Functions of the discourse marker So in the LINDSEI-AR corpus
Sami Algouzi

Abstract: This article aims to explore the functions of the discourse marker so in the
LINDSEI-AR corpus. This corpus forms the first part of LINDSEI, representing English
used by Arab learners. The participants in this study were advanced Saudi learners of
English in their third and fourth years of undergraduate study. There are five discourse
marker functions of so in the learner corpus, occurring with different frequencies. The
most frequent use is so indicating result. The second most frequent function is so
introducing a new sequence. Then, comes so marking the main idea unit. The fourth
most frequent function is so marking transition. The least frequent function is so
introducing a turn. Based on these results, it is possible to conclude that the analyses
illustrate that the existing functions of the so found in the literature can be confirmed
in the LINDSEI-AR and that Saudi English language learners, both male and female,
could use the functions of so appropriately. With respect to gender, the results show
no statistical significance between Saudi males and females.

Subjects: Language & Linguistics; Applied Linguistics; Corpus Linguistics; Pragmatics

Keywords: Discourse markers; LINDSEI-AR; Saudi learner; English as a foreign language; So

1. Introduction
Discourse markers (henceforth DMs) have been one of the most controversial topics in linguistics
research. Studies have not only investigated the types of DMs most frequently used but have also
extended to the exploration of how speakers use them and what discursive functions they carry
(Schourup, 1999).

Discourse markers were defined by Schiffrin (1987: 31) as dependent elements functioning to
bracket units of talk. Fraser (1996) argued that DMs can be perceived as linguistic expressions that

ABOUT THE AUTHOR
Dr Sami Algouzi holds a PhD degree in Linguistics from Salford University – England. He is an assis-
tant professor of linguistics at the Department of English, College of Language and Translation at
Najran University. His main research interests include discourse analysis, corpus linguistics and pragmatics.

PUBLIC INTEREST STATEMENT
English Language learners have many aims in learning a language. One of these aims is to
speak English appropriately. Discourse markers are important as one of the common features of
spoken English and indeed very many languages.
So, this study is significant because it investigates discourse markers in the spoken English.
The results are useful because they are expected to provide an insight into how English language
learners use discourse markers. The analysis of the functions of discourse markers is expected to
raise awareness among language learners and consequently contributes to better usage.
signal the relationship between the utterance they introduce and the foregoing utterance. It must be acknowledged that there are several definitions and labels for DMs, such as discourse connectives, discourse operators (Redeker, 1990) and discourse particles (Hansen, 1997). This article employs the term “discourse marker” in line with the aforementioned definitions of Schiffrin (1987) and Fraser (1996) and particularly due to its popularity in most prior research (Schourup, 1999).

Regarding the characteristics of DMs, Schiffrin (1987: 31) pointed out that the removal of a DM from a sentence leaves it structurally intact and they can occur freely within a sentence, i.e. in initial, medial and final position. In addition to these, Brinton (1996: 33–35) added several other significant characteristics of DMs: they are used predominantly in oral discourse, form an independent tone, are difficult to relate to any grammatical word class and are multifunctional.

Hölker (1991) conducted an interesting study that discussed the features of DMs from both a semantic and pragmatic perspective. The study reported that DMs have no effect semantically on the truth conditions of an utterance and at the same time add nothing to the propositional content. Pragmatically, he argued that DMs are connected to the speech and utterance rather than the topic discussed. He delineated the functions of DMs as emotive and expressive, not referential, denotative, or cognitive. However, I argue against this notion that DMs have no referential function. One of the key functions of so, for example, is to mark the relationship between an utterance and the preceding one (Fung & Carter, 2007).

