data and authenticity, and (9) the principle of accountability in relation to sociolinguistic corpora. Thus, these corpus sociolinguistic principles, in addition to selecting DMs as a concrete example in this paper, should inform the rationale and the purpose of the design (subsection 4.1), sources, collection and structures of data (subsection 4.2), transcription protocols and the degree of precision and faithfulness of transcriptions (subsection 4.3), and the metadata that should be included in the corpus (subsection 4.4). Briefly, the discussions in the following subsections are mainly guided by the corpus sociolinguistic principles and the selection of DMs. The next discussion considers the rationale and the purpose of proposing and designing the sociolinguistic corpus of spoken Saudi Arabic.

4.1 The rationale and the purpose of the design

The very basic step for corpora construction is to delineate the significance of a corpus; otherwise, what is the point of proposing and designing a spoken corpus which might require substantial effort, be time consuming (Baker, 2010), and above all challenging? As mentioned earlier, some sociolinguists are still unsure about the connections between – and the usefulness of – corpus linguistics, corpora, and sociolinguistics; they fear the reliability of corpora, which is greatly reflected in the validity and credibility of claims arising from sociolinguistic investigations, and might have motivated such a reaction. Meanwhile, research questions of DMs have traditionally been investigated through compiling a traditional and 'unconventional' sociolinguistic corpus. However, as research interests and questions of Arabic DMs are constantly developing (e.g., questions of language variation and change), it was realized how advantageous the availability and accessibility of corpora could be for sociolinguistic questions related to the interaction of multiple social variables and the influence of those social variables on linguistic variants such as DMs.

The motivation for writing this paper and proposing a sociolinguistic corpus of spoken Saudi Arabic comes from the fact that there is no single Saudi Arabic corpus developed to meet sociolinguistic inquiries at the time of writing this paper. In order not to make any wild judgments and to provide sensible conclusions, first, the literature of Arabic scholarly works available on sociolinguistics were reviewed, along with corpus linguistics, discourse variation, and finally DMs in particular. Also, the available Arabic corpora were visited and evaluated. Several friends were also consulted aiming to find any corpora that met sociolinguistics inquiries. However, the efforts were in vain. What was highlighted was rather the complete lack of Saudi Arabic sociolinguistic investigations which utilize spoken corpora and corpus methods; this is not surprising given the lack of a single representative sociolinguistic corpus of spoken Saudi Arabic. In other words, this lack could also be due to the fact that spoken language is the hallmark of sociolinguistic research (Kendall, 2011), and what is available in Saudi Arabic comes predominantly from written rather than spoken resources. Note that also after surveying several of the available Saudi Arabic corpora, the following facts were noted: (1) the slight increase in Arabic learners' corpora, (2) the available Saudi Arabic corpora are exclusively written, (3) the materials in those corpora come only from a few written resources ignoring the significance of genre diversity as well as period diversity, (4) the materials were derived from selected newspapers, classic and current literary works that might be representative of very few varieties of Arabic (e.g., Egyptian, Kuwaiti, Classical Arabic, and Modern Standard Arabic), (5) the static nature (Davies, 2009) of most of these corpora in which no more texts and materials have been added after they were assembled, and (6) the poor interface features (Davies, 2009; 2010; 2013).

The advantages observed through the sociolinguistic investigations of DMs utilizing corpora and corpus linguistics methods (e.g., Andersen, 2001; Barbieri, 2008; Erman 2001; Levey, 2006; Miller 2009; Pichlar, 2009; Stubbe & Holmes, 1995; Torgesrsen et al., 2011; Tottie, 2011) have motivated this paper. These advantages consist of (1) the size of corpora, (2) the ease of comparability in terms of real-time and apparent-time studies of language change, across different genres and registers, and a complex variety of social factors, e.g., age groups, gender cohorts, societal classes, and others (Gregersen & Barner-Rasmussen, 2011; Kendall & Herk, 2011), (3) the theoretical, methodological, and practical and applied implications that come out of the sociolinguistic investigations, for example, of DMs, (4) the purpose of opening up numerous avenues for future research (Rühlemann & O'Donnell, 2012), (5) the provision of publicly available or relatively restricted accessible research materials for scholars for many years to come (Childs et al., 2011), (6) the support of researchers who contribute to our understanding of language variation and change from distinct speech communities (Childs et al., 2011), and (7) the search for new ethical challenges caused by linking sociolinguistics with corpus linguistics (Childs et al., 2011; King et al., 2011; Pope & Davis, 2011).

Note that the benefits of sociolinguistic research using spoken corpora and corpus linguistics methods are not only restricted to research projects and linguistic conferences. Rather, as King et al. (2011, p. 49) put it, "[we] intend to continue to find venues and ways to make our findings more widely known at the community level." According to Kendall (2011), most Maori and indigenous language teachers find the MAONZE project "helpful in validating their informal observations and confirming that an emphasis on pronunciation in language learning should not be just at the beginners' stages" (p. 49). Subsequently, the MAONZE project and its research team enabled Maori and indigenous language teachers to develop a pronunciation tool that helped bridge the differential generational gap between Maori older native speakers and Maori younger speakers whom were considered second language speakers (Kendall, 2011).

