



## Foreword

**DOI:**

[10.1075/scl.97.for](https://doi.org/10.1075/scl.97.for)

**Document Version**

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

**Citation for published version (APA):**

Denison, D. (2020). Foreword. In E. Jonsson, & T. Larsson (Eds.), *Voices past and present - Studies of involved, speech-related and spoken texts: In honor of Merja Kytö* (pp. xii-xiii). (Studies in Corpus Linguistics; Vol. 97). John Benjamins Publishing Company. <https://doi.org/10.1075/scl.97.for>

**Published in:**

Voices past and present - Studies of involved, speech-related and spoken texts

**Citing this paper**

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

**General rights**

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

**Takedown policy**

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact [uml.scholarlycommunications@manchester.ac.uk](mailto:uml.scholarlycommunications@manchester.ac.uk) providing relevant details, so we can investigate your claim.



Denison, David. 2020b. Explaining explanatory *so*. In Ewa Jonsson & Tove Larsson (eds.), *Voices past and present – Studies of involved, speech-related and spoken texts: In honor of Merja Kytö* (Studies in Corpus Linguistics 97), 207-25. Amsterdam: John Benjamins.

**This is the 'author accepted manuscript'. The published version (with a couple of trivial re-arrangements) cannot be posted on a website, but I am allowed to send a PDF offprint to an individual if requested.**

## Explaining explanatory *so*

David Denison

University of Manchester

### Abstract

This chapter examines a recent use of *so* in spoken British English, namely as a discourse marker conveying acceptance of an invitation to take the floor and give an explanation. I demonstrate a long-term increase in turn-initial *so*, dating the specifically 'explanatory *so*' to the 2010s in Britain. Evidence comes from corpora of academic discourse, of media language and especially of conversation. I argue that the usage is a coalescence of several well-attested discourse uses of *so*, perhaps strengthened by transatlantic influence. I explain the often hostile public reaction by the sentence grammar of *so*, also offering a general hypothesis about what makes an innovation salient and objectionable to conservative speakers.

## 1 Introduction<sup>1</sup>

A young academic, asked over a beer why he lived in Manchester when his post-doctoral fellowship was in Bristol, began a relaxed account of the reason as follows:

(1) **So** my girlfriend lives here. [...] (27 March 2019, attested DD)

Such uses of *so* are my topic. They tend to involve explanations, hence the title of this chapter, but that is only to give the principal context of occurrence, not the function. To anticipate my findings, explanatory *so* is a discourse particle in turn-initial position which has such functions as accepting an invitation to take the floor and prefacing an explanation. This chapter offers the first corpus-based approach to the recent development of explanatory *so* in British English.

Section 2 is a sketch of those uses of *so* in clause-initial position which can be classed broadly as belonging to sentence grammar, of which result clauses in particular will be relevant further on to both discourse functions and metalinguistic attitudes. Section 3 takes the discussion from the grammatical to the discourse level and reviews *so* as a discourse marker. Section 4 shows the rise of discourse *so* by means of a corpus investigation, conversation in the British National Corpus (BNC) and Spoken BNC2014, and academic English in the British Academic Spoken English (BASE) corpus. I also look briefly at

---

<sup>1</sup> I hope Merja Kytö, indefatigably generous contributor to the world of English linguistics, will appreciate this sketch, with its various dialogues: between speakers, between syntax and discourse or pragmatics, between usage and attitudes. I am grateful to reviewers and editors for pushing me in the right, or at least a better, direction. I wish also to thank audiences at the British Library ('English Grammar Day', July 2019) and Université Paris 3 ('New perspectives on language change and variation in the history of English', October 2019) for comments on different oral versions of the material.

broadcasting. Section 5 traces the origins of explanatory *so* in British English in the light of the corpus data. Section 6 explains the extraordinary reaction it has provoked in the media, and more generally why certain linguistic innovations but not others draw the (f)ire of the complaint tradition. Section 7 is a brief envoi.

In surveying the distribution of such a common word,<sup>2</sup> I confine my discussion to *so* at the start of a clause and with clausal scope. The function of *so* is largely unchanged when initial position is shared with an agreement marker. Throughout the chapter, therefore, the label ‘initial’ also covers a *so* preceded by an interjection, up to two words usually of assent or disagreement, as in the interviewee’s response in (2):

(2) Interviewer: It feels a bit out of sync with public policy, doesn’t it, though, I mean if we’re trying to move to zero carbon by 2050, in 30 years’ time, how does this play into (pause) to that goal?

Interviewee: **Yes, so**, there is a broader context within UK energy policy, we [...] (BBC Radio 4 *Today*, 17 June 2019)

## 2 *So* in sentence grammar

In sentence grammar, each item has a syntactic slot and makes some contribution to meaning, albeit not always amenable to glossing as we move down the cline from lexical to function words. The following sketch is based on the standard modern grammars of English (Quirk et al. 1985, Biber et al. 1999, Huddleston & Pullum 2002).

---

<sup>2</sup> The BNC recognises several *so* lexemes with different part-of-speech tags. Taken together, they constitute the 31st most frequent form in its spoken portion.

As a conjunction (at least in traditional terminology), *so* introduces a clause. Huddleston & Pullum (2002: 726) observe that ‘[s]o is used to indicate either purpose or result’, with the differences between the two functions discussed at (2002: 733), not just in relation to *so*. (See also Quirk et al. 1985: 1108–9.) If asked to give an example of clause-initial *so*, many people would probably come up with either purpose or result. I take them in turn.<sup>3</sup>

In purpose clauses like (3) (the first example is from Huddleston & Pullum 2002: 725), *that* is optional, and the whole subordinate clause can be placed either before or after its main clause.