Having discussed the definition of DMs and their major characteristics, I turn to their importance in speech. DMs have always been perceived to comprise an important feature in native speakers’ discourse (Hellermann & Vergun, 2007). Regarding the effect of the use of DMs in speech, it is believed that it makes speech friendly and sociable, whereas DM-free speech might sound strange (Weydt, 2006: 208). It has been argued that a high frequency of DM use is one of the signs of a fluent speaker (Sankoff et al., 1997). However, it seems rather that there is an acceptable frequency of usage of DMs in speech. Indeed, it has been acknowledged that a scarcity or underuse of DMs makes speech sound unnatural or non-native-like, but so does overuse (Müller, 2005: 13; Siepmann, 2005: 245). According to Müller (2005), repeated use of DMs can be experienced as “irritating” and this applies to native speakers and non-native speakers alike. This leaves a crucial question unanswered in terms of who decides what might be acceptable and what is not acceptable in terms of frequency of DM use. There should perhaps be a common sense of what constitutes the normal frequency of DM use, but I argue that the acceptable frequency of usage should be judged by the hearer: what sounds irritating to one hearer might be acceptable to another.

Out of many other interesting discourse markers, so was chosen to be investigated because it appeared in the corpus frequently with variant functions. Therefore, the focus of this article is going to be on so and how it is found in the LINDSEI-AR corpus. The article consists of five sections. Section 2 looks at how so has been reported in relevant prior research. Section 3 outlines the data and the methodology employed in this study. In section 4, the qualitative and quantitative results for so are presented. Section 5 summarises the findings from the corpus and draws conclusions.

2. Previous studies of “so”

The discourse marker so has drawn the attention of a number of researchers (e.g., Aijmer, 2002; Blakemore, 1998; Bolden, 2008, 2009; Fraser, 1990, 1999, 2006; Müller, 2005; Östman, 1981; Redeker, 1990; Schourup, 1985). Van Dijk (1979) conducted one of the very first studies on this particular marker. He observed that so “links two speech acts of which the second functions as a ‘conclusion’ with respect to the first speech act” (1979: 453). Schiffrin (1987) conducted a detailed study of because (with a causal meaning) and so (with a result meaning). She contended that both “have semantic meanings realised on sentence and discourse levels” (1987: 201). Similar to Schiffrin’s treatment of Redeker (1990, 1991) noted that it was one of the connectives most used in her data, in addition to and. She observed that so was found to occur “between successive elements” of events and to indicate “summing-up or conclusion” (1990: 373).
Following the discussion of transitional, or what he calls “stand-alone” so, Raymond (2004: 185) argues that participants use so in conversations “to prompt action by a recipient”. In his study, the analysis of so indicated that it helps invite hearers to establish connections between the preceding turn and the course(s) of action in which they participate. He demonstrated that the stand-alone so is deployed to manage overlapping contingencies imposed by the organization of turn-taking, sequencing and the overall structural organization of conversation as a unit (Raymond, 2004: 212).

Some previous studies have compared the use of so by native speakers of English to that by non-native speakers. Anping (2000) examined the use and frequency of so in Chinese EFL learners’ written English. As expected, Anping discovered that so was applied more frequently by Chinese learners than by native speakers of English. Anping (2000) referred the overuse by Chinese learners to the stylistic differences between written and spoken English, the limited exposure of the Chinese learners to English and the mode of instruction at school and negative L1 transfer.

One of the most comprehensive studies of so is that conducted by Müller (2005). Her qualitative and quantitative analysis of the use of so by English native speakers and German non-native speakers suggested that learners of English use the functions of so as native speakers do, but with notably lower frequency (Müller, 2005). She examined the functions of so at two levels, textual and interactional. On the textual level, so is a marker of result or consequence, a marker of a main idea unit, summarizing/rewording/giving an example and a sequential marker. So is also a marker of a boundary between instructions and the beginning of narrative (Müller, 2005: 80–81). On the interactional level, so operates as a discourse marker that indicates a speech act of questioning or request, of opinion and a marker of an implied result (i.e. directly addressing the hearer and challenging her/him to establish what the speaker is implying) (Müller, 2005: 86). The last function of so on the interactional level is as a marker of a transition relevant place. Interestingly, from a qualitative perspective, all nine functions of so were found to occur in the native and the non-native data. However, native speakers of English used so almost twice as often as German non-native speakers.