Up till this point of the paper, readers should have recognized the recurrent emphasis on proposing a *spoken corpus* of Saudi Arabic, but not a spoken and written one (or even only written). This decision might provoke the following question: why spoken rather than written?

It is important to clarify here that it is not intended to say that written corpora cannot aid sociolinguistic queries. Baker (2010) and Säily (2011) have shown how written corpora can be well suited for sociolinguistic analysis. Rather, the decision of proposing and designing a sociolinguistic corpus is similar to Davies' when he wrote exclusively about spoken corpora (Davies, 2013) in the seminal sociolinguistic volume edited by Mallinson et al. (2013). Although, given Davies' experience with the Corpus of Contemporary American English (COCA), he can write about both spoken and written corpora. Historically speaking, the traditions of sociolinguistic research (Eckert, 2012) have overwhelmingly rested on spoken unconventional corpora through which sociolinguists trace non-standard varieties or the vernacular (Kendall, 2011). Another reason for the repeated emphasis on proposing and designing a sociolinguistic corpus of spoken Saudi Arabic comes from the goals and the focus for which such a corpus will be built— DMs and the sociolinguistics of DMs. DMs are a typical and distinctive feature of spoken rather than written language (Baker, 2010); and for the sociolinguistic investigations of DMs to be conducted in Saudi Arabic, a sociolinguistic corpus of spoken Saudi Arabic needs to be carefully mapped out and designed. Therefore, the following subsection considers sources, collection and structures of data. This is an important step for mapping out and constructing the sociolinguistic corpus of spoken Saudi Arabic.

4.2 Sources, collection and structures of data

One of the most important initial steps for corpora construction is looking for sources of materials that fit the aims a corpus is built for. This section offers some sources of spoken materials that help in building the proposed sociolinguistic corpus of spoken Saudi Arabic. For ethical and practical considerations related to data collection (Childs et al., 2011; King et al., 2011), it is advisable to view the work in this section as several stages of data collection in which every stage features procedures that could systematically initiate the sociolinguistic corpus of spoken Saudi Arabic.

The first stage of data collection will focus on what is publicly available and easily accessible from the existing spoken resources. For instance, sources of spoken language of unscripted speech on television and radio programs represent one possibility (Davies, 2009; 2010; 2013; Holmes, 1988; Stubbe and Holmes, 1995). Speech that is publicly available online, for example on YouTube, represent another possibility (Davies, 2009; 2010; 2013; Holmes, 1988; Stubbe and Holmes, 1995). Note that a good number of these programs have freely downloadable transcripts or transcripts available on CDs/DVDs with their broadcast episodes. It is very important to clarify that these transcripts are available in Arabic standard orthography and this raises issues related to transcription protocols. These issues are discussed in the section designated for transcription protocols. According to Davies (2013, p. 211), there are two limitations with such transcripts concerning "the naturalness of the language" and "the difficulty in coding them for demographic information, e.g., age, gender, ethnicity, or socioeconomic status." Despite these potential limitations, these programs and their transcripts still show colloquial linguistic features such as DMs *like* and *you know*, in other words, typical features of everyday conversational settings. As such, the use of DMs in the spoken language drawn from television and radio programs might motivate sociolinguistic questions related to the frequency and pragmatic functions of DMs across different forms of mediated language and/or sociolinguistic questions across different genres of spoken language.

King et al. (2011) added that the archives of national and local radio and television stations could be accessible; these could serve as good sources of spoken language. Spoken language from such resources can crystallize interesting sociolinguistic questions in terms of language variation and change in which recordings for "historical speakers" might be obtained and compared with the speakers of younger generations (King et al., 2011). Nevertheless, recordings extracted from television and radio archives usually face the challenge of finding good quality sound, suitable for transcription and sociolinguistic analysis, since these recordings use broadcast conventions, and as such they mostly have background music (King et al., 2011). Table 1 shows a sample of initial data incorporated from Saudi TV shows in the sociolinguistic corpus of spoken Saudi Arabic.