- (3) a. We booked early **so that** we could be sure of getting good seats.
- b. We booked early **so** we could be sure of getting good seats.
- c. **So (that)** we could be sure of getting good seats, we booked early.

In result clauses like (4) (examples from Huddleston & Pullum 2002: 725, 1320, respectively), the *so*-clause can only follow the main clause. However, Huddleston & Pullum (2002: 1320–21) argue that *so* in (4) is close to being a coordinator. For Quirk et al. (1985: 442) it is a conjunct adverb that resembles a conjunction, but differs because its clause cannot be fronted without change of meaning, while *so* can be preceded by a coordinating conjunction, but not by a subordinating one (Quirk et al. 1985: 645–6). Finally, Quirk et al.

---

<sup>3</sup> A third variant, the manner clause, is described only in Huddleston & Pullum (2002: 968). It is said to be ‘comparatively rare [...] and usually interpreted as’ result or purpose – in fact obscuring the distinction between result and purpose. Their examples are

- i. He’d arranged the programme **so that** we had lots of time to discuss the papers.
- ii. I apply the hay **so that** only the tops of the plants show above it.

(1985: 1070) state that a result clause like (5) is a disjunct rather than an adjunct – that is, a speaker comment on the preceding main clause. A disjunct can be part of discourse structure. I have given their example verbatim, though in my own usage, *so* would be more natural here without *that*.

- (4) a. The airline had overbooked, **so that** two of us couldn't get on the plane.
- b. There was a bus strike on, **so** we had to go by taxi.
- (5) We know her well, **so that** we can speak to her on your behalf.

*So* has uses as an intensifier, part of a VP- or other substitute, and part of more complex adverbs and conjunctions, illustrated respectively in (6)a–c), all with *so* in initial position:

- (6) a. **So** cruelly did he treat them, that ...
- b. **So** I believe.
- c. **So long as** you insist on ...

*So* can also be an adverb, as in this example from Huddleston & Pullum (2002: 1319):

- (7) The mill could be sold off, **so** providing much-needed capital.

In (6) and (7), *so* does not introduce and have scope over the clause.

Corpus investigation of *so* requires string searches, as part-of-speech tags are not reliable, and any of the sentence grammar patterns (3)–(7) may turn up. Only result clauses like (4) will play some part in the story of explanatory *so*. The rest are irrelevant and must be filtered out manually.

### 3 *So* as discourse marker

Linguistic items may have discourse functions and thereby help to organise different sections of a text, or, at a lower level, make connections between adjacent utterances. They can also express speaker attitudes, and they can help to negotiate interpersonal relations between speaker and hearer. (See, for example, the references in Brinton 2006.) Unfortunately, items often serve more than one of these functions at the same time, in proportions which cannot be measured systematically, and according to classification schemes that vary from scholar to scholar.

Initial position is usually regarded as crucial for discourse markers. In spoken discourse, speakers generally take turns, and what is crucial about explanatory *so* is that it is *turn*-initial, as for example in (1).

Quirk et al. (1985: 633) observe that a conjunct such as *so* can actually be *discourse*-initial:

(8) **So** you're leaving, then! [intonation marking not reproduced here – DD]

[...] Discourse-initiating items can be less easy to account for plausibly, but it seems significant that such items are usually those that have a well-established conjunctive role in mid-discourse use. [...] It would seem that, in discourse-initial use, these items seek to enforce by implication some continuity with what *might* have gone before. Silence is difficult to break without some such convention. (Quirk et al. 1985: 633–4)

This observation neatly captures the behaviour of *so*, in (8) and more widely. The conjunct *so* appears under the semantic heading 'resultive' (label from Quirk et al. 1985: 635; cf. also 638) and is labelled 'informal'.

What are the discourse functions of *so*? Among the different but often overlapping classifications available, Buysse's survey is helpful for our purposes. Drawing on Halliday & Matthiessen (2004), he lists ten functions in three categories (Buysse 2012: 1767, Table 1), exemplified in my Table 1 below. If more than one can be discerned in the same example, he argues that one function is always more salient than the others, and in classifying his data, he allows an ideational or an interpersonal function to trump a textual one. 'Ideational' would include what I have called a result clause in sentence grammar, which Buysse defends including among discourse functions as well 'because it is both syntactically and semantically optional in this context' (2012: 1768), whereas purpose is not a discourse function (2012: 1767). If the dual status of result in both sentence grammar and discourse grammar seems analytically messy, I would contend in response that discourse markers have generally grown out of sentential functions, hence it is only to be expected that there should be some overlap between the two.<sup>4</sup>

I adopt Buysse's list of functions, adapting it slightly for my data.<sup>5</sup> Can it accommodate explanatory *so*? Two of the textual items on the list bear a close relationship to explanatory *so*

---

<sup>4</sup> The difficulty of assigning a word class to certain uses of *so*, noted by all three major grammars, testifies to its often liminal status.