Aiming to explain the function of so in implementing incipient actions, Bolden (2009) examined a large corpus of recorded conversations collected from the everyday talk. The analysis showed that so was used by speakers to signal an action that is about to happen “emerging from incipiency” (p. 988). Her study showed that so could be implemented to create certain actions to outline “interactional agenda”. This function of so helps to create discourse coherence that contributes to attaining understanding (Bolden, 2009, p. 996).

Buysse (2012) investigated the use of so among non-native and native speakers of English. His corpus consisted of informal interviews with 40 undergraduate Belgian native speakers of Dutch, 20 of them majoring in Commercial Sciences and 20 in English Linguistics. His data were compared with native English speaker data. In all, 10 functions of so were identified in the study, pertaining to three domains: ideational, interpersonal and textual. All 10 functions were used by both parties, language learners and native speakers. The results of the study demonstrated that the language learners used so significantly more than the English native speakers and that the students in English Linguistics used so slightly more than those in Commercial Sciences.

In academic advice sessions conducted in an English as a lingua franca context, House (2010) reported that the occurrence of the discourse marker so was extraordinarily frequent. So was identified as performing different functions, e.g., signalling causal and inferential connections between clauses and introducing a new topic. So was also employed as a deictic element that speakers used both to buy time for their upcoming moves and to help them “look backwards”, summing up previous stretches of discourse.

Liu (2017) compared ideational and pragmatic functions of but and so used by native and non-native speakers of English. Data were gathered using individual sociolinguistic interviews. There were five native English speakers and ten L1 Chinese speakers. The results demonstrated that
Chinese speakers of English underused the pragmatic functions of *but* and *so*. The results also confirmed that there was a gap between native and non-native English speakers in communicative competence in the use of *but* and *so*. Liu related her findings to the reason that speakers’ first language and their overall oral proficiency in oral discourse affected their use of *but* and *so*.

To conclude, the review of the literature shows that the discourse marker *so* is chiefly associated with inference or consequence. In addition, the studies demonstrate how non-native speakers of English use it (Anping 2000; Buyssse, 2012; Müller, 2005). This study, therefore, is significant because it aims to investigate the use of the DM so in the speech of Saudi non-native speakers, contributing to the existing literature and research on Saudi speakers of English. It is expected that this study will benefit not only linguists in the field of DM research but also the English teaching community in Saudi Arabia.

3. Data and methodology

This study applies the methodology of corpus linguistics. A corpus is “a collection of texts or parts of texts upon which some general linguistic analysis can be conducted” (Meyer 2009: xi). The LINDSEI-AR sub-corpus, employed in this study, is the official component of the Louvain International Database of Spoken English Interlanguage (LINDSEI) and as such the research design was determined by the general LINDSEI model. This corpus forms the first part of LINDSEI, representing English used by Arab learners. The LINDSEI project was launched in 1995 at the Université Catholique de Louvain (UCL), the result of the collaboration between several international universities. Its aim was to collect oral data produced by advanced learners of English (De Cock, 2004; Gilquin, 2008). The proficiency level of the participants is established based on years of English study at university and the criterion is that all participants be in their third or fourth year of study (De Cock, 2004; Gilquin, 2008; Gilquin et al., 2010).

In principle, following the standard format of LINDSEI gives LINDSEI-AR the advantage of using data from native speakers of English, LOCNEC being a reference corpus. LOCNEC was compiled by De Cock (2004) and had the same structure as LINDSEI. The native speakers in LOCNEC were British students majoring in English language and/or linguistics programmes at the University of Lancaster. The interviews were administered by members of the teaching staff according to the format of LINDSEI (Gilquin et al., 2010). The interviews comprised three tasks: set topic, free discussion and picture description. Each interview lasted approximately 15 minutes and followed the same pattern. The interviews began with the set topic. The interviewee was asked to talk about an experience that had affected him/her, a country that he/she had visited and liked/disliked, a film or a movie that he/she had watched and liked and why, etc. This led to a free and engaging discussion with the interviewer. Finally, every interview concluded with a short picture-based story-telling activity.