Table 1: A Sample of Saudi TV Shows in the Corpus

TV show	Episode	Number of participants including host (gender)	Age range	Show category	Topics discussed	Publication date	Length Min: Second	Broadcast channel
Althaminah	1	6 (2 females; 4 males)	13-18	Talk show	Saudi young authors	4/1/2015	45:06	MBC1
	2	4 (1 male; 3 females)	> 30*	Talk show	Saudi southern food	1/25/2016	42:30	
Alusbuʕ fi Saʕah	3	5 (4 males; 1 female)	> 30	Talk show	Topics of the week	3/12/2016	46:26	Rotana Khalijiah
	4	5 (4 males; 1 female)	> 30	Talk show	Topics of the week	10/29/2016	48:19	
Alkura Tatakallam	5	3 (males)	> 30	Sports show	Saudi soccer topics	4/6/2016	54:22	Sports 24
	6	3 (males)	> 30	Sports show	Saudi soccer topics	11/20/2016	55:57	
Almustashar Attaʕlimi	7	2 (2 males); one caller (male)	> 30	Education/ academic show	Academic success	12/5/2015	59:55	iEN TV
	8	2 (2 males); 3 callers (males)	> 30	Education/ academic show	Saudi studying abroad	2/14/2016	49:50	
Fatawa	9	2 (males) and several callers (males and females)	> 30	Religious/ social show	Religious matters	11/30/2015	55:20	Saudi TV1
	10	2 (males) and several callers (males and females	> 30	Religious/ social show	Religious matters	6/19/2016	45:24	

*** > 30 refers to age range more than 30 years old.**

The second stage of data collection will focus on eliciting spoken language collected in more natural and informal settings. That is, language informed by the sociolinguistic treatments of DMs, and the sociolinguistic investigations intersected with corpus linguistics. The research team will seek naturalistic and authentic speech through sociolinguistic fieldwork and sociolinguistic data collection methods such as sociolinguistic interviews. Speaking from an insider Arabic cultural point of view and from a scholarly point of view, there is a remarkable proliferation of heroic and nostalgic stories (Eckert, 2008) about the old life in the desert. There is also a noticeable proliferation of heroic and nostalgic stories about the life of work such as military life. Such types of narratives, if systematically prompted, elicited, and documented, would benefit sociolinguistic investigations in general, and the field of narrative analysis in particular. Additionally, oral narratives of personal experience, a feature of everyday conversational settings, can be highly useful for getting naturalistic speech. Casual informal conversations prompted by sociolinguistic interview methods (Barbieri, 2008; Erman, 1992; Holmes, 1986; 1990; Stubbe & Holmes, 1995; Tagliamonte, 2005; Wouk, 1999) focusing on informal topics such as “school activities, hobbies, sports, friends, and lots of commiseration about problems with parents,” (Tagliamonte, 2005, p. 1899) and husbands and wives, for instance, serve as good sources for building a corpus that meets sociolinguistic queries.

The collection of spoken language using sociolinguistic field methods brings challenges related to speakers’ access and recruitment. For that, King et al. (2011) and Levon (2013) suggested the use of the sociolinguistic ‘friend of a friend’ technique (Milroy, 1987). Thus, the research team of this project will make initial contact with their friends and acquaintances who could help recruit participants to take part in this project. Furthermore, the *snowball sampling* (Goodman, 1961; Levon, 2013) technique, the most commonly used method in social sciences, will be appropriate in many of the Saudi contexts that value social relationships and social networks. Data elicitation will also include different age groups such as subjects drawn from elderly and younger generations. Table 2 presents only a sample distribution of the (expected) elicited spoken data in the corpus.

Table 2: A Sample Distribution of Elicited Spoken Data

Saudi Region	Gender	Age range	Number of participants	Collection method	Length
Asir	5 males	20 – 40	20	Sociolinguistic interviews	30 minutes – one hour
	5 females	20 – 40			
	5 males	40 – 60			
	5 females	40 – 60			
Najran	5 males	20 – 40	20		
	5 females	20 – 40			
	5 males	40 – 60			
	5 females	40 – 60			
Riyadh	5 males	20 – 40	20		
	5 females	20 – 40			
	5 males	40 – 60			
	5 females	40 – 60			

The readers should note that Table 2 is only a sample of the suggested elicitation of spoken data. Whenever possible and accessible, elicitation of the spoken data is planned to incorporate the 13 Saudi regions (Riyadh, Makkah, Madinah, Jazan, Najran, Asir, Al-Bahah, Tabuk, Eastern Region, Qassim, Hail, Jouf, and Northern Borders). The data will be collected from a total of 20 participants in each Saudi region. The distribution of participants includes 5 males and 5 females (age range 20 — 40) and 5 males and 5 females (age range 40 — 60). The sociolinguistic interview is the method of data collection. The length of each sociolinguistic interview is expected to range between 30 minutes and one hour with each participant. The elicitation of spoken data from these 13 Saudi regions is an ongoing and gradual process.

Despite the implementation of the above techniques of in order to secure participants' recruitment, as a result of cultural and religious reasons, there is still a big challenge of recruiting female participants if the research team members are males: how to recruit female participants? This is not to say that recruiting female participants is impossible; it is just to seek and spell out appropriate methods for data collection from female participants. The solution offered for this challenge is to follow the methodological tradition of sociolinguists where they “often try to match interviewers to interviewees in terms of age and gender” (Labov 2000, cited in King et al., 2011, p.43). Therefore, trained female fieldworkers utilizing participants’ recruitment techniques and data collection can easily access Saudi female population and elicit naturalistic speech of importance to sociolinguists. These suggested techniques are expected to be useful in compiling a database and creating a corpus that models representativity and balancedness in the initial stages of the corpus construction.

