Brno Studies in English
Volume 43, No. 2, 2017
ISSN 0524-6881, e-ISSN 1805-0867
DOI: 10.5817/BSE2017-2-2

Carlos Prado-Alonso

A Corpus-based Analysis of *Do So* Anaphora in Fictional and Non-fictional English

Abstract
This paper presents an in-depth corpus-based analysis of 'do so' verbal anaphora constructions in different fictional and non-fictional written English texts taken from different computerised corpora of British and American Present-day English, comprising texts from the 1960s, 1990s and 2000s. 'Do so' verbal anaphora, as in 'I ate an Apple yesterday in the park, and Peter did so last week', has received extensive attention from a theoretical perspective. Research has focused mainly on the analysis of the categorical factors – i.e. semantic and syntactic – that determine the use of the construction. Little research, however, deals with the analysis of 'do so' anaphora in real written English. The present analysis, based on tested criteria of multidimensional linguistic variation, sheds light on the linguistic and textual factors that drive the pragmatic use and the distribution of the construction. It will be shown that, in addition to semantic and grammatical factors, genre variation also plays an important role in the use of 'do so' anaphora in written discourse.

Keywords
Verbal anaphora; 'do so' constructions; textual variation; corpus-based studies; written English

1. Introduction

Do so constructions, as in '*The medical profession would like to test patients but cannot do so without informed consent*', are verbal anaphors that have received extensive attention from a theoretical perspective. Research has focused mainly on the categorical factors – i.e. semantic, and grammatical – that determine the use of the construction. It has been argued, for instance, that the extent of application

of *do so* anaphora depends principally on factors such as: (a) non-stativity of the antecedent (Lakoff 1966); (b) antecedent not headed by *be* (Levin 1986); (c) coreferentiality of subjects in the antecedent and *do so* clauses (Souesme 1987); (d) adjunct status of any "orphan" in the *do so* clause (Culicover and Jackendoff 2005); (e) non-contrastive status of any adjunct in the *do so* clause (Stirling and Huddleston 2002); (f) antecedent embedded rather than matrix predicate (Levin 1986); (g) voice or category differences between antecedent and the *do so* clause (Stirling and Huddleston 2002); and (h) adverse connotations of the antecedent (Bolinger 1970).

Little research, however, deals with the analysis of *do so* anaphora in naturally occurring discourse. In particular, there is currently no work on the textual factors affecting the distribution and pragmatic use of *do so* constructions in Present-day English, other than some isolated hints here and there (cf. Miller 2011). In order to fill this gap, the present paper presents an in-depth corpus-based analysis of the factors that drive the pragmatic use and distribution of *do so* constructions in different contemporary fictional and non-fictional written English texts. The data for the study are taken from six computerised corpora of British and American Present-day English, namely the LOB, FLOB, Brown, FROWN, BE06, and AmE06 corpora, comprising texts from the 1960s, 1990s and 2000s (for details see Hofland et al. 1999 and Baker 2009).

The paper is organised as follows. Section 2 offers some preliminaries, such as the structural and semantic patterns of *do so* constructions. Section 3 reviews the literature on *do so*, the study of which has been neglected from a corpus-based perspective. Section 4 offers a corpus-based analysis of *do so* constructions in fictional and non-fictional English. Section 5 seeks to shed light on the linguistic and textual factors that determine the distribution and pragmatic use of this construction in written English. Finally, section 6 offers some concluding remarks.

2. The structural and semantic patterns of 'do so' anaphora

The scope of anaphora in English has been discussed widely over recent decades, and references to it can be found in many of the best-known descriptions of English (cf. Halliday and Hasan 1976: 314, Lakoff and Ross 1976, Quirk et al. 1985 §12.21–26, Stirling and Huddleston 2002: 1529–1532, Culicover and Jackendoff 2005, among others). Verbal anaphors include *verb phrase ellipsis*, *do it* anaphora, *do that/this* anaphora and *do so* anaphora, as illustrated in (1)–(4) respectively, which share features such as the need of an antecedent for full understanding.

(1) Neville and Hibbert both had to see red, but Kuyt and Carragher should have, too. (BrE06, Press Editorial. B20)

(2) Enlightenment portraits make mental notes; they often do it in writing. (BrE06, Press Review. C10)

(3) You have to learn to trailer your horse because you cannot expect Quint to do that for you. (AmE06, Adventure and Western. N18)

(4) Users can see their own personal files but they can only see other people's files when they've been given permission to do so. (AmE06, Popular Lore. F45)

Do so anaphora, which will be our concern here, has been treated as a complex pro-form that serves as an anaphoric verb phrase (cf. Huddleston and Pullum 2002: 1529). Viewed independently, however, both *do* and *so* have been seen as versatile clause substitute devices (cf. Miller 1990). There is general agreement in the literature as regards viewing *so* as an adverbial modifier in *do so* constructions (cf. Hankamer and Sag 1976, Kehler and Ward 2004, Bos and Spenader 2011, among others). Arguments supporting this view include the fact that, in contrast to the *it* of *do it* anaphora, the *so* of *do so* cannot be the subject of a sentence with a passive verb (cf. 5), and that it cannot be the object of a preposition (cf. 6) because it does not display nominal properties.

(5) Someone broke our front window, and we think that it/*so was done sometime around noon. [Bouton 1970: 22]

(6) Jeremy had been planning to propose to Marilyn for several weeks, but the doing of it/*so in public he hadn't counted on. [Bouton 1970: 25]

The *do* of *do so*, in turn, has been regarded as the head of the pro-form and has generally been considered a main verb rather than an auxiliary (cf. Lakoff and Ross 1976 or Kehler and Ward 1999, among others). An argument supporting this claim is that, in this type of construction, *do* has semantic content and is compatible with non-stative or eventive antecedents. This is illustrated in (7), where the use of *do so* is incompatible with the stative mental verb *to know*.[1] Further evidence for considering the *do* of *do so* a main verb is that *do so* does not undergo subject-auxiliary inversion in polar questions: rather, do-support provides the auxiliary, as shown in (8).

(7) *I know the Easter Bunny is real, and Kent does so, too. [Houser 2010: 7]

(8) I ate my sandwich in one sitting, but did Grant do so? [Houser 2010: 14]

Even if the most common manifestation of the construction is precisely the form *do so*, some authors have noted a case of inversion in the positioning whereby the constituents of the combination are reversed (see Hankamer and Sag 1976, Stirling and Huddleston 2002: 1532, Kehler and Ward 2004). As illustrated in (9) and (10), this can happen in both *to*-infinitive constructions and gerund-participles.

Such variants, regarded as rather exceptional, are infrequent and occur almost exclusively in formal registers.

(9) We are left helpless to cope with it because we do not dare speak of it as anything real for to do so (so to do) would imply a commitment to that which has already been discredited and proved false. (Brown, Religion. D01)

(10) I had no intention of writing, but, if I had, the appointment of Mr. Muggeridge would have seemed to me to rule out any possibility of successfully so doing.
(LOB, Belles-Lettres. G14)

Semantically speaking, some studies have claimed that *do so* anaphora is not compatible with all types of antecedents. Lakoff (1966), for instance, maintains that *do so* is possible with non-stative antecedents (cf. 11a), but not with stative antecedents (11b). Likewise, Kehler and Ward (1999:14) conclude that *do so* is only compatible with antecedents that denote events and not states. A similar view is expressed by Culicover and Jackendoff (2005: 284) who – distinguishing between states (12a), non-action events (12b), and actions events (12c) – argue that *do so* anaphora is only compatible with actions events. However, recent corpus-based studies (cf. Michiels 1978, Houser 2010, or Miller 2013) have shown that *do so* constructions are possible with stative antecedents in naturally occurring discourse, though they are dispreferred. Miller (2013) in fact shows, by using an acceptability experiment, that the use of *do so* with stative antecedents is judged to be only slightly less acceptable than the use of *do so* with eventive antecedents.