The LINDSEI-AR sub-corpus comprises 55 informally recorded interviews in English between Saudi students taking English as a foreign language. All 55 recordings were transcribed and marked up according to the LINDSEI protocols (see Appendix 1). The participants were all third- and fourth-year university students (30 males and 25 females, aged 20–25 years) who were studying English language at different universities and colleges. The total number of words used by the interviewers and interviewees collectively was 103,170, of which 77,075 were from the interviewees. The results for the frequency of the discourse marker *so* were normalized, i.e. standardized according to a consistent text length of 10,000 words. Based on this information, I calculated the occurrence of the discourse marker tokens in each speaker’s interview. To obtain the frequency, the number of the tokens was multiplied by 10,000 and divided by the total word count.

Aiming to present a comprehensive account of the use of the discourse marker *so*, the approach followed in this article is mostly corpus-based. The qualitative aspect consists of a system of categorization for the functions of *so*, classifying the tokens according to the function so was found to represent. To differentiate between *so* as a DM and other uses (such as so as an adverb of degree or manner, intensifier, in fixed expression, preform, or as a conjunction to express purpose),
the functions were identified building on characteristics found in previous studies (e.g., Brinton, 1996; Fraser, 1990, 1999, 2006; Schiffrin, 1987). Examples of these characteristics of *so* are as follows: it is typically used in oral discourse rather than in writing; it forms a separate tone unit; it is syntactically optional; it is difficult to place within a traditional grammatical word class; it is multifunctional (Brinton, 1996: 33–35).

4. Results
As a discourse marker, *so* is commonly found to fulfil five functions. The first function is *so* indicating a result. It may also be used to introduce a main idea unit as a recapitulation of something that has been said or discussed earlier. Another function of *so* is to help the speaker introduce successive elements in a chain of events. This function is different from *so* introducing a turn. In the latter, *so* introduces a turn which is new/does not relate to a previous discussion and is mostly turn initial as a discourse opener. As a transition organiser, *so* helps the speaker to give the floor to the listener.

4.1. “So” indicating a result
The use of the discourse marker *so* to report results is one of its most well-established functions (Anping 2000; Blakemore, 1998; Buysse, 2012; Fraser, 1990, 1999; Müller, 2005; Schiffrin, 1987). In line with Müller (2005: 72), this function of *so* is optional not only syntactically but also semantically, as exemplified using an extract from the LINDSEI-AR data:

(1)

```
<A> <starts laughing> it’s a fact it’s a fact <stops laughing> </A>
<B> <starts laughing> yeah it seems so <stops laughing> </B>
<A> okay </A>
<B> but honestly I haven’t seen encountered them face-to-face (erm) a lot so I can’t judge <foreign> yaani <foreign> </B>
```

(AR45)

In this excerpt, AR45 is speaking about his experience when travelling to an Asian country. The interviewer asks him about some aspects of the culture the people of this country are well known for. The interviewee AR45 says that he does not have that much knowledge of the culture because he did not meet with people enough. He uses *so* to indicate the result that he can’t judge. Based on the chronological sequence (Biber et al., 1999) of the cause “I haven’t seen encountered them face-to-face (erm) a lot” and the result “I can’t judge”, the analysis suggests that the use of *so* could well be interpreted as indicating a result. Another similar example of the use of *so* was found in the following lines, when the interviewee, AR16, compared and contrasted New Zealand, Canada and the United States. Here, he uses *so* to express the result that out of these countries, New Zealand is the cheapest and therefore he would go there to learn English.