Transcription protocols and conventions represent the next step after data collection procedures. The next subsection discusses transcription protocols and conventions.

4.3 Transcription protocols

Transcription protocols appeared to be a recurrent topic of great importance in the surveyed studies of DMs and the scholarly treatments of sociolinguistic corpora construction. These studies and scholarly discussions strongly asserted the importance of establishing, developing, and maintaining transcription protocols that aid sociolinguistic analysis. There are several methodological reasons for bringing up the issue of transcription in such a detailed fashion here: first, “the development of a transcription protocol that encourages quick transcription while providing enough structure that all of the transcribers are producing consistent documents has been a negotiation” (Childs et al., 2011, p. 171). Second, uniform transcription conventions and protocols in corpora construction, or any sociolinguistic treatment dealing with discourse-pragmatic variation, have always been a challenge for sociolinguists because of the use of idiosyncratic and differential transcription conventions that influence the overall word counts, raw frequency, and normalized frequency tabulations. Such influences lead, on many occasions, to misleading sociolinguistic results and interpretations (Pichlar, 2010). Such a process of thinking, both hypothetically and practically, greatly helps in establishing the reliability of the proposed corpus.

Consideration of adopting transcription conventions and developing transcription protocols raises the issue of either employing and following standard Arabic spelling and orthography or following the Roman alphabet's style of transcription. This issue is highlighted here because there is a fairly good number of transcripts for radio and TV programs available in Arabic standard orthography. The serious question is: what style of transcription should be adopted? With technological advancements and computing features, however, there could be two possible solutions: (1) transcribers can follow a standard Arabic spelling and orthography accompanied by the use of Arabic transliteration conventions, while using and keeping the standardized transcription codes and conventions for paralinguistic features such as pauses and intonation contours with the Arabic transliteration model, or (2) transcribers can follow an Arabic transliteration model accompanied by the use of standard Arabic spelling and orthography, while using and keeping the standardized transcription codes and conventions with the standard Arabic orthography.

The solution to the above problem should become clear by answering the question: what style of transcription will be most useful and effective? Determining the usefulness and effectiveness of one of the suggested transcription models over the other should involve considering many factors of great importance to corpora, including, but not limited to: (1) the envisioned audience of the proposed corpus, (2) the long-term availability and accessibility of the proposed corpus, (3) adaptations of the proposed corpus for use with concordance programs (Childs, et al., 2011) such as *AntConc* (Baker, 2010; Barbieri, 2008) or Praat and Microsoft applications and spreadsheets (King et al., 2011), (4) the availability or otherwise of the original sound files, (5) the issue of readability, (6) the issues of speed and standardization (Childs et al., 2011, p. 171), (7) dealing with local, and sometimes, subtle linguistic features (Childs et al., 2011, p. 171) such as "*weredn't* 'was/were not', *luh* 'look (you know)', *wuh* 'what', *b'y* 'boy', *yeer* 'your', and *hees* 'his' in the Newfoundland English in the Petty Harbor project, (8) consistency with the available shared corpora and/or linguistic literature, (9) convenience and conformity with the technological and computerized features, (10) taking into account what Arabic transcripts are already available that use standard Arabic spelling and orthography, (11) familiarity with transcription codes and conventions, (12) the degree of commitment and collaboration among the interested research members and among the interested governmental agencies and academic institutions for launching the proposed corpus, (13) the (expected) sociopolitical and financial pressures that are usually exerted by the interested governmental agencies and academic institutions, (14) the goal of sharing transcripts and making them accessible for different kinds of sociolinguistic analysis, (15) negotiations among members of interested research teams (Childs et al., 2013), (16) how faithful a transcription should be (Keune et al., 2005), and (17) uniformity that enables search and comparison across different variables and different genres or subgenres (Gregersen & Barner-Rasmussen, 2011).

From another angle, the sociolinguistic treatments of DMs (Andersen, 2001; Erman, 1992; Grant, 2010; Holmes, 1986; 1988; 1990; Pichler, 2009) have shown that these linguistic items are grammatically, functionally, and distributionally sensitive to their prosodic environments, contextual patterns, utterance positions, pauses, intonational patterns, false starts, hesitations, overlapping, and combinations with other linguistic elements. For instance, in a series of studies examining the functional distribution of DMs used by men and by women, Holmes (1986; 1988; 1990) has shown how prosodic features (e.g., stress, pauses, and intonation), syntactic positions, and contextual information helped identifying the functions and distribution of *you know, I think,* and *of course* respectively. Within these successive discussions (Holmes, 1986; 1988; 1990), Holmes argued that although ‘form’ is important for analysis, it cannot alone be a sufficient basis for categorization. According to Holmes (1990, p. 185), such features make it “important to devise a methodology to protect against avoidable bias in the [stages of] data analysis.” Therefore, while developing transcription protocols for the proposed corpus, transcribers should take into consideration the maintenance of prosodic features, filled and unfilled pauses, fillers, general extenders, minimal responses, interjections, and false starts adjacent to DMs. It is important to note that the articles that introduce and discuss sociolinguistic corpora construction rarely, if ever, touch on such methodological protocols. Such information was found only in the sociolinguistic treatments of DMs cited in this paper.