<sup>5</sup> I considered adding a function such as 'Introduce a question', a context that is frequent in my data but unlikely to come from Buysse's interviewees. Prefaced questions cannot be assigned to his 'Prompt' function as a way for a speaker to yield the floor, as Prompt is said to occur at the **end** of a segment. I have tentatively decided not to include introducing questions in my tabulations, since most such examples can also be assigned to one of Buysse's ten functions. However, I have annotated *so*-questions as such in my data in case useful for further research. The decision does not affect my main line of argument.



without quite being suitable descriptions. The closest is ‘Introduce a section of the discourse’, where *so* can open an interviewee’s first turn, perhaps preceded by a short prompt from the interviewer (Buysse 2012: 1771–2). The other is ‘Introduce a new sequence’. In discussing that function, Buysse writes that *so* ‘can indeed introduce [...] a new step in an explanation’ (2012: 1768). I suggest that explanatory *so* goes further, in that the speaker actually *begins* an explanation at that point, and that the usage has the further function of accepting an invitation to take the floor. Hence explanatory *so* must be located among the interpersonal discourse uses. Note furthermore that it must be distinct from *so* in purpose or result clauses, because finite purpose clauses require a modal auxiliary, unlike explanatory *so*, while result clauses have different semantics and can begin with *and*, again unlike explanatory *so*. Explanatory *so* is not ideational, therefore, nor merely textual in function. I have added it to Buysse’s classification under the heading ‘Introduce an invited explanation’. My Table 1 reproduces Buysse’s Table 1 faithfully, apart from that one additional interpersonal function (indicated by italics), and with abbreviated examples from my corpora.

**Table 1: Discourse marker functions of so, adapted from Buysse (2012: 1767).**

<b>Type of relation</b>	<b>Discourse marker function</b>	<b>Abbreviated example</b> (from Spoken BNC2014 texts unless otherwise noted)
Ideational	Indicate a result	[Cos mummy made the tea ...] <b>so</b> daddy does the washing up. (BNC, KBW)
Interpersonal	Draw a conclusion	yeah <b>so</b> she was born in forty-three (SKDX)
	Prompt	but erm <b>so</b> (SM6B)
	Hold the floor	<b>so so</b> that 's that <b>so</b> but not much else has really happened (S26N)
	<i>Introduce an invited explanation</i> [explanatory <i>so</i> ]	<b>So</b> , I did a little research on what the conditions are like in the mines (TV Corpus, Big Bang Theory)
Textual	Introduce a summary	<b>so</b> they 're not free [...] (SHTM)
	Introduce a section of the discourse	<b>so</b> did you guys have a good week ? (SKHW) [out of the blue]
	Indicate a shift back to a higher unit of the discourse	<b>so</b> what time is the flight tomorrow ? (SBB2) [discussed earlier]
	Introduce a new sequence	<b>so</b> I was like oh my god I have to get this for her like (S23A) [continuing a narrative]
	Introduce elaboration	<b>so</b> that was brilliant you know cos [...] (S2T6)

Mark self-correction	So I had to <- -> [...] yeah, <b>so</b> I did have to guess a great deal (BNC, KC0)
----------------------	-------------------------------------------------------------------------------------

The classification of Table 1, if sufficiently comprehensive, should permit us to explore corpus data systematically in search of explanatory *so* and its precursors.

#### 4 Corpus data

Explanatory *so* belongs to recent spoken English. The most important reservoir of speech patterns is everyday conversational usage. Anecdotally, explanatory *so* is prevalent in academic speech and in radio or TV interviews (see e.g. Liberman 2010, Creighton 2015). Accordingly I have tried to find British corpus data that represents conversation, academic speech and broadcasting. Ideally we would have searchable diachronic corpora for all three domains that straddle the advent of explanatory *so*, or else comparable synchronic corpora from before and afterwards. Not all the desiderata can be met.

For conversation ‘before’, I took the demographically sampled spoken component of the British National Corpus, henceforth ‘Spoken BNC1994DS’. Recordings were mostly made in 1992, and the text-type is similar to the purely conversational Spoken BNC2014,<sup>6</sup> recorded 2012–16, permitting a comparison of usage about a generation apart. These corpora have the advantage of generous markup (BNC 2007, Love et al. 2017) and excellent search engines.<sup>7</sup>

---

<sup>6</sup> I am grateful to an anonymous referee for this point.

<sup>7</sup> BNC is available with BNCweb search software (Hoffmann et al. 1996–2010) at Manchester and Lancaster Universities. ‘Word counts’ are in fact token counts, as is normal with tagged corpora (e.g. *don’t* is split into a stem and a clitic token). The software CQPweb (Hardie 2013) was developed from BNCweb, and the CQPweb server at Lancaster hosts a

Compared to academic speech and media interviews, relaxed conversation among friends and family is not a context where explanatory *so* would be especially likely to occur, but we can at least expect information about the frequency of turn-initial *so*, and about which uses might allow an indigenous development of explanatory *so*.

For academic speech I used British Academic Spoken English (BASE, Nesi & Thompson 2000–2005), a corpus of transcriptions of lectures and seminars at two UK universities in 1998–2005.<sup>8</sup>

For broadcast speech it is easy to illustrate explanatory *so* with anecdotal data, for example in BBC Radio 4's daily *Today* programme. Unfortunately the BBC does not offer recordings of this programme beyond a rolling four-week catch-up period, and I have not located a suitable corpus. Some historical information on broadcast usage can be gleaned from the TV Corpus (Davies 2019) and from the work of Schlegl (2018).

#### 4.1 Conversation: Spoken BNC1994DS and Spoken BNC2014

I searched within a single turn for start of turn followed by up to two optional words, an optional piece of punctuation, and finally *so*.<sup>9</sup> The same CQP query was run on each corpus, wide range of corpora, including both BNC and Spoken BNC2014, but note that its token counts include punctuation. Search syntax and results are identical in both packages; only the frequencies differ.

<sup>8</sup> The corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council. The file pssem009.xml is missing from the claimed 200 files, never having been transcribed (Hilary Nesi, pers. comm., 1 September 2019).

<sup>9</sup> Search term “<u> ([ ] ){0,2} ([pos="PUN"])? [word="so"%c] within u”.

with the results shown in Table 2. The searches actually find turn-initial *so* (and not necessarily discourse *so*). Recall should have been very high, and precision in the samples extracted from each set of hits was surprisingly high too at 82–83%.<sup>10</sup> This means that the more than doubling in frequency shown in the table is probably a safe indication of a steep rise of turn-initial discourse *so*.