(11) a. I learned the answer, although Bill told me not to do so.
b. * I knew the answer, although Bill told me not to do so.
[Lakoff 1966: 45]

(12) a. * Robin dislikes Ozzie, but Leslie doesn't do so.
b. * Robin fell out the window, but Leslie didn't do so.
c. Robin read the newspaper today, but Leslie didn't do so.
[Culicover and Jackendoff 2005: 284]

From a semantic and pragmatic perspective, it is also widely acknowledged that *do so* performs an anaphoric function, avoiding the repetition of identical verb phrases, and corefers with the antecedent from which it takes it meaning. It has been shown that *do so* can stand anaphorically for an entire verb phrase, a verb, its complement, and/or an adjunct that is not string-contiguous (cf. Culicover and Jackendoff 2005: 125). One of the semantic issues surrounding *do so* is whether it is a case of deep or surface anaphora. Following the seminal division by Hankamer and Sag (1976), surface anaphors are syntactically restricted and need a syn-

tactic antecedent, while in deep anaphors a referent of the appropriate semantic type suffices. Given the construction's behaviour, and with a few exceptions (e.g. Houser 2010), broad consensus exists that *do so* qualifies as a type of surface structure anaphora because, as the above examples illustrate, the pro-form requires an explicit antecedent (see Ward and Kehler 2002, Kehler and Ward 2004).

Adding to its anaphoric function, Ward and Kehler (2002: 11–12) argue that *do so* can also be used for standard hyponymic reference. They illustrate their argument by evaluating variants of the following sentence: *The hit man dispensed with his mob boss by shooting him in broad daylight, with plenty of witnesses around*. According to them, the progression in (13) goes from specific to general in the sense that (13a) makes uses of the same verb as the original sentence (*shoot*), (13b) turns to a more general verb (*murder*), and (13c) is the broadest due to the presence of *do so.*

(13) a. By so shooting him, the hit man established himself as his victim's likely successor.
b. By so murdering him, the hit man established himself as his victim's likely successor.
c. By doing so, the hit man established himself as his victim's likely successor. [Ward and Kehler 2002: 11–12]

As Huddleston and Pullum (2002: 1532) note, *do it* and *do that/this* differ semantically from *do so* in that they require an agentive interpretation, whereas *do so* can denote a non-agentive dynamic situation, as shown in (14). In this example, the use of *do so* is grammatical because the tree's falling is dynamic rather than static, but the tree does not have the role of the agent and this makes the use of *do it* and *do that/this* ungrammatical. Huddleston and Pullum's analysis, based on corpus data, contradicts Cullicover and Jackendoff claim that non-actional antecedents are impossible in *do so* constructions (cf. 12), and shows that, even if there is a preference for actional antecedents, well-formed examples of *do so* anaphora with non-actional eventive antecedents may be also attested.

(14) When the tree fell, it did so / *did it with a loud crash. [Huddleston and Pullum 2002: 1532]

(15) A: Rover is scratching the door.
B: Yes, he always *does so*/*does it*/*does that* when he wants attention. [Quirk et al. 1985: 876]

Despite this semantic restriction – because they are kinds of verbal anaphora – *do it* and *do that/this* can be used in linguistic contexts where *do so* is common, as shown in (15) above. Recent research (cf. Miller 2011, 2013) has shown that textual variation stands as a key factor in the usage of these types of verbal anaphora. The usage of *do so* has in general been regarded as more typical of formal registers

(cf. Macía-Vega and Payne 2007 or Miller 2013), whereas *do it* and *do that/this* as more typical of informal registers in conversational and fictional texts (cf. Biber 1992).

In the following review of the literature on *do so* anaphora it will be noted that there is currently an absence of corpus-based studies, and that a more fine-grained analysis, based on tested criteria of linguistic variation, is needed in order to provide more representative conclusions as to the textual environment of the construction. Such an analysis will be provided in section 4.

3. Previous research on 'do so' anaphora

Studies on anaphora in relation to textual variation are not rare, but they have been limited mainly to the occurrence of anaphors in individual text-styles (cf. Lord and Dahlgren 1997, among others). Research dealing with the dispersion of anaphora across genres is less commonly found. Most studies dealing with anaphora across genres are synchronic analyses, looking at how textual information is distributed, sometimes by taking a corpus as their basis (e.g. the LOB corpus in the case of Kurzon 1985) and sometimes by studying discursive differences between two specific genres (e.g. telephone conversations vs. written narratives, in Fox 1987). In such studies, anaphora has been approached from a variety of perspectives ranging from the analysis of its situational characteristics to the study of its rhetorical functions in the text, and/or its specific linguistic features (see Atkins and Biber 1994).

Most studies on the *do so* verbal anaphora in particular are of a theoretical nature, that is, are descriptive accounts of the syntactic and semantic features of *do so* and/or reports on the characteristics of the elements surrounding the construction that favour or hinder its use (cf. Stirling and Huddleston 2002, Culicover and Jackendoff 2005). Most interest in *do so* anaphora has therefore been generated by the study of its categorical – i.e. semantic, syntactic and grammatical (cf. Section 1) – dynamics of use, and, given its intrinsic anaphoric nature, it has even been studied as a test for *nouniness* (cf. Ross 1973).

In the last decade these categorical factors that determine the use of *do so* have also been assessed, albeit infrequently, through the use of corpora (cf. Houser 2010 and Miller 2011). Houser (2010), for instance, analyses the semantic restrictions on *do so* in 1,000 examples extracted from the American National Corpus (cf. Reppen et al. 2005), and shows that in 98% of his instances *do so* strongly prefers to occur with non-stative antecedents. Similarly, Miller (2011) shows that there is a preference for *do so* to occur with the same subject as its antecedent and to occur with non-contrastive adjuncts.

In general, however, corpora have been used very rarely in the study of *do so* across genres, and those few corpus-based studies on the topic "have focused on the analysis of some particular genre, such as newspaper texts, popular fiction, or conversation" (cf. Biber 1992: 216). The very few corpus-based analyses dealing

with *do so* across genres – cf. Miller (2011) – also tend to suffer from limitations of various kinds.

Miller (2011:2), for instance, contends that genre variation stands as a key factor in the usage of *do so* and shows that the construction is more frequently attested in academic prose and newspapers than in spoken and fictional text-styles. His analysis, however, is based on only 100 random occurrences collected from the Corpus of Contemporary American English (cf. Davis 2008). As Biber (1988: 191) has noted, "there are systematic patterns of variation within the major genre categories of a corpus". Hence, a more comprehensive cross-genre analysis of *do so* which examines all these systematic variations is needed, as only then can a clear picture of the distribution, uses and functions of the construction in different genres begin to emerge.

Despite the existing body of research, then, some aspects of the construction have been entirely neglected or demand further clarification. Biber et al. (1999: 432), for example, is a highly regarded grammar with a focus on corpus data, yet whereas it provides a cross-register comparison of the distribution of verbal anaphors such as *do it* and *do that/this*, it does not do this for *do so*. Such a noteworthy absence, I believe, deserves careful attention. The present study is a first step in that direction: section 4 provides a detailed corpus-based study of *do so* in fictional and non-fictional British and American English genres, and section 5 deals with the textual dimensions of the construction.

4. A corpus-based analysis of 'do so' in fictional and non-fictional British and American English

4.1 The corpora

The corpora used here to analyse the behaviour and distribution of *do so* anaphora in written discourse are: 1) the *Lancaster-Oslo-Bergen Corpus of British English* (LOB; compilation date: 1961), 2) *the Brown Corpus of American English* (Brown; compilation date: 1961), 3) *the Freiburg-Lancaster-Oslo-Bergen Corpus of British English* (FLOB; compilation date: 1991), 4) *the Freiburg-Brown Corpus of American English* (FROWN; compilation date: 1992), 5) *the British English 2006 Corpus* and the *American English 2006 Corpus* (see Hofland et al. 1999 and Baker 2009). The six corpora are parallel in structure and each comprises 500 samples of approximately 2,000 words each, thus totalling 1,000,000 running words organised into fifteen textual categories, of which the following have been selected for the present analysis: *Science Fiction*, *Adventure and Western*, *Mystery and Detective fiction*, *Romance and Love Story*, *General Fiction*, *Official Documents*, *Press Editorial* and *Belles-lettres* and *Biographies*. These categories have been further grouped into fictional and non-fictional texts. A total sample of 3,108,000 words was analysed, distributed as indicated in Table 1.[2]

Table 1. Sources and distribution of the corpus texts from LOB, Brown, FLOB and FROWN, BE06, AmE06

Fictional categories	**Samples**	**Words**	**Non-fictional categories**	**Samples**	**Words**
Adventure and Western	174	348,000	Official Documents[3]	180	360,000
Mystery and Detective	144	288,000	Press Editorial	162	324,000
Romance and Love Story	174	348,000	Belles-lettres and Biographies	456	912,000
General Fiction	174	348,000			
Science Fiction	36	72,000			
Humour	54	108,000			
Total	756	1,512,000	**3,108,000 words**	798	1,596,000

4.2 Do so *constructions in British and American English written texts*

The analysis of LOB, Brown, FLOB, FROWN, BrE 2006 and AmE 2006 yielded a total of 861 instances of *do so* constructions. From a synchronic perspective, and on the basis of these corpora, no dramatic discrepancies can be found in the distribution of these syntactic constructions in British (465 instances) and American English (396 instances) written discourse, as illustrated in Table 2.