(2)

```
<A> New Zealand is </A>
<B> (uhu) </B>
<A> is less expensive than (er) Canada and the[i:] US </A>
<B> I think yes some = somebody told me that </B>
<A> right ok </A>
<B> so maybe I will go there </B>
```

(AR16)

4.2. “So” marking a main idea unit
The second category is using *so* as a marker of a main idea unit (Müller, 2005; Schiffrin, 1987). Illustrating how this form of *so* functions, Schiffrin (1987: 195) states that “… stories can be told not
just to report a specific experience, but to warrant a general point being made by the speaker. The entry to such a story can be marked with because, and the return to the main point marked with so”.

In the same vein, Müller (2005) categorizes so as the marker of a main idea unit when the speaker returns to it after a digression. The following example illustrates so as it refers to a main idea unit:

(3)

<B> (er) I now understand that you know the call of prayer is not just to call for prayer it’s you know a call of a a (er) a call to live a call for life. and </B>
<A> nice </A>
<B> a call you know to build to work and to learn </B>
<A> (mhm) </A>
<B> so yeah I’ve always known how to perform prayers of course you know you know I’m a Muslim and stuff </B>

(AR11)

Interviewee AR11 speaks about a book she reads and loves about prayers and the call to prayer. Earlier in the interview, she states that the book does not teach her how to pray, but rather explains the essence of prayers and the call for prayer and what they mean to her. In her last turn, she returns to the main idea that she has always known the steps of prayers because she is Muslim. There are some textual cues that researchers previously considered to strengthen their conclusions (cf. Ajimer 1996). In this example, I argue that one use of yeah after so is to signal a return to the main idea.

4.3. “So” introducing a new sequence

The function of so to introduce a new sequence enables the speaker to accomplish coherence “between successive elements in a chain of events” (Redeker, 1990: 373–374). Dictionaries define this function of so as one that introduces the next event/part in a story (Oxford Advanced Learner’s Dictionary, 2000; Sinclair 1995). Buysse (2012) argues that the sequential so can be distinguished from resultative so as the former starts a new sequence within the turn, whereas the latter brings a sequence or a turn to a close.

The following excerpt illustrates the use of the DM so introducing a new sequence. In Excerpt (4), the interviewee speaks about a novel he has read called Doctor Faustus. He introduces it to the interviewer and explains its main idea. He then moves on to elaborate on the nature of the language used in the novel. To introduce a new sequence flawlessly about some examples of vocabulary from what he calls old language, he uses so.

(4)

<B> but I didn’t try to do that actually the[i:] language was hard and you know it’s an old language </B>
<A> yeah </A>
<B> and we are we do we understand the new English so (er) I try to understand it I don’t like that I used the dictionary too </B>
<A> you should </A>
<B> yeah so some words like thou thee that (er) I don’t </B>
<A> old English yes </A>
<B> yes <XX> yeah I have a good dictionary he it it translates old English of words </B>

(AR48)
4.4. “So” marking transition
In her participation framework, Schiffrin (1987) states that the DM so can function in the organization of transitions. She believes that transitions in participation have two characteristics: the speaker shifts responsibility to the hearer and/or there is a shift from the speaker to hearer. Either of these are intended to fulfil a certain interactional task. Sacks et al. (1974: 697) maintain that “[o]nce a state of talk has been ratified, cues must be available for requesting the floor and giving it up, for informing the speaker as to the stability of the focus of attention he is receiving”. This is one of the functions so achieves in discourse. So can indicate if a speaker is willing to cede the floor, or more directly encourage the addressee to take the floor (Lam, 2010: 670). This function has various labels in the literature, for example, in Müller (2005) it is “marking implied result” and in Buyse (2012) it is a “prompt”. One example of this function of so is illustrated in the following example. Interviewee (27) chooses two movies and speaks about how they unfold. It could be argued that the lengthened so, situated towards the end of the conversation, fulfills its function by handing over the floor to the interviewer. The evidence for this is when the interviewer takes the floor and starts asking a general question about the movies.