Putting in mind the aforementioned challenges, potentials, factors, reasons, and justifications, the transcribers preferred and developed a particular model of transcriptions for the suggested sociolinguistic corpus of spoken Saudi Arabic. After several attempts of testing and trials, transcribers followed the standard Arabic spelling and orthography accompanied by the use of transliteration conventions. The transliteration conventions are adopted from the *Encyclopedia of Arabic Language and Linguistics* (EALL; Versteegh, 2006, p.viii) with minor modifications adapted from the International Phonetic Alphabet (IPA). The table in appendix A illustrates the used transliteration conventions. In addition, transcribers adapted transcription conventions for paralinguistic features (e.g., intonation, pauses, etc.) from Jefferson (2004) with a few additions and modifications. The table in appendix B demonstrates the used transcription conventions. The spoken data in the corpus is presented following this layout:

(a) The first line is a rough Arabic transcription of the spoken data.
(b) The second line is a transliteration of the spoken data with the transcription conventions for paralinguistic features.
(c) For convenient reference, numbers were provided for the original Arabic lines of the spoken data.

The following is an illustrative sample for the model of transcription protocols and conventions used in the proposed sociolinguistic corpus of spoken Saudi Arabic:

Illustrative sample

1 AH: الزيادة وينها، وين الزيادة هههه

zziyadah wiinha, wiin zziyadah hhhh=

2 NA: يعني جيب بيبسي شاورما يعني

=***y↑aʕni↓*** jiib bebsi šawarma ***yaʕni***.=

3 AH: هههه طيب مشاهدينا الكرام

=hhh **ṭayyib** mušahdina lkiram...

In the above illustrative sample, the readers can observe examples of DMs used in Saudi Arabic. There are two instances of the DM ***yaʕni*** (يعني) in line 2, one instance is turn-initial and another token is turn-final. There is also one instance of the DM **ṭayyib** (طيب) in line 3. **Bold** and *italics* indicate the instances of Saudi Arabic DMs in the corpus.

The reason for suggesting, developing, and preferring such a model of transcription protocols and conventions (among others) is that it is the most useful and most effective one given the 17 factors and aspects mentioned and discussed previously. The transcription and coding processes should also consider metadata. This is another important step for motivating sociolinguistic queries of DMs.

4.4 Metadata

The review of the sociolinguistic treatments of DMs and the consultation of corpus sociolinguistic articles have highlighted the significance of providing metadata of the spoken collected materials included in the proposed corpus as well as metadata about the participants contributing to the corpus. Importantly, these scholarly treatments have helped uncover the kinds of meta-information necessary for sound sociolinguistic research. These treatments have shown that adding the appropriate metadata and including that as an input in a corpus will stimulate new routes for sociolinguistic research using corpora and corpus methods. Table 3 in the following gives an overview of the expected metadata included in the corpus.

Table 3: A Summary of the Metadata Included in the Corpus

Participant metadata	Text (transcript/spoken) metadata
1. Age 2. Gender 3. Geographic location (origin) 4. Number of years in the geographic origin 5. Other geographic locations 6. Number of years in other geographic locations 7. Language contact 8. Level of education 9. Number of languages spoken 10. Ethnicity 11. Tribal affiliation	1. Speech genre (e.g. narrative; casual conversation; interview) 2. Place of recording 3. Time of recording 4. Length of recording 5. Number of words 6. Number of participants recorded 7. Same sex or mixed sex interaction 8. Medium of collection

Note that table 3 elucidates the canonical social variables of importance to sociolinguists and sociolinguistic investigations: age, gender, geography, social class, language contact, and ethnicity. Other social factors were also added. Such social factors might have significant impact on the use of language in Saudi Arabic contexts, specifically the stratified use of DMs for instance tribal affiliation as a social factor in Saudi Arabic. Tribal affiliation can be used as an indication of dialect areas in which dialect variation is expected to appear in the use of DMs. It is important to note that, linguistically speaking, dialect variation can be minimal between adjacent areas and maximal between relatively distant and distant areas. By glancing at Table 3, one might think of including lots of information. Reliability of corpora and the intention to build a corpus that stimulates a wide variety of sociolinguistic questions of DMs, for example, are reasons to include such detailed and subtle metadata.

In the following, Table 4 provides some examples of the suggested metadata scheme for the context of the proposed sociolinguistic corpus of spoken Saudi Arabic.