Table 2. Distribution of *do so* constructions in British and American English

Fictional categories	**British English**	**American English**	**TOTAL**
Official Documents	96	60	156
Press Editorial	54	51	105
Belles-lettres, Biographies, Essays	123	144	267
Adventure and Western	36	18	54
Mystery and Detective	30	42	72
Romance and Love	51	48	99
General Fiction	48	18	66
Science Fiction	9	9	18
Humour	18	6	24
TOTAL	**465**	**396**	**861**

These results seem in line with Algeo (1988: 2) who claims that syntactic differences between the two varieties of English are "the least numerous, the least salient, and the least confusing to speakers of one variety encountering a text composed in the other variety". Likewise, as Biber (1987: 116) observes, "linguistic differences [especially syntactic differences] among genres [or textual categories] are likely

to be larger than differences between British and American English of the same genre". In contrast to dialectal differences in pronunciation or vocabulary, for instance, syntactic constructions such as *do so* clauses are not likely to be perceived as Briticisms or Americanisms. This indeed seems true about these constructions here in British and American English: they occur in both varieties, they do not differ significantly in frequency, and hence they do not mark either variety as such.

Further evidence for the uniformity in the use of *do so* constructions in British and American English is found in the diachronic analysis of their distribution. As illustrated in Figure 1, the data show that there are very slight differences in the distribution of the construction in British and American English in the three periods under consideration – i.e. 1960, 1990, and 2006 – and prove that such differences are mitigated through time. The equal number of *do so* occurrences (117 tokens) in British and American English in the 2000s clearly suggests that the use of such clauses is not a matter of dialectal convention or interspeaker variation. Rather, as will be noted, the frequency and distribution of *do so* clauses in written English is related to the types of texts in which they occur and in the pragmatic function these constructions serve in discourse.

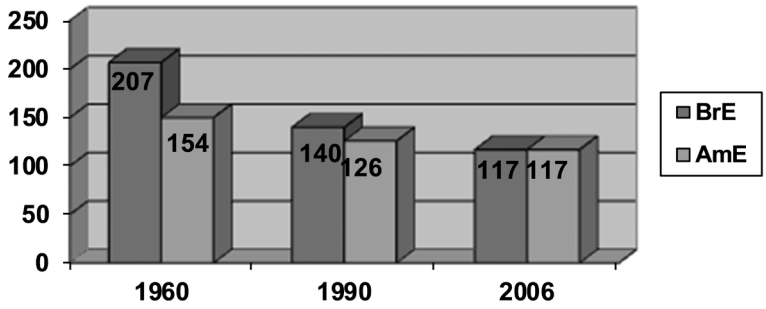

Figure 1. Diachronic distribution of *do so* clauses in British and American English in the 1960s, the 1990s, and the 2000s

4.3 Do so *constructions in fictional and non-fictional written English*

As can be seen in Table 3, *do so* clauses are more frequently attested in non-fictional discourse (528 instances / normalised frequency per 100,000 words: 33.01) than in fictional discourse (333 instances / normalised frequency per 100,000 words: 22.02), for reasons which will be explained presently.

Table 3. Normalised and raw distribution of *do so* constructions in the fictional and non-fictional categories

Fiction	**Raw**	**Normalised**	**Non-fiction**	**Raw**	**Normalised**
Adventure and Western	54	15.5	Belles-lettres, Biographies, Essays	267	29.2
Mystery and Detective	72	25	Press Editorial	105	32.4
Romance and Love	99	26.4	Official Documents	156	43.3
General Fiction	66	18.9			
Science Fiction	18	25			
Humour	24	22.2			
TOTAL	**333**	**22.02**		**528**	**33.1**

This higher frequency is even more notable in the individual analysis of the different categories in the corpora. As illustrated in Table 3, frequency of occurrence is consistently higher in non-fictional than in fictional categories, with *Official Documents* (43.4) and *Press Editorial* (32.4), showing the highest frequencies.

On the same lines, Kjellmer (1998: 160) points to a sharp divide between the same fictional and non-fictional categories, a distinction which can be interpreted in terms of the dichotomy formality vs. informality. The non-fictional texts from LOB, FLOB, FROWN, Brown, BrE 2006, and AmE 2006 are considered to be more formal in nature that the fictional texts. Moreover, the different non-fictional categories can be said to represent different degrees of formality. For instance, among formal text-types, it is possible to distinguish between highly formal categories, such as *Official Documents*, and relatively less formal ones, such as *Press Editorials* and *Belles-lettres*, *Biographies and Essays.*

Given this information, it would be tempting to argue that the distribution of *do so* constructions in the written genres analysed here is related to the degree of formality of the texts in question: those texts which are more formal – that is, the non-fictional texts – would favour the use of these constructions. Such a tendency has also been noted by Macía-Vega and Payne (2007) who, in their analysis of the distribution of different verbal anaphors in the FLOB corpus, find *do so* to be markedly more frequent in formal text-styles, and also by Miller (2013: 133), who claims that "*do s*o might overall be interpreted as a marker of higher register and thus an example of 'good speech'." However, in order to examine the exact reasons for the marked difference in distribution seen in Table 3, a more fine-grained analysis of the distribution of *do so* constructions in both genres is required, and will be provided in what follows. As will be noted below, this textual analysis has not been done randomly but, rather, is based on tested criteria of linguistic variation.

5. *'Do so'* constructions: Textual dimensions and relations

Biber (1988) analyses linguistic variation in the textual categories of the LOB and the Brown corpora. As pointed out in section 4, these two corpora were compiled in the 1960s and match the structure of FLOB, FROWN, BrE 2006 and AmE 2006 (for details see Hofland et al. 1999, Mair 2002). Studies prior to that of Biber analysed linguistic variation in terms of single parameters; for example, texts were traditionally considered to be related according to isolated parameters such as formal/informal, interactive/non-interactive, literary/colloquial, or restricted/elaborated. By contrast, Biber argues that linguistic variation is too complex to be analysed in terms of any single dimension, and claims that the relations among texts cannot be defined unidimensionally because comparison of texts with respect to any single dimension gives way to incomplete and sometimes misleading text typologies. Biber's work, in fact, confirms that most texts must be seen as multidimensional and not as pure text types.

The textual categories comprised in the LOB and BROWN corpora are analysed by Biber in terms of six parameters or dimensions. Dimension 1, which he labels *Involved versus Informational Production*, distinguishes discourse with interactional, affective or involved purposes and which is associated with strict real-time production and comprehension constraints, from discourse with highly informational purposes. Dimension 2, *Narrative versus Non-narrative Concerns*, distinguishes discourse with primary narrative purposes from discourse with non-narrative purposes, hence dealing with the difference between active, event-oriented discourse and more static descriptive or expository types of discourse. Dimension 3, *Endophoric versus Situation-Dependent Reference*, distinguishes between discourse that identifies referents fully and explicitly through relativisation, and discourse that relies on non-specific deictics and reference to an external situation for identification purposes. This dimension thus corresponds closely to the distinction between endophoric and exophoric reference (cf. Halliday and Hasan 1976). Dimension 4, *Overt Expression of Persuasion*, refers to those features associated with the speaker's expression of point of view or with argumentative styles intended to persuade the addressee. Dimension 5, labelled *Abstract versus Non-abstract Information*, distinguishes between texts with a highly abstract and technical informational focus and those with non-abstract focus. Finally, dimension 6, *On-line Informational Elaboration*, distinguishes between informational discourse produced under highly constrained conditions in which the information is presented in a relatively loose, fragmented manner, and other types of discourse, be it informational discourse that is highly integrated or discourse that is not informational in nature.