(5)

<B> not that man who is that man you are looking at yourself </B>
<A> (mhm) yeah </A>
<B> because he saw himself in a different way </B>
<A> right </A>
<B> so: I felt very interested and the movie was very touching </B>
<A> is this are both American movies </A>

(AR27)

4.5. “So” introducing a turn/discourse
Another function of so is to start a turn. The reason this function comes last in this section is that it appears with the least frequency in the data; a total of two tokens/times in B turns in the LINDSEI-AR.¹ However, so is used to fulfill this function in A turns at least once by almost all interviewers. I argue that this has to do with the nature of the interview, or as Buyse (2012: 1772) puts it, the fact that “the specificity of the environment in which so takes on this function heavily restricts the frequency of ‘discourse initial’ so”. Along the same line, Buyse’s (2012) study found that this function occurred 15 times in the whole corpus, 13 in his non-native speaker data and only two in the native speaker data. He took this frequency in the non-native speaker data to be “relatively high” and explained it in terms of how the interviewers prompted the interviewees to start the conversation (2012: 1772).

A similar function of so is reported in another study conducted by Johnson (2002). She argues that when so prefaces questions, it performs as a topic developer: “We could describe the function of the so-prefaced questions as topic developers or topic sequencers whilst at the same time marking the discoursal act and topic boundary” (Johnson 2002: 103).

Using a different label, Müller (2005: 80) refers to this function as a “boundary marker”, denoting that it establishes the boundary between instructions and the beginning of a narrative. She claims that this function of so has not received attention from Schiffrin, Blakemore or Fraser. Defining this function, Müller (2005: 81) maintains that “so does not relate propositional ideas […] rather, it structures the spoken material into types of speech […] and thus can also be seen as functioning at the textual level”.

The interviewer in the following extract introduced the conversation and displayed the options from which the interviewee should choose. As can be noticed in the first turn by the interviewee,
the discourse marker so was used to start her turn ("ok so") or in other words to signal "the beginning of the narrative" (Müller, 2005: 80).

\[(6)\]

\(<A>\) play or play you had seen which you thought was particularly good or bad. bad describe the[:i:] film or the[:i:] play and say why (er) why you thought it was a good it was good or bad <</A> \\
\(<B>\) (uhu) ok so <</B> \\
\(<A>\) so which topic you want do you want to talk about <</A> \\
\(<B>\) (erm) well (erm) i can choose a film but can i choose a book instead <</B> \\
\(<A>\) yeah sure <</A> \\

(AR11)

4.6. Quantitative results for functions of “so”

Table 1 shows an overview of the main functions of so and their corresponding abbreviations in this study.

Table 2 and Figure 1 illustrate the distribution of the raw frequency of so, the non-discourse marker (non-DM) so, the discourse marker (DM) so and unclear instances in LINDSEI-AR.

Regarding the DM functions of so, from Table 3 and Figure 2 it can be observed that the most frequent DM function used by Saudis was so indicating a result. Saudi speakers used so indicating result 211 times or 27 times per 10,000 words. The second most frequent use of DM so was to introduce a new sequence, with 145 tokens, almost 19 times per 10,000 words. So marking the main idea unit was the third most frequent function with 76 tokens, 9 times per 10,000 words. So marking transition was used 31 times, or 4 times per 10,000 words. The function that occurred the least was so introducing a turn, with only two tokens in the entire corpus.

In terms of the comparison of the use of the functions of the DM so between males and females, none of the differences were statistically significant. This could be due to the size of the corpus, which is relatively small compared to other corpora, such as the British National Corpus (BNC).