**Table 2: Turn-initial *so* in two conversational BNC corpora**

	Spoken BNC1994DS	Spoken BNC2014
Total tokens including punctuation (CQPweb)	5,014,655	11,422,617
Total ‘words’ excluding punctuation <sup>11</sup>	4,233,962	11,209,172
Turn-initial <i>so</i>	8,743	42,007
Frequency per million tokens (CQPweb)	1,743.49	3,677.53

<sup>10</sup> The precision figures taken from Table 3 below are simply the ratio of genuine examples of turn-initial discourse *so* in each sample (100) to the number of hits that had to be examined before 100 genuine examples had been found (120 and 122, respectively). The percentages serve as an estimate of the precision of the full search results reported in Table 2.

<sup>11</sup> The figure for Spoken BNC1994DS is given by the BNCweb interface. The figure for Spoken BNC2014 is calculated by subtracting the number of tokens tagged as *\_Y\**, namely 213,438, of which all but 7 are question marks, virtually the only punctuation used in that corpus.

The hits were randomly thinned down to 200 from each corpus, then classified according to the scheme of Table 1 (or as non-discourse or non-initial *so*).<sup>12</sup> Examination proceeded until 100 examples of turn-initial discourse *so* had been found. The results are shown in Table 3.

---

<sup>12</sup> Hits were discounted as being non-initial if something other than an agreement marker or similar preceded *so*, or if the turn was in fact a continuation interrupted by overlap with another speaker.

**Table 3: Comparison of turn-initial discourse *so* in Spoken BNC1994DS and Spoken BNC2014**

	Spoken BNC1994DS	Spoken BNC2014
Indicate a result	2	1
Draw a conclusion	18	11
Prompt	0	0
Hold the floor	10	20
Introduce an invited explanation [explanatory <i>so</i> ]	0	0
Introduce a summary	6	4
Introduce a section of the discourse	3	1
Indicate a shift back to a higher unit of the discourse	11	3
Introduce a new sequence	30	16
Introduce elaboration	14	42
Mark self-correction	2	0
Not classified	4	2
Subtotal	100	100
Non-discourse or non-initial <i>so</i>	20	22
Total hits examined	120	122
Precision in sample examined	83%	82%

The two directly causal functions, indicating a result and drawing a conclusion, have between them dropped by 40%, though numbers are too small for precise claims. However, a fall is consistent with a continued shift from *so* as a marker of logical inference towards a

more general core meaning, something like ‘in the light of what has been said previously’.<sup>13</sup> As already noted, overall use of turn-initial *so* has more than doubled between the two BNC corpora.

An interesting example in my Spoken BNC1994DS sample is (9), involving overlapping speech (signalled by <-|->) between mother and son:

- (9) Jane 1661 Which reminds me you still haven't written to Geoffrey and  
Jean to thank them for your birthday money <-|-> have we?  
David 1662 <-|-> **So** mum, I know <-|-> I haven't.  
Jane 1663 <-|-> Or have <-|-> you, which is more to the point. (KCH  
1659)

David's *so* seems to carry something of the interpersonal emotive charge seen in (8), and it does not follow in an obvious way from the previous discourse. As its function is at present unclear to me, I marked it as ‘not classified’.

There were no examples of explanatory *so* in my thinned set of 200 Spoken BNC1994DS hits, nor did I notice any in other, relatively cursory searches of the BNC. Although there were no examples in my thinned set of 200 Spoken BNC2014 hits, explanatory *so* does occur in the full set of hits, for example as in (10):

- (10) S0018: but do you put the water inside the ball ?  
S0146: **yeah so** you put all the potatoey stuff in and then you just fill up the  
rest of it with the tamarind water [...] (S4L9)

---

<sup>13</sup> Or perhaps sometimes ‘in the light of our shared contextual knowledge’, given the occasional use of discourse-initial *so*, as in (8), though this is probably uncommon, to judge from the low figures for introducing a section of the discourse in my samples.



#### 4.2 Academic speech: BASE

The whole BASE corpus of 1,644,942 tokens<sup>14</sup> was loaded into the MonoConc Pro concordancer and a regex search run for *so* as first word of a turn, possibly preceded by one or two other words and/or pauses.<sup>15</sup> A word preceding *so* was typically a marker of assent such as *okay*, *right*, *yeah*, *yes*, but also *sorry* or *possibly*, or else *so* was repeated. There were 313 hits in the whole corpus, a rate of 190.28 pmw, which is far lower than in either of the conversation corpora. Hits were examined and filtered until 100 examples of turn-initial discourse *so* had been found, classified as before. The results are shown in Table 4.

---

<sup>14</sup> MonoConc Pro gives a total of 1,633,617 words, including filled pauses; the difference is unimportant.

<sup>15</sup>Search term “<u[^>]+?> ((#|<pause[^>]+?>)[ ])?([a-z]+[ ])((#|<pause[^>]+?>)[ ])?{0,2}so\W”. Note that ‘<u ...>’ (for ‘utterance’) marks the start of a speaker turn, and ‘#’ represents a filled pause. In BASE there can be consecutive turns by the same speaker resuming after an audio or video demonstration or long pause.