In addition to multidimensionality, variation is treated as continuously scalar in Biber's analysis. The six parameters, then, define continua of variation rather than discrete poles. For example, although it is possible to describe a text as simply abstract or non-abstract, it seems more accurate to describe it as more or less abstract. The similarities and differences among textual categories can therefore

be considered with regard to each of the six dimensions mentioned above. Some textual categories can be similar with respect to some dimensions but quite different with respect to others.

In the last thirty years, Biber's (1988) multidimensional analytical framework has been regarded a powerful tool for approaching register variation and genre analysis. Its results have been considered tested criteria of linguistic variation and they have allowed linguists to investigate language in use and to formulate detailed descriptions, which in turn encapsulate how language users make concrete language choices in particular linguistic contexts.

What follows provides a comparison of the distribution of *do so* constructions in the textual categories of the corpora analysed here with Biber's analysis of the same categories in terms of different dimensions of linguistic variation. If the distribution of *do so* clauses is sensitive to any of these linguistic dimensions, we can assume that this should be seen clearly in the present study. As will be shown, this will certainly be the case regarding the distribution of *Do so* in relation to dimensions 2, 3 and 5.

5.1 Do so *constructions in narrative and non-narrative written English*

The fictional categories analysed here are highly narrative in nature (cf. Hofland et al. 1999). Biber (1988: 137) has shown that the narrative texts included in the categories analysed here differ from the non-narrative ones in that they do not normally "include the presentation of expository and procedural information or the description of actions actually in progress". The non-fictional texts, by contrast, are highly nominal in nature: they are descriptive and argumentative rather than verbal or narrative (cf. Biber 1988: 140).

As illustrated in Figure 2 below, the fictional categories, namely *Humour*, *Adventure and western*, *Science fiction*, *Mystery and detective*, *General fiction*, and *Romance and love story* score high on the narrative pole of dimension 2 in Biber's analysis. By contrast, the non-fictional categories, *Belles-lettres, biographies and essays*, *Official documents*, and *Press editorial* have much lower scores on this dimension.

Following Biber (1988), textual categories with high scores on dimension 2 typically exhibit a high incidence of past tenses, perfect aspect verbs, third person pronouns, communication verbs (e.g. *say*, *discuss*, *explain*, *suggest*), present participial clauses and synthetic negation, together with markedly infrequent occurrences of present tense verbs. On the contrary, textual categories which rank low on the narrative pole of dimension 2 have the opposite characteristics. As Biber (1988) asserts, "the large separation of the fiction genres from all other genres indicates that the proposed interpretation of a narrative versus non-narrative dimension is an accurate description underlying the function here" (1988: 137; emphasis added).

+ NARRATIVE

7- **Romance and love story**
6- **Mystery, science** and **general fiction**
Adventure and western, Humour
5-
4-
3-
2- **Belles-lettres, Biographies and Essays**
1- Prepared speeches
Personal letters
0- Press reportage
Popular lore

Face-to-face conversation
Religion; **Press editorials**
-1- Interviews

Press review

-2- Telephone conversations
Professional letters
Academic prose
Official documents
-3- Hobbies
Broadcast

- NARRATIVE

Figure 2. Mean scores of dimension 2 – Narrative vs. Non-narrative Concerns – in Biber (1988: 136; emphasis added)

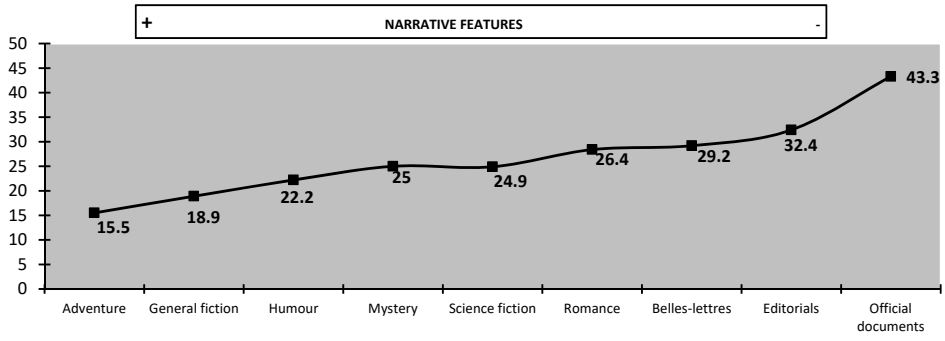

Figure 3. Normalised distribution of *do so* constructions and the degree of narrative features in the textual categories.

On the basis of the corpus analysis, *do so* clauses can be considered constructions that are disfavoured in narrative text-styles. The comparison of Biber's findings for the narrative or non-narrative features of texts (cf. Figure 2, and Biber 1988: 122–124) with the data retrieved from the corpora (cf. Figure 3) confirms that

there is a tendency for those categories with a higher degree of narrative features, that is, namely *Humour*, *Adventure and western*, *Science fiction*, *Mystery and detective*, *General fiction*, and *Romance and love story* to disfavour the use of this construction. By contrast, categories which are more non-narrative in nature, such as *Belles-lettres, biographies and essays*, *Official documents*, and *Press editorial* favour the use of *do so* clauses.

As shown in Table 4, this tendency is seen even more clearly if we measure the correlation between the mean scores on Dimension 2 and the normalized frequencies of *do so* constructions by calculating a (Pearson) correlation coefficient.[4] The result, which is significant even at the p<=0.001 level, is *-0.81048* and confirms that the more narrative oriented a text is, the less *do so* constructions are to be expected.[5] The frequency of *do so* constructions is therefore dependent on the narrative nature of the text in which these constructions occur. As will be explained in what follows, the reasons for this dependence are motivated by the type of anaphoric function performed by these constructions in discourse.

Table 4. Pearson correlation coefficient for the distribution of *do so* clauses and Biber's (1988: 122–124) mean scores on Dimension 2

	Mean scores of selected categories on Dimension 2	**Normalised frequencies for *do so* clauses**
Official Documents	-2.9	43.3
Press Editorial	-0.8	32.4
Belles-lettres, Biographies, Essays	2.1	29.2
Adventure and Western	5.5	15.5
Mystery and Detective	6	25
Romance and Love Story	7.2	26.4
General Fiction	5.9	18.9
Science Fiction	5.9	24.9
Humour	5.8	22.2
CORRELATION COEFFICIENT	**-0.81048**	

5.2 Do so *constructions and the degree of endophoric reference*

Taking dimension 3 into account, *Explicit versus Situation-Dependent Reference*, Biber (1988:142) shows that the texts in the category of *Official documents* exhibit a high degree of explicitness and text-internal reference, whereas the categories of *Belles-lettres* and *Press editorial* show intermediate values, and the fictional categories – i.e. *Adventure and Western*, *Romance and Love Story*, *General Fiction*, *Mystery and Detective, Humour* and *Science Fiction* – rely more on situation-dependent or exophoric reference with a lower degree of explicitness. This is illustrated in Figure 4, where the categories *Official documents, Belles-lettres* and *Press editorial* score higher on the explicit/endophoric reference pole of dimension 3 than the fictional categories.