### Table 1. Functions of so and their abbreviations

| Function                      | Abbreviation |
|-------------------------------|--------------|
| So indicating a result        | Result       |
| So marking main idea unit     | Main idea    |
| So introducing a new sequence | New seq.     |
| So marking transition         | Transition   |
| So introducing a turn         | Introductory |

### Table 2. Numbers of tokens of so in LINDSEI-AR (frequencies per 10,000 words in parentheses)

| Total number of words in B turns | Total frequency | Non-DM use | DM use | Unclear |
|----------------------------------|-----------------|------------|--------|---------|
| (77,075 words)                   | 715 (93)        | 264 (34)   | 423 (55)| 44 (6)  |
5. Summary and conclusion
The discourse marker so proves to be worth investigating due to its diverse functions. In the LINDSEI-AR sub-corpus, the analysis yields five significant functions. Similar to previous studies, the five functions are found in Saudi English language learners’ speech. Similar to Müller (2005) and Buyse (2012), the most frequent use of so in the LINDSEI-AR is so indicating result. However, unlike the most frequent so indicating result, the order of the frequency of the rest functions of so

---

Table 3. Functions of so in the LINDSEI-AR (frequencies per 10,000 words in parentheses)

| Total number of words in B turns | Result | New Seq. | Main Idea | Transition | Introductory |
|----------------------------------|--------|----------|-----------|------------|--------------|
| (77,075 words)                   | 211 (27) | 145 (19) | 67 (9)    | 31 (4)     | 2 (0.25)     |

---

Figure 1. Distribution of so in LINDSEI-AR.

Figure 2. Distribution functions of the DM so in LINDSEI-AR.
was found inconsistent in the literature. Therefore, even though previous findings have acknowledged the functions in this study but it seems unfeasible to relate each function’s frequency to its equivalent found in the literature.

The second most frequent function is so introducing a new sequence. Then comes so marking the main idea unit. The fourth most frequent function is so marking transition. The least frequent function is so introducing a turn. Based on these results, it is possible to claim that Saudi English language learners, both male and female, can use the functions of so appropriately. Regarding gender differences, the results show no statistical significance between Saudi males and females.

Last, but not least, there are some limitations that must be noted but are at the same time recommendations for future studies. First, it is important to acknowledge that the results of this study are restricted to a particular genre, namely informal interviews between staff members and their students. I assume that findings from informal conversations between peers or daily life conversations might yield different findings. The second limitation is the corpus size. Even though LINDSEI-AR can be claimed to be rather representative of universities in Saudi Arabia, I believe a larger corpus could produce different results, both qualitatively and quantitatively. Third, this study is specific to non-native speakers. Thus, it is recommended that future research compare and contrast the findings of this study using analysis of data from native speakers of English. Finally, it would be interesting to examine possible correlations between the use of discourse marker functions and L1 (Arabic) transfer and formal school education (textbooks) to provide a better mapping between the findings and underlying factors affecting the use of so.

Acknowledgements
The author would like to express his gratitude to the Ministry of Education and the Deanship of Scientific research – Najran University Kingdom of Saudi Arabia for their financial and technical support under code number (NU/SHED/17/201).

Funding
This work was supported by the Najran University [NU/SHED/17/201].

Author details
Sami Algouzi
E-mail: sialgouzi@nu.edu.sa
ORCID ID: http://orcid.org/0000-0002-0853-7203

1 English, Najran University, Najran, SAUDI ARABIA.

Citation information
Cite this article as: Functions of the discourse marker So in the LINDSEI-AR corpus, Sami Algouzi, Cogent Arts & Humanities (2021), 8: 1872166.

Note
1 The <B> turns are the interviewees’, whereas the <A> turns are the interviewers’.