Table 4: An Example of the Suggested Metadata Scheme for the Corpus

Alkura Tatakallam (Saudi Sports TV show)								
Speaker	**Role**	**Age**	**Gender/ sex**	**Work**	**Specific Discussed Topics**	**Publication date**	**Length Min:Second**	**Broadcast channel**
Ahmed (AH)	Host	37	Male	Journalism and media				
Episode 1 (KEP1)						4/6/2016	54:22	Sports 24
Khalid (KH)	Guest	43	Male	Journalist, lawyer, consultant	- Elections for SAFF*; Candidates for the elections; - Derby between Al-Nassr and Al-Ittihad			
Fahad (FA)	Guest	49	Male	Business, sports analyst,				
Episode 2 (KEP2)						11/20/2016	55:57	Sports 24
Sultan (SU)	Guest	43	Male	Team manager and a previous football coach	Mutual responsibilities and obligations between players and their teams			
Nawaf (NA)	Guest	35	Male	Lawyer, sports business agent				

* Saudi Arabian Football Federation

Table 4 presents the available and accessible participants' metadata including speakers' identities, age, gender, role of participants, and work. Note that the metadata of *work* might have some implications for the level of education and the socioeconomic status of the speakers. Table 4 also offers the available and accessible texts' metadata including speech genre (spoken Saudi sports TV show), publication date, length of episodes, place of recording (broadcast channel Sports 24), medium of collection (broadcast channel Sports 24), same sex interactions (male participants), the number of participants (3 in each episode), and the specific discussed topics.

5. Conclusion

In conclusion, no one single paper can encapsulate all the guiding principles and the important details for building a sociolinguistic corpus of spoken Saudi Arabic that can support sociolinguistic investigations.

Rather, such principles will be of great interest for subsequent papers. Although this paper is an attempt to propose a sociolinguistic corpus of spoken Saudi Arabic that fulfills sociolinguistic queries, this attempt might be a starting point for stimulating sociolinguistics experts to build and launch connected corpora based on other varieties of Arabic similar to those of the International English Corpora (ICE) where different varieties of English are represented such as Australian English and New Zealand English (Miller, 2009) and South African English (Huddlestone & Fairhust, 2013).

Building and launching an Arabic spoken corpus will not only be of value for sociolinguistics; it will also be of great value to many different linguistic fields such as discourse analysis, pragmatics, historical linguistics, syntax, and morphology. It will be clear that potential benefits of building and launching an Arabic spoken corpus will be worth the extra work and risk. Such a corpus is expected to be usable by many people including researchers, educators, and even members of interested Arabic communities. For the utility of this corpus to succeed, it should not be limited to certain social variables; rather, it is crucial for this corpus to incorporate diverse Saudi Arabic varieties, spoken genres, and speakers, to name a few. Finally, it is hoped that this proposal finds encouragement, support, and collaboration from the Saudi research community as well as Saudi communities themselves, and the people who share an interest in Saudi Arabic linguistic studies in general, and Arabic sociolinguistics studies in particular. It is also hoped that this proposal motivates people who care about discovering new directions and research ideas/projects that illuminate our understanding of various linguistic phenomena, methodologies, and/or theories.

* * *

6. References:

(1) Andersen, G. (2001). *Pragmatic Markers and Sociolinguistic Variation. A Relevance- Theoretic Approach to the Language of Adolescents*. Amsterdam/ Philadelphia: John Benjamins.

(2) Aijmer, K., & Simon-Vandenbergen, A. M. (2011). Pragmatic markers. In J. Zienkowski, J. Ostman, & J. Verschueren (Eds.), *Discursive pragmatics*, *8*, (pp. 223-247). Philadelphia: John Benjamins.

(3) Babel, A. M. (2014). Stereotypes versus experience: Indexing regional identity in Bolivian Valley Spanish. *Journal of Sociolinguistics*, *18*(5), 604-633.

(4) Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh University Press.

(5) Barbieri, F. (2008). Patterns of age-based linguistic variation in American English. *Journal of sociolinguistics*, *12*(1), 58-88.

(6) Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275-311.

(7) Childs, B., Van Herk, G., & Thorburn, J. (2011). Safe harbour: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory*, *7*(1), 163-180.

(8) Dailey-O’Cain, J. (2000). The sociolinguistic distribution of and attitudes toward focuser like and quotative like. Journal of Sociolinguistics 4, 60–80.

(9) Davies, M. (2009). The 385+ million-word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*(2), 159-190.

(10) Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, 18(4), 447- 464.

(11) Davies, M. (2013). Establishing corpora from existing data sources. In C. Mallinson, B. Childs, & G. Herk (Eds.), *Data collection in sociolinguistics* (pp. 210-212). New York: Routledge.

(12) Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453-476.

(13) Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, *41*, 87-100.

(14) Erman, B. (1992). Female and male usage of pragmatic expressions in same-sex and mixed-sex interaction. *Language Variation and Change*, *4*(02), 217-234.

(15) Erman, B. (2001). Pragmatic markers revisited with a focus on you know in adult and adolescent talk. *Journal of Pragmatics*, *33*(9), 1337-1359.