+ EXPLICIT TEXT-INTERNAL REFERENCE (+ENDOPHORIC REFERENCE)

7-	**Official documents**
	Professional letters
6-	
5-	Science; Press review
4-	Religion
3-	Popular lore
2-	
	Press editorials; Belles-lettres, biographies, essays
1-	Spontaneous speeches
	Prepared speeches; Skills, trades, hobbies
0-	
	Press reportage
-1-	**Humour**
	Science Fiction
-2-	
-3-	
	General fiction,
	Mystery and Adventure fiction
-4-	Face-to-face conversations; **Romantic fiction**
-5-	Telephone conversations

– EXPLICIT TEXT-INTERNAL REFERENCE (+SITUATION-DEPENDENT)

Figure 4. Mean scores of dimension 3 – Endophoric vs. Situation Dependent Reference – in Biber (1988:143; emphasis added)

As discussed in Biber (1988: 142–148), texts ranking high in this dimension are characterized by explicit, elaborated and endophoric reference. A strong correlation is then expected between *do so* as an anaphoric device and categories with high scores in Dimension 3, in line with what has previously been observed (Souesme 1987, Culicover and Jackendoff 2005, Stirling and Huddleston 2002), and this is precisely what is found.

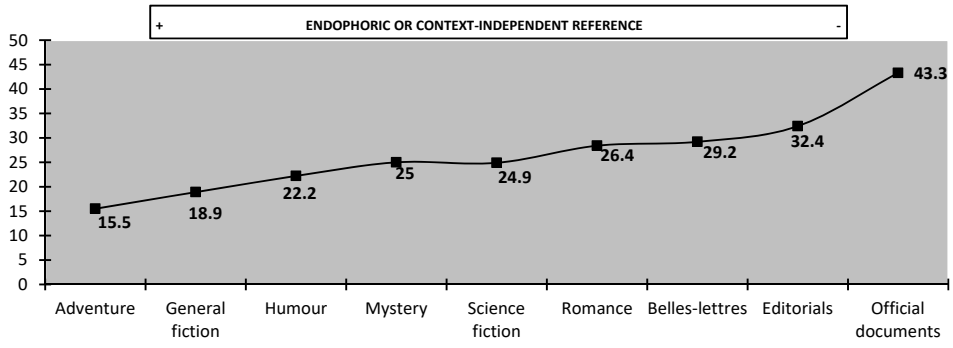

Figure 5. Normalised distribution of *do so* constructions and the degree of endophoric reference in the textual categories

As illustrated in Figure 5, the corpus-based results show that the frequency *do so* is higher in *Official Documents* (43.3), *Editorials* (32.4.) and *Belles-lettres* (29.2) than in the fictional categories, namely *Romance and Love Story* (26.4), *Mystery and Detective* (25), *Science Fiction* (24.9), *Humour* (22.2), *General Fiction* (18.9) and *Adventure and Western* (15.5).

The comparison of Biber's findings for the endophoric or situation-dependent referential nature of texts (cf. Figure 4) with the distribution of the *do so* construction in the corpora analysed here (cf. Figure 5) suggests that there is a tendency for those categories with a higher degree of endophoric or "context-independent" reference (cf. Biber 1995:156), namely *Official Documents*, *Press Editorial*, and *Belles-lettres* to favour the use of the construction. By contrast, the *do so* construction is less favoured in texts, such as the fictional texts, which are more situationally-based and contain "direct reference to the physical and temporal situation of discourse" (Biber 1988: 145).

This is even more clearly noticed if we measure the correlation between Biber's mean scores on Dimension 3 and the normalised frequencies for the distribution of *do so* structures in the corpora texts by calculating a (Pearson) correlation coefficient, as illustrated in Table 5. The result is *0.85117* which, again, is significant even at the p<= 0.001 level. The distribution of *do so* constructions in the textual categories analysed here therefore seems to be related to the degree of endophoric reference of its texts: the more endophorically oriented the text, the more *do so* clauses are to be expected.

Table 5. Pearson correlation coefficient for the distribution of *do so* clauses and Biber's (1988: 122–124) mean scores on Dimension 3

	Mean scores of selected categories on Dimension 3	**Normalised frequencies for *do so* clauses**
Official Documents	7.3	43.3
Press Editorial	1.9	32.4
Belles-lettres, Biographies, Essays	1.7	29.2
Adventure and Western	-3.8	15.5
Mystery and Detective	-3.6	25
Romance and Love Story	-4.1	26.4
General Fiction	-3.1	18.9
Science Fiction	-1.4	24.9
Humour	-0.8	22.2
CORRELATION COEFFICIENT	**0.85117**	

Biber's work shows that categories that score highly on Dimension 3 characteristically make frequent use of WH relative clauses, pied-piping constructions, phrasal coordination, and nominalizations, that is, constructions whose function is to "explicitly identify referents or to provide elaborating information concerning referents" (cf. Biber 1995:157): relative clauses, for instance, pack informa-

tion into noun phrases instead of expressing the information as separate independent clauses, WH relative clauses explicitly identify nominal referents, and phrasal coordination (as in "the clowning and the prettiness") allow for the packing of large amounts of information into phrases or clauses.

On the basis of the present corpus-based analysis, *do so* clauses can also be considered constructions which are used to create a highly wrought and explicit textual reference in the non-fictional categories, and allow for the packing of large amounts of information. As pointed out in section 2, the complex pro-form *do so* serves as an anaphoric verb phrase and is used to perform a text-structuring function. In these types of clauses, anaphoric *so*, which stands for given information in discourse, functions syntactically as complement or as adjunct and its interpretation can only be determined via the antecedent, as illustrated in (16) and (17) below.

(16) With his four interestingly diverse upper-class individuals he was able to construct at least the scaffolding of the larger entity behind and around them, and do so while minding his p's and q's as a biographer, not as a sociologist

(FROWN, Belles-lettres, Biographies, Essays. G21)

(17) I fully understand and appreciate your desire not to give reasons in general, but on this occasion you might consider it worth your while to do so.

(LOB, Official Documents. H19)

In example (16), an anaphoric link between *so* and the preceding clause is expressed, and *so* is interpreted as "*being able to construct at least the scaffolding of the larger entity behind and around them"*, which is its antecedent. In other words, *do so* expresses a parallelism with the preceding clause and establishes a comparison between two referents. Example (16) implies that the information coded by the antecedent in the first clause is also applied to the second clause. Similarly, example (17) entails that "*to give reasons in general*" is "*worth your while on this occasion*". Also in this case, the *do so* construction is used in the second clause to express that the same type of event has occurred as that expressed in the first clause. *So* has a cohesive effect since it stands for given information and, as Potts (2002: 640) notes, it "adjoins directly to the linguistic material from which it obtains its meaning [its antecedent]".

The fact that the *do so* construction and its text-structuring function is far more frequent in non-fictional than in fictional text-styles shows that, in non-fictional discourse, it is an important construction for making clauses fit with the explicit and elaborated discourse context. The category of *Official documents*, for instance, contains texts that are highly informative and abstract in nature (cf. section 5.3) which, as Biber (1988: 132) notes, are produced under conditions permitting careful word choice. Although official texts are highly informational, they tend to contain particularly long utterances and need considerable lexical repetition of vocabulary

because of the exact technical meanings associated with particular terms, as shown in (18) below: i.e. *officer*, *official*, *agency*, *organization*, *Legislative Research Commission*, *subcommittee*, *task force*, *fees*, *charges*, *cost*, etc. In contrast to lexical repetition, the *do so* represents a strategy for avoiding the repetition of identical verb phrases: *to furnish information* in example (18). It is precisely this grammatical reduction that allows for the packing of information in non-fictional discourse and builds up a coherent text that eases the reading process for the receiver.

(18) 7.112 Information to be provided free of charge.
Any public officer or official, agency, or organization of state or local government required or requested to furnish information or data to the Legislative Research Commission, a subcommittee, task force, or other body associated therewith shall do so without any fees or charges of any kind for the information or data or any cost associated with its gathering, processing, or production.
Effective: July 15, 1988
History: Created 188 Ky. Acts ch. 231, sec. 2, effective July 15, 1988.
(AmE 2006, Official Documents. H05)

The fictional texts, by contrast, show a lower proportion of *do so* clauses but rather make use of other verbal anaphors, namely *do it*, *do this*, *do that* (also cf. Miller 2011: 2), with an inherent deictic meaning. This was demonstrated by Biber et al. (1999: 432), whose corpus-based analysis shows that pro-verb *do* combined with a following pronoun *it* is more common in conversation and fiction than in non-fictional genres such as news reports or academic prose. In speech, the preponderance of *do it* takes place because spoken texts are coupled with online production needs where speakers can rely on the shared situation together with the possibility for immediate clarification to identify the implied meaning, as illustrated in (19). This example clearly shows the extreme reliance on implicit meaning retrieved from the context by both addressor and addressee in conversation.