References
Ajijmer, K. (1996). Conversational Routines in English. Convention and Creativity. London, Longman.
Ajijmer, K. (2002). English discourse particles: evidence from a corpus. John Benjamins.
Anping, H. (2000). On the discourse marker so. In P. Peters, P. Collins, & A. Smith (Eds.), New frontiers of corpus research (pp. 41–52). Rodopi.
Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English. Longman.
Blokemore, D. (1998). On the context for so-called discourse markers. In J. Williams & K. Malmkjøer (Eds.), Context in language understanding and language learning (pp. 44–60). Cambridge University Press.
Bolden, G. B. (2008). “So what’s up?” using the discourse marker so to launch conversational business. Research on Language and Social Interaction, 41(3), 302–337. https://doi.org/10.1080/03851810802237909
Bolden, G. B. (2009). Implementing incipient actions: The discourse marker ‘so’ in English conversation. Journal of Pragmatics, 41(5), 974–998. https://doi.org/10.1016/j.pragma.2008.10.004
Brinton, L. J. (1996). Pragmatic Markers in English. Grammaticalization and Discourse Functions. Mouton de Gruyter.
Buyse, L. (2017). So as a multifunctional discourse marker in native and learner speech. Journal of Pragmatics, 44(13), 1764–1782. https://doi.org/10.1016/j.pragma.2012.08.012
De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. Belgian Journal of English Language and Literatures (BELL), New Series, 2(2), 225–246. http://hdl.handle.net/2078.1/75157
Fraser, B. (1990). An approach to discourse markers. Journal of Pragmatics, 14(3), 383–395. https://doi.org/10.1016/0378-2166(90)90096-V
Fraser, B. (1996). Pragmatic markers. Pragmatics, 6(2), 167–190. https://doi.org/10.1075/prag.6.2.03fra
Fraser, B. (1999) What are discourse markers?, 31, 931-952. doi:10.1016/S0378-2166(98)00101-5
Fraser, B. (2006). Towards a theory of discourse markers. In K. Fischer (Ed.), Approaches to discourse particles (pp. 189–204). Elsevier.
Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. Applied Linguistics, 28(3), 410–429. https://doi.org/10.1093/applin/amnm030
Gilquin, G. (2008). Hesitation markers among EFL learners: Pragmatic deficiency or difference? Pragmatics and Corpus Linguistics: A Mutualistic Entente, 2, 119. http://hdl.handle.net/2078.1/775826
Appendix 1. LINDSEI transcription guidelines

1. Speaker turns

Speaker turns are displayed in vertical format, i.e. one below the other. Whilst the letter “A” enclosed between angle brackets always signifies the interviewer’s turn, the letter “B” between angle brackets indicates the interviewee’s (learner’s) turn. The end of each turn is indicated by either A or B.

2. Empty pauses

Empty pauses are defined as a blank on the tape, i.e. no sound, or when someone is just breathing. The following three-tier system is used: one dot for a “short” pause (1 second), two dots for a “medium” pause (1–3 seconds) and three dots for “long” pauses (3 seconds).

3. Filled pauses and backchanneling

Filled pauses and backchanneling are marked as (eh) [brief], (er), (em), (erm), (mm), (uhu) and (mmhm).

4. Unclear passages

A three-tier system is used to indicate the length of unclear passages: X represents an unclear syllable or sound up to one word, XX represents two unclear words, and XXX represents more than two words.
Unclear names of towns or titles of films, for example, may be indicated as the name of city or title of the film.

5. Anonymization
Data should be anonymised (names of famous people like singers or actors can be kept). Transcribers can use tags like the first name of interviewee, first name and full name of interviewer or name of professor to replace names.

6. Truncated words
Truncated words are immediately followed by an equals sign (=).

7. Foreign words and pronunciation
Foreign words are indicated by foreign (before the word) and foreign (after the word).

8. Phonetic features
(a) Syllable lengthening A colon is added at the end of a word to indicate that the last syllable is lengthened. It is typically used with small words like to, so or or. Colons should not be inserted within words.

(b) Articles—when pronounced as [ei], the article a is transcribed as a[ei]. When pronounced as [i:], the article the is transcribed as the[i:].

9. Prosodic information: voice quality
If a particular stretch of text is said laughing or whispering, for instance, this is marked by inserting starts laughing or starts whispering immediately before the specific stretch of speech and stops laughing or stops whispering at the end of it.