**Table 4: Turn-initial discourse *so* in BASE**

	BASE
Indicate a result	2
Draw a conclusion	10
Prompt	0
Hold the floor	6
Introduce an invited explanation [explanatory <i>so</i> ]	0
Introduce a summary	22
Introduce a section of the discourse	1
Indicate a shift back to a higher unit of the discourse	9
Introduce a new sequence	19
Introduce elaboration	29
Mark self-correction	0
Not classified	2
Subtotal	100
Non-discourse or non-initial <i>so</i>	6
Total hits examined	106
Precision in sample examined	94%

There were some difficulties of classification. For example, although (11) did indicate a shift to a higher unit, repeated hesitation by staff member *nm0285* led me to code his *so* as primarily holding the floor. The trickiest case was (12), in the middle of a group presentation, where student *sf5092*'s *so* was open to several different analyses: introducing a new sequence, or a summary of preceding utterances, or elaboration. I even considered

explanatory *so*, given that she received an invitation to take over at the end of the preceding turn, but since this was not her opening contribution, and since introducing a new sequence was the clear function of a fellow-student's turn-initial *so* later in the presentation, that was the primary function recorded for (12).

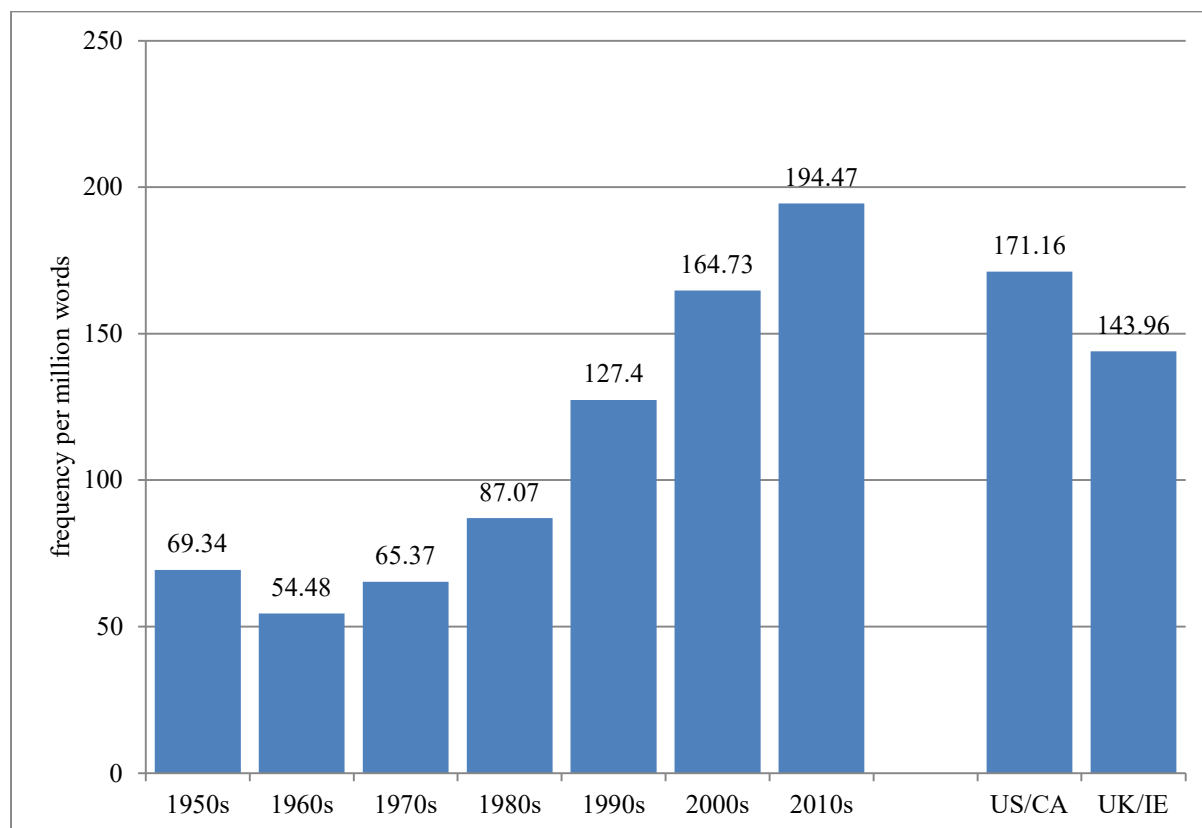
- (11) <u who="nf0289" trans="pause"> yes but you are advising <gap  
reason="inaudible due to overlap" extent="1 sec"/> </u>  
<u who="nm0285" trans="overlap"> **so so** we we we tell you what our  
opinion is <pause dur="0.2"/> and you can modify your proposal <pause  
dur="0.2"/> if you wish [...] </u> (ahsem008)
- (12) <u who="sm5093"> yeah we all decided in one voice one feeling one  
soul we decided that we can't judge yet the the Cuban Revolution because # it  
hasn't finished yet it's still running Castro will wake up this morning he will  
have his cigar and everything so it hasn't finished that's the point so continue  
</u>  
<u who="sf5092"> **okay so** basically we decided that according to the initial  
aims that we thought the revolutionaries had when they first came into power  
in nineteen-fifty-nine # seems to have been a failure # one of the main aims  
they seem to have had was to restore democracy to restore the constitution of  
nineteen-forty nineteen-forty </u> (ahsem003)

*So* was hardly ever used to introduce a response (4%), and no examples were classified as explanatory *so*, whether in the sample of 100 or the remaining 213 hits. Absence from a modest-sized corpus is not proof of non-existence, of course, but it is suggestive.