(16) Fraser, B. (1999). What are discourse markers?. *Journal of Pragmatics*, *31*(7), 931-952.

(17) Fraser, B. (2009). Topic Orientation Markers. *Journal of Pragmatics*, 41, pp. 892-898.

(18) Fuller, J. M. (2003). Use of the discourse marker like in interviews. *Journal of Sociolinguistics*, *7*(3), 365-377.

(19) Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics,* 32, 148-170.

(20) González, M. (2004). *Pragmatic markers in oral narrative: The case of English and Catalan* (Vol. 122). John Benjamins Publishing.

(21) Grant, L. E. (2010). A corpus comparison of the use of I don't know by British and New Zealand speakers. *Journal of Pragmatics*, *42*(8), 2282-2296.

(22) Gregersen, F., & Barner-Rasmussen, M. (2011). The Logic of comparability: On genres and phonetic variation in a project on language change in real time.*Corpus Linguistics and Linguistic Theory*, *7*(1), 7-36.

(23) Holmes, J.(1986). Functions of you know in women's and men's speech. *Language in Society,* 15, 1–22.

(24) Holmes, J. (1988). Of course: A pragmatic particle in New Zealand women's and men's speech. *Australian Journal of Linguistics*, *8*(1), 49-74.

(25) Holmes, J. (1990). Hedges and boosters in women's and men's speech. *Language and Communication*, 10, 185–205.

(26) Huddlestone, K., & Fairhurst, M. (2013). The pragmatic markers anyway, okay, and shame: A South African English corpus study. *Stellenbosch Papers in Linguistics Plus*, *42*, 93-110.

(27) Huspek, M.(1989). Linguistic variability and power: An analysis of YOU KNOW/I THINK variation in working-class speech. *Journal of Pragmatics*, 13, 661–683.

(28) Innes, B. (2010). "Well, that's why I asked the question sir": Well as a discourse marker in court. *Language in Society*, *39*(01), 95-117.

(29) Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp.13-23). Philadelphia: John Benjamins.

(30) Kendall, T. (2011). Corpora from a sociolinguistic perspective. *Revista Brasileira de Linguística Aplicada*, *11*(2), 361-389.

(31) Kendall, T., & Van Herk, G. (2011). Corpus linguistics and sociolinguistic inquiry: Introduction to special issue. *Corpus Linguistics and Linguistic Theory*, *7*(1), 1-6.

(32) Keune, K., Ernestus, M., Hout, R. V., & Baayen, R. H. (2005). Variation in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory*, *1*(2), 183-223.

(33) King, J., Maclagan, M., Harlow, R., Keegan, P., & Watson, C. (2011). The MAONZE project: Changing uses of an indigenous language database. *Corpus Linguistics and Linguistic Theory*, *7*(1), 37-57.

(34) Labov, William. 2000. *Principles of linguistic change. Volume II: Social factors.* Oxford: Blackwell.

(35) Lenk, U. (1998). Discourse markers and global coherence in conversation. *Journal of Pragmatics*, 30, 245-257.

(36) Levey, S. (2006). The sociolinguistic distribution of discourse marker like in preadolescent speech. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*, *25*(4), 413-441.

(37) Levon, E. (2013). Ethnographic fieldwork. In C. Mallinson, B. Childs, & G. Herk (Eds.), *Data collection in sociolinguistics* (pp. 69-79). New York: Routledge.

(38) Macaulay, R. (2002). You know, it depends. *Journal of Pragmatics*, *34*(6), 749-767.

(39) Mallinson, C., Childs, B., Van Herk, G. (2013). *Data collection in sociolinguistics: Methods and applications*. New York: Routledge, Taylor & Francis Group.

(40) Meyerhoff, M. (1994). Sounds pretty ethnic, eh?: A pragmatic particle in New Zealand English. *Language in Society*, *23*, 367-367.

(41) Miller, J. (2009). Like and other discourse markers. In P. Peters, P. Collins, & A. Smith (Eds), *Comparative studies in Australian and New Zealand English: Grammar and beyond* (pp. 317-337). Amsterdam: John Benjamins.

(42) Milroy, L. (1987). *Observing and analyzing natural language*. Oxford: Blackwell.

(43) Muller, S. (2005). *Discourse Markers in Native and Non-native English Discourse*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

(44) Pichler, H. (2009). The functional and social reality of discourse variants in a northern English dialect: I DON'T KNOW and I DON'T THINK compared. *Intercultural Pragmatics*, *6*(4), 561-596.

(45) Pichler, H. (2010). Methods in discourse variation analysis: Reflections on the way forward. *Journal of Sociolinguistics*, *14*(5), 581-608.

(46) Pope, C., & Davis, B. H. (2011). Finding a balance: The Carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, *7*(1), 143-161.