(19) A: Well I haven't got one of those.
B: Yes.
C: Go.
B: Come on. You **do it** too
A: No, you **do it**.
B: No, you **do it**.
C: I've got a wicked joke
C: Go.
[Taken from Biber et al. 1999: 432]

Dependence on the context is also found in the fictional texts analysed here. Further, in fictional texts, as shown in (20), conversational features may also be natural when the addressor writes as if he or she were speaking, or wants to reflect

direct speech situations, and therefore a frequent use of the complex pro-form *do it* is also attested. In both conversation and fiction *do it* can be identified as a substitute for a lexical verb, *to ring* in example (21), or can function as a substitute for a series of actions or events, as in (20) where the pro-verb expression *does it* refers to different actions such as *speaking, looking, smiling* and *frowning.*

(20) "Oh, he knows," she replied acidly. "He doesn't speak – won't even answer direct questions. Never looks you in the eye. Never smiles. Never frowns. No expression at all. He does it on purpose just to infuriate us."
(BrE 2006, Romance and Love Story. P26)

(21) Still in her coat, Mary stood up and went into the living-room to the telephone. Most of the people she rang had been checked already but Alan let her do it.
(LOB, Romance and Love Story. P26)

Similarly, as demonstrated by Biber et al. (1999: 432), pro-verb *do* followed by the demonstratives *this* or *that* is commonly attested in conversation and fiction, cf. (22)-(23), and rarely found in non-fictional texts such as news reports or academic prose.

(22) "You two sure know how to gang up on a guy. Now listen. I'm going to spend my time with the corporate folks, and you've got to try to meet the guy who heads up the insurance defence group, right? She caught Joel's anxious glance." –"I can do this**,** I really can."
(FROWN. Mystery and Detective. L24)

(23) "This is my uncle, Mr. MacNally, and this my nephew, Michael."
The hand came past her and rested on the boy's head.
Before she could acknowledge the introductions he went on:
"Take Miss Metcalfe up to the house, Uncle Shane."
"Aye, Ralph. Yes I'll do that."
(LOB. Romance and Love Story. P21)

This and *that* can serve as either deictic or anaphoric markers in discourse, as shown in (24) and (25) below. In (24), *that* will be interpreted deictically as referring to some object present in the fictional situation, whereas in (25) *that* obtains its interpretation anaphorically, from the antecedent *$40*. It has been shown (Huddleston and Pullum 2002: 1532) that the deictic use of *this* and *that* carries over to their use in combination with *do*, as illustrated in (26), with *do that* denoting the action previously performed: *biting his hand off at the wrist*. Beyond that, the deictic and anaphoric uses of these demonstratives are clearly related, and it is plausible to regard the anaphoric use as derivative from the deictic, and in fact it "retains some residual deictic meaning" (cf. Huddleston and Pullum 2002: 1455).

(24) With an unabashed curiosity he took a mental inventory of the room: its lighting, its shelves, its chairs, its pictures, the jumble of knick-knacks along the mantelpiece; then started on a tour of investigation, taking up a book, peering into an etching, lifting a cigarette-box; without comment, as though he were visiting an exhibition, till suddenly, with a note of real interest in his voice, "What's that doing here?" he asked.
(LOB. Romance and Love Story. P19)

(25) The check is $40. That's too much.
(FROWN. Mystery and Detective. L13)

(26) And the fact remains that when Johnny tried to round him up, Ellaway bit his hand off at the wrist. Most lunes don't do that. Lunes aren't usually savage enough to hurt you that badly before you get them tranked.
(AmE 2006, Adventure and Western. N29)

The deictic meaning discussed above is a characteristic feature of conversation and of fiction that shows a particular concern for time and place (cf. Huddleston and Pullum 2002: 548). These fictional and spoken texts encourage reference to the physical and temporal situation of discourse (cf. Biber 1988: 145) and contain situated texts that depend directly on direct reference and knowledge of time and place for their understanding. Conversational language, in particular, typically involves participants who share the same location in space and time, and who alternate in their roles (cf. Chafe 1994: 44). In speech, exophoric reference is generally present since oral communication normally deals with events actually in progress, where the hearer is forced to construct a mental map of the situation in order to understand the text (cf. Biber et al. 1999: 1042–1044). Speech therefore shows a strong preference for the use of spatial and temporal phrases and deictic elements conveying a locative meaning.

Descriptions with time and place deictics are also particularly common in fiction, where scenes are constantly introduced. Locativity is indeed inherent to fiction and spatial and temporal reference is not an optional or peripheral feature of narration but a core property that helps constitute narrative domains (cf. Herman 2001). However, the case of fiction is slightly different from that of conversation since reference is made to text-internal physical and temporal situations. Formally speaking, this reference seems exophoric because it refers directly to the situation of events (cf. Biber 1988: 148) but, in the case of fiction, there is a fictional situation that is referred to directly in the text. The reader understands this exophoric reference in terms of the physical and temporal situation developed in the fictional context within the text. In fictional texts, the referent is therefore not physically present in the situation of utterance but is located in the discourse context itself. This *textual deixis* or *discourse deixis*, to use Huddleston and Pullum's (2002: 1460) terminology, also has to do with "using the deictic procedure to point to part of a pre- or post- existing textual or memory representation" (cf.

Cornish 2006: 633), and is one of the basic properties that most clearly separates the fictional and non-fictional genres analysed here.

5.3 Do so *constructions and the degree of abstractness*

Biber's (1988) investigation demonstrated that the textual categories included in the present investigation differ in their degree of abstractness. His findings show that the non-fictional categories – namely *Official Documents*, *Press Editorials* and *Belles-lettres, Biographies and Essays* – have higher scores on the abstract pole of dimension 5 than the fictional categories (cf. Figure 6). *Official Documents*, for instance, are markedly constrained in linguistic form since most of these texts involve the discussion of abstract or impersonal topics.

+ ABSTRACT

	Science
5-	
	Official Documents
4-	
3-	
2-	
	Religion
	Hobbies
1-	
	Press reviews
	Press reportage
	Professional letters; **Press editorials**
0-	Popular lore
	Humour; **Biographies, belles-lettres, Essays.**
-1-	
	Broadcasts
-2-	Prepared speeches; interviews
	General, science, and adventure fiction
	Mystery fiction; Romantic fiction
-3-	
	Personal letters; face-to-face conversations
	Telephone conversations

– ABSTRACT

Figure 6. Mean scores of dimension 5 – Abstract vs. Non-abstract Information – in Biber (1988: 152; emphasis added)

Surprisingly enough, the comparison of Biber's findings for the abstract and non-abstract features of texts (cf. Figure 6, and Biber 1988: 122–124) with the distribution of *do* so in the corpora (cf. Figure 7) shows that there is a tendency for those categories with a higher degree of abstractness or technical focus, that is,

Official Documents, *Press Editorials* and *Belles-lettres*, *Biographies and Essays*, to favour the use of this construction. By contrast, the fictional categories, which are less abstract in nature, make less use of *do so* clauses.

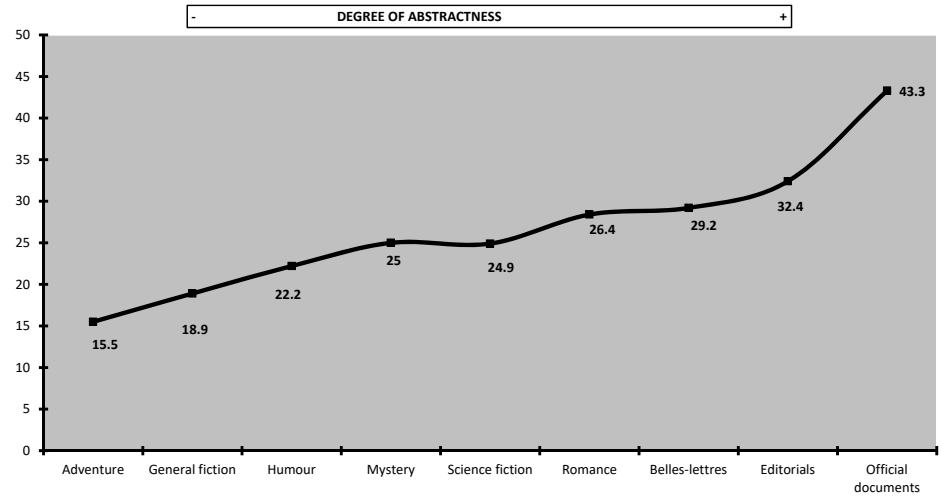

Figure 7. Normalised distribution of *do so* constructions and the degree of abstractness in the textual categories

As shown in Table 5, this tendency can be even more clearly noted if we measure the correlation between the mean scores on Dimension 5 and the normalized frequencies of *do so* constructions by calculating a (Pearson) correlation coefficient. The result, which is significant even at the p<=0.001 level, is *0.81149* and shows that the more abstract a text is, the more *do so* constructions may be expected.