### 4.3 Broadcasting

A recently-released historical corpus of TV episodes from six English-speaking regions (Davies 2019) offers a promising 325 million words. The left-hand portion of Figure 1 plots occurrence over time in all TV shows in all regions of the string ‘? So’, used as a very crude proxy for a context where explanatory *so* might appear. There is a fall after the 1950s, but a steady increase from the 1960s to a frequency well over three times higher in the 2010s. The last two columns compare hits over the whole period in two of the three regional groupings: the overall frequency is 19% higher in USA/Canada than in UK/Ireland. Inspection suggests that most of the 54,472 hits are of discourse *so* (a higher proportion in the last decade of the corpus than the first), but that the overwhelming majority of hits are not explanatory *so*, and indeed many are themselves questions. It is not practical to do any statistical work on discourse types. Search tools are limited, and crucially, the corpus does not mark speakers, turns or even scene changes, so the linguistic context of a sentence-initial *so* can only be inferred and may not even be a response to the previous question.

**Figure 1: Initial *so* after question mark in the TV Corpus; figure adapted from (Davies 2019)**



A ‘virtual corpus’ of UK documentaries from the TV Corpus attests to the currency of explanatory *so*. There are about a dozen examples, all recent:

(13) A. And you're getting these things up to what sort of height ?

B. **So**, we're interested at the height of about 10km, but they continue on up to about 25–30km. (2016 BBC TV *Horizon*, episode ‘Ice Station Antarctica’)

For Canadian (and to some extent US) English, Schlegl (2018: 14–16) managed to assemble a corpus of broadcast material that spans five time periods ranging from 1951 to

2018, size not stated.<sup>16</sup> She tracks a number of discourse markers in a sample from the corpus, including *well* and *so*, and applies a mixed-effects regression model to the data. Among her interesting observations are (i) a striking decrease from the 1970s to 2018 in *well* as a marker of spontaneous new topics (50% to 16%), with a corresponding increase in *so* (25% to 66%) (Schlegl 2018: 30); and (ii) that utterance-initial *so*,<sup>17</sup> although present in her oldest data to mark ongoing speech, is being extended ‘primarily by young women’ into marking new turns and new topics, and that ‘the same uses [...] are apparently metalinguistically salient to members of the [Canadian] speech community’ (2018: 36–7).

## 5 Origins of explanatory *so*

Our analyses of recent British corpora will help to date the advent of explanatory *so* and give evidence of its precursor functions. Explanatory *so* is probably absent from Spoken BNC1994DS and from BASE (despite example (12)) but is present in Spoken BNC2014 in examples that fall outside the sample counted for Table 3. I briefly review some secondary literature.

Explanatory *so* is not described in the path-breaking study of discourse markers by Schiffrin (1987), which includes other uses of *so*, nor is it recognised in any of the standard reference grammars (Quirk et al. 1985, Biber et al. 1999, Huddleston & Pullum 2002), nor is precisely that usage yet in the *Oxford English Dictionary* (but then it is not typically a written

---

<sup>16</sup> I am grateful to Lisa Schlegl for letting me see her dissertation, and to Heike Pichler (pers. comm. 14 March 2019) for putting me in touch with her. Note too that Pichler was interviewed on BBC Radio 4’s *Feedback* programme on 3 November and 10 November 2017 about listeners’ complaints about language annoyances in broadcasts, among others the use of *so* (2017).

<sup>17</sup> ‘Utterance’ is not defined but is not necessarily a whole turn by one speaker.

usage). Shortly after the turn of the millennium, Bolden (2003a, 2003b, 2008, 2009 etc.) began a series of conference papers and publications on *so* in American conversation, though not specifically explanatory *so*. Blevins (2015), apparently American, has a level-headed survey of all kinds of ‘sentence-initial *so*’ in history. She offers the following as a ‘commonly touted theory’ about its recent spread, a theory which apparently targets something close to explanatory *so*:

This was first noted by Michael Lewis in *The New New Thing* (1999)- “When a computer programmer answers a question, he often begins with the word ‘so’.” As to how this came about, it is thought that given the international composition of the typical Silicon Valley work site, where a large number did not speak English as their first language, it became the simple “catchall” word of transition. Over time and frequent usage, it eventually became like a tic and just part of the common speech pattern of those in that industry and then spreading beyond. (Blevins 2015)

All of this suggests that explanatory *so* is largely a twenty-first-century development there. Public reaction in the UK doesn’t start till well into the 2010s.

I have no data of my own to test the plausibility of a computer programmers’ *lingua franca* in California for the initiation of the usage, nor of Schlegl’s better-supported suggestion (Section 4.3 above) of a spread led by young women, at least in Canada. My concern in this chapter is with British English. It is quite possible that explanatory *so* in Britain was introduced or reinforced by North American influence, whether in academic usage or via the media. The alternative explanation, not incompatible, is that it arose naturally within British English. Indeed a dual source seems likely.

For a possible internal origin I suggest the following scenario. Turn-initial *so* has been on a long-term trajectory of increase at the expense of *well* and other discourse particles in speech. For a speaker conveying expertise, *so* has the particular advantage over *well* of not sounding vague or tentative, instead briskly suggesting a logical basis for what follows. In my discourse *so* data, the function ‘Introduce a new sequence’ is quite common (16 to 30% in the three corpora), whereas ‘Introduce a section’ is not (1 to 3%). Explanatory *so* seems to combine elements of these two textual functions: from the former, a step within an exposition; from the latter, an opening gambit. Its interpersonal function of ‘Accept an invitation to take the floor’ is an extension of the interpersonal function ‘Hold the floor’ (see Table 1), which itself appears to have become more common, rising to 20% in Spoken BNC2014. This whole complex of functions has come to be associated with the giving of explanations. Whatever the precise source of explanatory *so*, my strong impression is that it has become common in academic circles for explanations and research presentations, which may explain in part its appearance in broadcast interviews.