(47) Rühlemann, C., & O'Donnell, M. B. (2012). Introducing a corpus of conversational stories. Construction and annotation of the Narrative Corpus. *Corpus Linguistics and Linguistic Theory,* 8(2), 313-350.

(48) Säily, T. (2011). Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, *7*(1), 119-141.

(49) Sankoff, G., Thiboult, P., Nagy, N., Blondeau, H., Fonollosa, M.-O., & Gagnon, L. (1997). Variation in the use of discourse markers in a language contact situation. *Language and Change*, *9,* 191-217.

(50) Schiffrin, D. (1987). *Discourse Markers*. Cambridge: Cambridge University Press.

(51) Schiffrin, D. (2001). Discourse markers: Language, meaning, and context. In D. Schiffrin, D.Tannen, & H. Hamilton (Eds.), *The handbook of discourse analysis*, (pp. 54-75). John Wiley & Sons.

(52) Schourup, L. (1999). Discourse markers. *Lingua*, 107, 227-265.

(53) Schourup, L. (2011).The discourse marker now: a relevance theoretic approach. *Journal of Pragmatics*, 43, 2110-2129.

(54) Stubbe, M., & Holmes, J., 1995. You know, eh and other ‘exasperating expressions’: an analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English. *Language and Communication* 15 (1), 63–88.

(55) Tagliamonte, Sally, 2005. So who? Like how? Just what? Discourse markers in the conversations of young Canadians. *Journal of Pragmatics* 37, 1896–1915.

(56) Torgersen, E. N., Gabrielatos, C., Hoffmann, S., & Fox, S. (2011). A corpus-based study of pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory*, *7*(1), 93-118.

(57) Tottie, G. (2011). Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, *16*(2), 173-197.

(58) Versteegh, K. (Ed.). (2006). *Encyclopedia of Arabic Language and Linguistics* (Volume 1). Leiden: Brill.

(59) Wouk, F. (1999). Gender and the use of pragmatic particles in Indonesian. *Journal of Sociolinguistics*, *3*(2), 194-219.

* * *

Appendix A
TRANSLITRATION CONVENTIONS

The transliteration conventions used in the sociolinguistic corpus of spoken Saudi Arabic are adopted from the *Encyclopedia of Arabic Language and Linguistics* (EALL; Versteegh, 2006, p.viii) with minor modifications adapted from the International Phonetic Alphabet (IPA). The following table illustrates the used transliteration conventions.

The Arabic sound/letter	The transcription symbol	Examples
همزة ،ء/glottal stop	ʔ	**ʔ**šya ‘things’
ب	b	**b**ukrah ‘tomorrow’
ت	t	**t**aʕliim ‘education’
ث	th	mi**th**al ‘example’
ج	j	**j**amiil ‘beautiful’
ح	ḥ	muḥami ‘lawyer’
خ	x	**x**amsa ‘five’
د	d	**d**uktur ‘doctor’
ذ	ḏ	ʔstaa**ḏ** ‘mister’
ر	r	**r**amz ‘code’
ز	z	fu**z** ‘win’
س	s	ʔ**s**a**s**i ‘basic’
ش	š	ma**š**ruuʕ ‘project’
ص	ṣ	**ṣ**ut ‘vote’
ض/ظ	ẓ	**ẓ**alaam ‘dark’
ط	ṭ	**ṭ**awil ‘tall; long’
ع	ʕ	la**ʕ**ib ‘player’
غ	ġ	**ġ**arib ‘strange’
ف	f	**f**ariq ‘team/club’
ق	q	**q**alb ‘heart’
ك	k	**k**urah ‘ball’
ل	l	raju**l** ‘man’
م	m	naj**m** ‘star’
ن	n	sa**n**ah ‘year’
هـ	h	**h**adaf ‘goal’
و	w	**w**araqa ‘paper’
ي	y	**y**imkin ‘probably’
Shadda	double consonants	di**qq**ah ‘accuracy’
Short vowels	a, i, u	q**a**lb ‘heart’
Long vowels	aa, ii, uu	**ẓ**al**aa**m ‘dark’

* * *

Appendix B
TRANSCRIPTION CONVENTIONS

Transcription conventions were adapted from Jefferson (2004) with a few additions and modifications.

(.)	Pauses
=	Equal sign marks latched talk when used between two contributions of turns by two different speakers. The equal sign also marks a continuation of talk/turn by one speaker
.	Final intonation
,	Continuing intonation
↑	Rise in intonation
↓	Drop in intonation
?	Questioning intonation
Underlining	Emphasis/stress
-	A hyphen indicates an abrupt interruption/abrupt in talk/speech
#	Creaky voice
yaʕni	Bold and italics indicate the analyzed instances of *DM*
ya::ʕni::	Colons mark elongated speech and/or a stretched sound
[ ]	Square brackets mark overlapped speech
.hh	Audible inhalation
hhh	laughter in speech/talk
…	Three dots indicate either previous and/or following deleted talk/speech
(( ))	Transcriber's/researcher's comment

* * *