Table 5. Pearson correlation coefficient for the distribution of *do so* clauses and Biber's (1988: 122–124) mean scores on Dimension 5

	Mean scores of selected categories on Dimension 2	**Normalised frequencies for *do so* clauses**
Official Documents	4.7	43.3
Press Editorial	0.3	32.4
Belles-lettres, Biographies, Essays	-0.5	29.2
Adventure and Western	-2.5	15.5
Mystery and Detective	-2.8	25
Romance and Love Story	-3.1	26.4
General Fiction	-2.5	18.9
Science Fiction	-2.5	24.9
Humour	-0.4	22.2
CORRELATION COEFFICIENT	**0.81149**	

On the basis of the corpus analysis, *do so* clauses can be considered constructions that are favoured in abstract text-styles. These results, nevertheless, are surprising since it has been shown (cf. Biber 1988: 111) that texts with a high degree of abstractness make use of syntactic constructions such as passives, constructions, adverbial past participial WHIZ deletions, and adverbial subordinators or conjuncts, which are used to present an open proposition with reduced emphasis on the agent. Agentless passives, for instance, are commonly used in procedural discourse, where the same agent is presupposed across several clauses and the specific agent of a clause is not important to the discourse purpose and is consequently not mentioned. In other words, these constructions are used to give prominence to the patients of the verb. *Do so* clauses, by contrast, do not share this function, that is, they do not reduce the emphasis on the agent and their presence in abstract texts should be expected to be less frequent overall.

A possible explanation for the strong correlation between the frequency of *do so* anaphora and the abstract texts analysed here is that *do so* anaphora can stand for a string of constituents from the verb phrase, ranging from entire verb phrase (cf. 27) to a complement of the verb (cf. 28). Hence, it does not subordinate the subject per se but places emphasis on the constituents of the verb phrase rather than the subject. In other words, the semantics and structural patterns of *do so* anaphora allow the addressor to give prominence to the constituents of the verb phrase, which are presented again in discourse and are grammatically reduced in form in order to ease the reading process for the addressee.

(27) Finally, whether petitioners are right or wrong, our prompt review will diminish the legal uncertainty that now surrounds the application to Guantanamo detainees of this fundamental constitutional principle. **Doing so** will bring increased clarity that in turn will speed review in other cases. (AmE 2006, Official Documents. H16)

(28) Given the urgent need for concrete measures to support biomass heat, we should not have to wait until 2007 for the Biomass Strategy, and recommend that the Government make clear in its response exactly when it anticipates publishing this strategy, and further suggest that it **does so** at the earliest possible opportunity.
(BrE 2006, Official Documents. H06)

It has been shown that abstract discourse is not very much concerned with spatial and temporal descriptions, nor is it very much concerned with the agent, but it does put a premium on explicitness of cohesion, which may be enhanced by the use of *do so* anaphora. In (28), for instance, the *do so* construction represents an element of comparison with the preceding clause: it points anaphorically to the antecedent: *publishing this strategy*. The use of *do so* anaphora directs the addressee's attention towards the information provided in the verb phrase of the preceding clause, and allows a semantic repetition that cohesively ties the sentences together.

6. Summary and conclusions

In the long history of the analysis of *do so* verbal anaphors, it has been argued that *do so* constructions are semantically and grammatically restricted. It has been claimed, for instance, that *do so* is only compatible with antecedents that denote events and not states (cf. Kehler and Ward 1999, among others) or that the use of *do so* is dependent on the voice or category differences between the antecedent and the *do so* clause (cf. Stirling and Huddleston 2002). Most of the literature on *do so* has dealt with these and other semantic and syntactic factors that drive the use of the construction although there are also studies of the construction in real data (cf. Kehler and Ward 1999, Houser 2010, or Miller 2013).

The present paper is a further contribution to this line of research and has provided a corpus-based analysis of *do so* in written English. It has shown that, in addition to semantic and grammatical factors, genre variation also plays an important role in the use of the construction. Previous corpus-based analyses of *do so* have argued that it is more typical of formal registers (cf. Macía-Vega and Payne 2007 or Miller 2013) and that it is more frequently attested in non-fictional text-styles than in fictional text-styles and spoken texts (see Miller 2011). However, the more fine-grained corpus-based analysis of *do so* – in specific fictional and non-fictional textual categories – provided here has shown that the use of the construction is sensitive to more detailed dimensions of textual variation.

The analysis has shown that *do so* is at home in non-fictional text-styles, where it performs a text-structuring function and is an important construction for making clauses fit with explicit and elaborate discourse contexts. Beyond this, the comparison of the distribution of *do so* in the fictional and non-fictional texts analysed here with Biber's (1988) dimensions of linguistic variation has shown that there is a tendency for *do so* to be more frequent in those textual categories which exhibit a higher degree of endophoric reference: the more endophorically oriented the text, the more *do so* clauses are to be expected. Related to this, the statistical results have also shown that the distribution of *do so* in written English is dependent on the narrative nature of the texts in which these constructions occur. The data showed that the more narrative oriented a text is, the less *do so* constructions are to be expected. It has been argued that this is mainly a consequence of the anaphoric function performed by *do so*, which is used to code endophoric rather than exophoric reference. Narrative discourse is quite dependent on exophoric reference – that is, on the context of the situation and on spatial deictic references – and makes use of other types of verbal anaphors, such as *do it*, *do this* or *do* that (cf. Biber et al. 1999:432; Miller 2011:2), which better suit the deictic association of this type of discourse. Finally, the corpus has also shown that the more abstract or technical is a text, the more *do so* constructions are to be expected. This seems to be a consequence of the fact that, although it does not subordinate the subject per se, *do so* clauses place emphasis on the constituents of the verb phrase rather than the subject, following the principle of end-focus (cf. Halliday 1967).

The findings of this study have shown that genre variation is an important factor that strongly influences the use of *do so* anaphora in written English, and raise the question as to whether it also does so in the spoken language. Further corpus-based studies need to be undertaken to provide the empirical base necessary for the analysis of the general properties of *do so* and its distribution in the spoken mode. A full understanding of *do so* can only merge from a comprehensive understanding of the construction in both the written and spoken language, and the present study has been a first step in that direction.

Notes

[1] Notice that, as Houser (2010: 7) points out verb phrase ellipsis poses not such restriction on the antecedent, as in (i) below:

(i) I know the Easter Bunny is not real, and kent does too.

[2] The examples have been retrieved from the corpora by using Wordsmith Tools. The concordance software allowed for a general search of all forms of *do* followed by *so*, and the non-relevant occurrences, i.e. examples like "Kim did so many things", were then manually filtered out by hand.

[3] The *Official Documents* textual category comprises government documents, institutional reports, industry reports, college catalogues and in-house industry texts.

[4] The Pearson Correlation is a precise measure of the way in which two variables correlate. Its value indicates both the direction (positive or negative) and the strength of the correlation between two variables. The value +1 indicates a perfect positive correlation and the value –1 a perfect negative correlation, whereas a value of 0 indicates no correlation at all (cf. Butler 1985, Baayen 2008, and Johnson 2008, among others).