## 6 Public reaction to explanatory *so*

There is an interesting interaction between language change and language attitudes. Explanatory *so* hits a nerve in some conservative speakers, when other innovations pass unnoticed or at least unchallenged. The usage has attracted mostly hostile media discussion, as for example an interview with the radio presenter John Humphrys (Creighton 2015), a leader in *The Times* (anon 2017), and the radio/podcast appearances by Pichler (Dixon 2017) and Gopnik (2018);<sup>18</sup> compare Shariatmadari (2017) and Liberman (2010).

Dr Bernard Lamb, President of the Queen’s English Society, is quoted as follows:

---

<sup>18</sup> Reference due to Stefan Krasowski in a comment on (Liberman 2019).



I think [the use of ‘so’ at the beginning of a sentence] is a sign of someone who is not particularly fluent, it’s fulfilling the function of ‘ummm’ and ‘errrr’ and giving the person a bit longer. It’s not being used as a conjunction to join things up, which is how it should be used. I think someone started doing it and then other people have begun slavishly copying it, it becomes fashionable. It’s just carelessness, it doesn’t have any meaning when used this way. (Creighton 2015)

The quotation implies that in so-called ‘good’ or ‘correct’ usage, *so* is always and only a subordinating conjunction, which is not true. Indeed in everyday speech, conjunction uses are only a small minority of turn-initial *so*.<sup>19</sup> Nor is it sufficient to refer only to the word’s part of speech and syntax: the semantics, pragmatics and discourse functions must be considered as well. As for ‘giving the person a bit longer’, that is what a discourse grammarian would call ‘holding the floor’.

Why, though, *is* there such a reaction? I offer the following explanation, framed somewhat simplistically in terms of ‘conservative’ and ‘advanced’ users. The conservative user does not themselves have explanatory *so* in their repertoire, but it is highly salient to them when heard, coming typically as the very first word uttered by an advanced speaker.<sup>20</sup> The

---

<sup>19</sup> Among the hits examined there are 5 (4.0%) result or purpose clauses in Spoken BNC1994DS and 3 (2.5%) in Spoken BNC2014. There might be a higher proportion if *so*-clauses in mid-turn were included.

<sup>20</sup> Compare a use of *because* as a discourse marker by several UK radio presenters without any real causal link to the preceding clause:

(i) And um is that piece typical of what’s on the album, **because** it’s the biggest piece, it’s the centre of the album, around which the smaller pieces are in orbit, she says. um Is the

Recency and Frequency Illusions (Zwicky 2005) kick in to amplify the salience (and possible annoyance), leading to the assumption that the usage is more frequent, and perhaps more recent, than it actually is. Language peevers rarely distinguish between the grammars of spoken and written English, and the quotation above unrealistically demands that speakers ‘talk like a book’. In an informal setting it would actually be rather unnatural to launch into an explanation without an introductory discourse particle.

Note that the conservative user is unlikely to be riled by hearing *so* introducing a question from an interviewer, as in this example from the BNC:

- (14)        **So** are we are we talking er do you see this as a as a as a launch pad?  
               (J9X 460)

That usage has been common for years, optionally with a separate intonation contour and rising intonation, or unstressed as part of the intonation contour of the following clause. What seems to annoy the conservatives is *so* before the answer.

Now, explanations can involve clauses of reason, result or purpose. Among the multiple possible forms of expression, *so* and *because* can act as approximate converses:

- (15)        He went for a walk **because** he was bored.  
               (16)        He was bored, **so** he went for a walk.

---

whole thing effectively a single piece? (Andrew McGregor, in conversation with Sara Mohr-Pietsch, ‘Record Review’, BBC Radio 3, 2 March 2019, 1:59:08)

If there is any causal sense, it is metatextual (‘the reason I am asking is’). The usage aids fluency and continuity, and in mid-turn and indeed mid-sentence position is far less salient than explanatory *so*. It does not, as far as I know, provoke listener complaints.

See also Quirk et al. (1985: 1108–9), Schiffrin (1987: Chapter 7). Increasingly, however, a whole explanation can be prefaced by *so*, as we have seen. Recall example (1), repeated and amplified below as (17):

- (17)       A. Why do you live in Manchester if your job is in Bristol?  
               B. **So** my girlfriend lives here. [...]

Contrast the reply with what were previously more conventional alternatives, all still in use and in variation with *so*, such as

- (18)       a.       B. **Because** my girlfriend lives here.  
               b.       B. **Well/You see/Actually/The reason is/...**, my girlfriend lives here.  
               c.       B. My girlfriend lives here.

To the conservative language user, *so* in (17) is being misused. While the sentence pattern is like (18) – a subordinate clause whose main clause is understood from the preceding turn – the lexical choice of clause-introducer is apparently like (16). But in (16) the second clause expresses result, not reason.

I speculate that one or both of the following conditions must obtain for people to have a hostile reaction to a linguistic innovation – given, of course, that they have noticed it at all:

- (19)       a. They (think they) remember having been taught that it's an error.  
               b. They can use those words themselves, but not with that meaning or function.

For explanatory *so*, it is condition (19) that is satisfied. Note that an unfamiliar word, even though noticeable, satisfies neither condition and is probably less likely to be complained about.

Some conservative speakers have gone some way towards recognising the new use of *so*. A telling passage occurs in the following online advice, recommending the avoidance of initial *so* in business presentations (Thurman 2014):

That little head cock, slight furrowing of the brow, and set-up with “so” says to your audience, “I’m trying to dumb this down so someone like you may have at least a chance of comprehending the importance of what I do.”

In other words, explanatory *so* is said to come over as patronising and condescending. Gopnik (2018) has the same interpretation.

However, what I observe is that many people can use turn-initial *so* quite freely when giving explanations – which is, after all, one of the main jobs of someone in academic life – and without the condescension that Thurman and Gopnik detect. It is unstressed, apparently unremarkable (to them), a polite but crisply efficient signal that they accept an invitation to take the floor. I don’t think it necessarily comes with a pause. I do not find what Bob Ladd did some years ago (comment dated 25 August 2010 below Liberman 2010):

The pause mentioned by a couple of people is also relevant – the usage in question usually has a prolonged vowel, on a fairly steady level pitch, followed by a pause before the rest of the discourse begins.

The discourse semantics of explanatory *so* conveys the acceptance of an invitation to speak, and it signals that what follows is an explanation. It occurs frequently now without any sense on either side that a requested answer requires specialist knowledge or training. No doubt there are cases where a speaker feels they have a superior right to offer an explanation and perhaps also that their expertise should be acknowledged, but that would be situationally

contingent – what we might call discourse pragmatics. I suggest that it is not a function of explanatory *so*.

## 7 Envoi

Explanatory *so* has become widespread in North America, and in the last decade or so in Britain, where it is likely to have developed at least in part out of existing uses of discourse *so*, as explored in Section 5. However, those other uses remain available and in many cases frequent. The same speakers may use *so*, for instance, to resume a narrative, to give a summary, to hold the floor. The explanatory function appears in a context that is particularly noticeable to non-users, but it is not far removed from other functions of longer standing.

Explanatory *so* is a distinct, identifiable usage that deserves a label of its own – in the modern parlance, ‘explanatory *so* is a thing’ – but it is after all neither an isolated nor a surprising development.

## References

- anon. 2017. So What? Well, it’s time to brush up (and vary) your verbal tics. *The Times*, 15 November 2017. <not read in full; behind paywall>
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson.
- Blevins, Melissa. 2015. So, when did we start introducing sentences with so? Today I Found Out: Feed Your Brain. <<http://www.todayifoundout.com/index.php/2015/09/start-introducing-sentences/>>.
- BNC. 2007. The British National Corpus, version 3 (BNC XML Edition). Oxford University Computing Services on behalf of the BNC Consortium.

- Bolden, Galina B. 2003a. Self and the other: The use of “oh” and “so” in sequence-initial position. Paper presented at Annual Conference on Language, Interaction, and Social Organization, Santa Barbara CA.
- Bolden, Galina B. 2003b. The use of “so” and “oh” in sequence-initial position. Paper presented at Annual Conference of the National Communication Association, Miami FL.
- Bolden, Galina B. 2008. "So what's up?": Using the discourse marker *so* to launch conversational business [Jul–Sep]. *Research on Language and Social Interaction* 41(3): 302–37.
- Bolden, Galina B. 2009. Implementing incipient actions: The discourse marker 'so' in English conversation. *Journal of Pragmatics* 41(5): 974–98.
- Brinton, Laurel J. 2006. Pathways in the development of pragmatic markers in English. In Ans van Kemenade & Bettelou Los (eds.), *The handbook of the history of English* (Blackwell Handbooks in Linguistics), 307–34. Oxford: Blackwell.
- Buyse, Lieven. 2012. So as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics* 44(13): 1764–82.
- Creighton, Sam. 2015. SO wrong! Why John Humphrys is in a rage at such a little word after it invades everyday speech. *Mail Online*, 20 June 2015.  
<<https://www.dailymail.co.uk/news/article-3132195/SO-wrong-John-Humphrys-rage-little-word-invades-everyday-speech.html>> (30 March 2019).
- Davies, Mark. 2019. TV Corpus. Brigham Young University.
- Dixon, Kate. 2017. Feedback, presented by Roger Bolton, with guest Heike Pichler. In *Feedback podcast*: BBC Radio.
- Gopnik, Adam. 2018. On prefixes. In *A Point of View*: BBC Radio 4.

- Halliday, M. A. K. & Christian M. I. M. Matthiessen. 2004. *An introduction to functional grammar*. London and New York: Edward Arnold.
- Hardie, Andrew. 2013. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380–409.
- Hoffmann, Sebastian, Stefan Evert, Hans-Martin Lehmann & Peter Schneider. 1996–2010. BNCweb (CQP-Edition): A web-based interface to the British National Corpus. <<http://www.bncweb.info/>> (2009).
- Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- Lewis, Michael. 1999. *The new new thing: A Silicon Valley story*. New York: Norton. Repr. 2012.
- Liberman, Mark. 2010. So new? Language Log. <<http://languagelog.ldc.upenn.edu/nll/?p=2570>> (13 April 2019).
- Liberman, Mark. 2019. So. Language Log. <<https://languagelog.ldc.upenn.edu/nll/?p=45575>> (26 December 2019).
- Love, R., C. Dembry, A. Hardie, V. Brezina & T. McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3): 319–44.
- Nesi, Hilary & Paul Thompson. 2000–2005. BASE: British Academic Spoken English. Warwick: University of Warwick.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London and New York: Longman.
- Schiffrin, Deborah. 1987. *Discourse markers* (Studies in Interactional Sociolinguistics 5). Cambridge: Cambridge University Press.

Schlegl, Lisa. 2018. *Tracking change in Canadian English utterance-initial discourse markers*. MA dissertation, University of Toronto.

Shariatmadari, David. 2017. So, what's the problem with 'so'? [15 November]. *The Guardian*:

Thurman, Hunter. 2014. How a popular two-letter word is undermining your credibility. .

*Fast Company*. <<https://www.fastcompany.com/3029762/how-a-popular-two-letter-word-is-undermining-your-credibility>> (26 April 2019).

Zwicky, Arnold. 2005. Just between Dr. Language and I. Language Log.

<<http://itre.cis.upenn.edu/~myl/language-log/archives/002386.html>> (4 May 2006).