[5] In statistics, the *p-value* or "statistical significance" of a result is the probability that the observed relationship between variables in a sample occurred by pure chance. Results that are significant at the $p \leq .01$ level are commonly considered statistically significant, and $p \leq .005$ or $p \leq .001$ levels are often called "highly" significant.

Acknowledgement

I am grateful to the Spanish Ministry of Economy and Competitiveness for generous financial support (National Programme for Excellence in Scientific and Technical Research; grant FFI2014-52188-P).

References

Algeo, John (1988) 'British and American grammatical differences'. *International Journal of Lexicography* 1 (1), 1–31.

Atkinson, Dwight and Douglas Biber (1994) 'Register: a review of empirical research'. In: Biber, Douglas and Edward Finegan (eds.) *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press, 351–385.

Baayen, Rolf H (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge: Cambridge University Press

Baker, Paul (2009) *Contemporary Corpus Linguistics*. London: Continuum.
Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (1999) *Longman Grammar of Spoken and Written English*. London: Longman.
Biber, Douglas (1987) 'A textual comparison of British and American writing'. *American Speech* (62), 99–119.
Biber, Douglas (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
Biber, Douglas (1992) 'Using computer-based text corpora to analyze the referential strategies of spoken and written texts'. In Svartvik, Jan (ed.) *Directions in Corpus Linguistics*. Berlin: Mouton, 213–252.
Biber, Douglas (1995) *Dimensions of Register Variation: A Cross-linguistic Comparison.* New York: Cambridge University Press.
Bolinger, Dwight (1970) 'The meaning of do so'. *Linguistic Inquiry* 1, 140–144.
Bos, Johan and Jennifer Spenader (2011) 'An annotated corpus for study of VP ellipsis'. *Language Resources and Evaluation* 45, 463–494.
Bouton, Lawrence (1970) 'Do So: Do + adverb'. In: Sadock Jerrold and Anthony Vanek (eds) *Studies Presented to Robert B. Lees*. Alberta: Linguistic Research Inc, 17–38.
Butler, Christopher (1985) *Statistics in Linguistics*. Oxford. Blackwell.
Chafe, Wallace L. (1994) *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.
Cornish, Francis (2006) 'Discourse anaphora'. In: Keith Brown (ed.) *Encyclopaedia of Language and Linguistics*. Oxford: Elsevier, 631–638.
Culicover, Peter W. and Ray Jackendoff (2005) *Simpler Syntax*. Oxford: Oxford University Press.
Davies, Mark (2008) *The Corpus of Contemporary American English: 450 million words, 1990-present*.
Fox, Barbara A. (1987) *Discourse Structure and Anaphora*. Cambridge: Cambridge University Press.
Halliday, Michael Alexander K. (1967) 'Notes on transitivity and theme in English: part 2'. *Journal of Linguistics* 3, 199–244.
Halliday, Michael Alexander K. and Ruqaiya Hasan (1976) *Cohesion in English*. London: Longman.
Hankamer, Jorge and Ivan A. Sag (1976) 'Deep and surface anaphora'. *Linguistic Inquiry* 7, 391–426.
Herman, David (2001) 'Spatial reference in narrative domains'. *Text* 21, 515–541.
Hofland, Knut, Anne Lindebjerg and Jørg Thunestvedt (1999) *ICAME Collection of English Language Corpora*. 2nd edition, CD-ROM version. Bergen: The HIT Centre.
Houser, Michael J. (2010) *The Syntax and Semantics of Do So Anaphora*. PhD dissertation, University of California. Berkeley
Huddleston, Rodney and Geoffrey K. Pullum (2002) *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Johnson, Keith (2008) *Quantitative Methods in Linguistics*. Oxford: Blackwell.
Kehler, Andrew and Gregory Ward (1999) 'On the semantics and pragmatics of identifier so'. In: Ken Turner (ed.) *The Semantics/Pragmatics Interface from Different Points of View* (Vol. I). Amsterdam: Elsevier, 233–256.
Kehler, Andrew and Gregory Ward (2004) 'Constraints on ellipsis and event reference'. In Horn, Laurence R. and Gregory Ward (eds.) *Handbook of Pragmatics*. Oxford: Blackwell, 383–403.
Kjellmer, Göran (1998) 'On contraction in modern English'. *Studia Neophilologica* 69, 155–162.
Kurzon, Dennis (1985) 'Signposts for the reader: A corpus-based study of text deixis'. *Text* 5(3), 187–200.
Lakoff, George and John Robert Ross (1976) 'Why you can't do so into the kitchen sink'. In: McCawley, James D. (ed*.) Syntax and Semantics 7: Notes from the Linguistic Underground*. New York: Academic Press, 101–111.

Lakoff, George (1966) *Stative Adjectives and Verbs in English*. Computational Laboratory, Harvard University.
Levin, Nancy S. (1986) *Main Verb Ellipsis in Spoken English*. New York: Garland.
Lord, Carol and Dahlgren, Kathleen (1997) 'Participant and event anaphora in newspaper articles'. In: Bybee, Joan, John Haiman, and Sandra A. Thompson (eds.) *Essays on Language Function and Language Type*, *Dedicated to Talmy Givòn*. Amsterdam: John Benjamins, 323–356.
Macía-Vega, Beatriz and John Payne (2007) 'Do So'. Talk delivered at the Second International conference on the Linguistics of Contemporary English. University of Toulouse-le Mirail.
Mair, Christian (2002) 'Three changing patterns of verb complementation in Late Modern English: a real time study based on matching corpora'. *English Language and Linguistics* 6, 105–131.
Michiels, Archibald (1978) 'A note on the relation between agent and stativity'. *Neophilologus* 62, 172–177.
Miller, Philip (1990) 'Pseudogapping and do so substitution'. In: Ziolkowski, Michael, Manuela Noske, and Karen Deaton (eds.) *Papers from the 26th Regional Meeting of the Chicago Linguistic Society.* Chicago: Chicago Linguistic Society, 293–305.
Miller, Philip (2011) 'The choice between verbal anaphors in discourse'. In: Hendrickx, Iris, Devi Sobha Lalitha, Antonio Branco and Mitkov Ruslan (eds.) *Anaphora Processing and Applications.* Berlin: Springer, 82–95.
Miller, Philip (2013) 'Usage preferences: The case of the English verbal anaphor do so'. In: Stefan Müller (ed.) *Proceedings of the 20th International Conference on Head-Driven Phrase Structure Grammar*. Stanford: Center for the Study of Language and Information Publications, 121–139.
Potts, Christopher (2002) 'The syntax and semantics of As-parentheticals'. *Natural Language and Linguistic Theory* 20(3), 623–689.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1985) *A Comprehensive Grammar of the English Language*. London: Longman.
Reppen, Ranci, Nancy Ide, and Keith Suderman (2005) *American National Corpus*. Linguistic Data Consortium.
Ross, John Robert (1973) 'Nouniness'. In: Fujimura, Osamu (ed.) *Three Dimensions of Linguistic Theory*. Tokyo: TEC Corporation, 137–258.
Souesme, Jean-Claude (1987) 'Valeurs et emplois respectifs de DO et DO SO'. *Modèles Linguistiques* 9, 65–92.
Stirling, Lesley and Rodney Huddleston (2002) 'Deixis and anaphora'. In: Huddleston, Rodney and Geoffrey K. Pullum (eds.) *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press, 1449–1564.
Ward, Gregory and Andrew Kehle (2002) 'Syntactic form and discourse accessibility'. Available at http://faculty.wcas.northwestern.edu/~ward/DAARC02.pdf (Accessed 8 October 2017).

Carlos Prado-Alonso works as a full time lecturer in *English Language and Linguistics* at the University of Oviedo (Spain). He is a member of the Research Group *Variation Linguistic Change and Grammaticalization* based at the University of Santiago de Compostela. His main research interests cover the fields of discourse and syntax. Most of his published work to date deals with the analysis, from a functional corpus-based perspective, of several English syntactic constructions, including inverted constructions, verbal anaphoric structures and parenthetical clauses.

Address: Dr. Carlos Prado-Alonso, Department of English, French and German, Faculty of Philosophy and Letters, University of Oviedo, C/ Amparo Pedregal s/n, E-33011 Oviedo, Spain. [email: pradocarlos@uniovi.es]