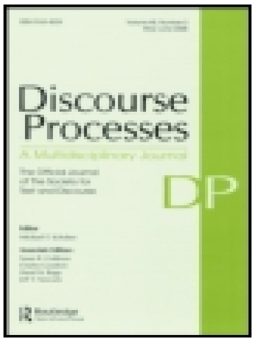

Discourse Processes

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hdsp20

Subjectivity (Re)visited: A Corpus Study of English Forward Causal Connectives in Different Domains of Spoken and Written Language

Marta Andersson & Rolf Sundberg

To cite this article: Marta Andersson & Rolf Sundberg (2021): Subjectivity (Re)visited: A Corpus Study of English Forward Causal Connectives in Different Domains of Spoken and Written Language, Discourse Processes, DOI: 10.1080/0163853X.2020.1847581

To link to this article: https://doi.org/10.1080/0163853X.2020.1847581

DISCOURSE PROCESSES
https://doi.org/10.1080/0163853X.2020.1847581

Routledge
Taylor & Francis Group

Subjectivity (Re)visited: A Corpus Study of English Forward Causal Connectives in Different Domains of Spoken and Written Language

Marta Andersson[a] and Rolf Sundberg[b]

[a]Department of English Stockholm University, Stockholm, Sweden; [b]Department of Mathematics Stockholm University, Stockholm, Sweden

ABSTRACT

Through a structured examination of four English causal discourse connectives, our article tackles a gap in the existing research, which focuses mainly on written language production, and entirely lacks attests on English spoken discourse. Given the alleged general nature of English connectives commonly emphasized in the literature, the underlying question of our investigation is the potential role of the connective phrases in marking the basic conceptual distinction between objective and subjective causal event types. To this end, our study combines a traditional corpus analysis with 'predictive' statistical modeling for subjectivity variables to investigate whether and how the tendencies found in the corpus depend on the systematic preferences of the language user to encode subjectivity via a discourse connective. Our findings suggest that while certain conceptual structures are quite fundamental to the usages of English connectives, the connectives *per se* do not seem to have a steady part in categorization of causal events. Rather, their role pertains to the level of intended explicitness bound to specific rhetorical purposes and contexts of use.

Introduction

The primary goal of this article is to investigate whether and how English language users make distinctions between different types of causal discourse relations in terms of subjectivity of the context. More specifically, the study focuses on four English discourse connectives as potential signals of subjectivity in CAUSE-RESULT relations (sometimes labeled 'forward causality' or formally defined as ‘A and as a result B’; Sanders et al., 1992) in different domains of both written and spoken English discourse.

Subjectivity is commonly understood as the degree of the speaker’s involvement in the relation construal realized as overt discourse manifestations of her or his point of view. The existing literature defines the 'speaker' as the entity whose intentional actions and/or mental activities constitute the source of causality, that is, a Subject of Consciousness (hence an SoC; see Pit, 2006; Sanders & Spooren, 2015; Stukker & Sanders, 2012; Traugott, 2010). Since subjectivity has been argued to be a basic cognitive principle that undergirds both production and comprehension of relations between discourse segments, it is therefore commonly accepted that language users categorize causal events into conceptually different objective and subjective types, which results in differences in their interpretations (based on Sweetser, 1990):

CONTACT Marta Andersson marta.andersson@english.su.se Department of English, Stockholm University, Universitetsvägen 10, 106 91 Stockholm, Sweden.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.
Supplemental data for this article can be accessed on the publisher’s website.

(1) It rained all night, *so* the streets are all wet. (non-volitional causality)
(2) It rained all night, *so* we decided to cancel the picnic. (volitional causality)
(3) It rained all night, *so* the streets must be all wet. (epistemic causality)
(4) It rained all night, *so* why don't we skip the picnic? (speech act causality)

While all relations above are instances of causality, in (1) and (2) the relation pertains to the domain of the states of affairs/events in the physical world, whereas both (3) and (4) convey the speaker's point of view. Utterances of this kind are regarded as subjective, as they pertain to the speaker's mental realm and are grounded in her or his beliefs and attitudes (Traugott, 2010). Subjectivity is therefore central to studies of the language aspects that express opinions, evaluations, assessments, and personal perspective. In recent years, an increased interest in this area has been observed, followed by an "affective turn" in philosophy, sociology, political science, and affective computing in artificial intelligence (Benamara et al., 2017). This trend obviously pertains to the rise of the social web and the possibility of widely broadcasting one's point of view.

Nevertheless, as argued in the literature, humans tend to use the vocabulary from the external (sociophysical) domain in speaking of the internal (emotional and psychological) domain (e.g., Sweetser, 1990). This tendency is believed to have led to a polysemous ambiguity of the meanings of discourse connectives, which commonly can cover a whole range of senses (*so* in (1)–(4) above). However, cross-linguistic research has demonstrated that language users consistently make specific choices to signal different causality types via specific connectives. In some languages, the connective specialization is strong. For instance, the Dutch *daardoor* ('as a result') occurs only with objective non-agentive events (such as (1); Stukker & Sanders, 2009). Less constrained yet significant preferences have been indicated also inGerman, French, and Chinese (e.g., Degand & Fagard, 2012; Li, 2014; however, see Santana et al. (2017) for the findings on Spanish, where the connectives were found to lack specific preferences).

Perhaps surprisingly, the question whether the functions of English connectives can be modeled in terms of subjectivity has not been sufficiently explored—likely because English connectives are assumed to lack the ability to signal the distinctions between causal event types. For instance, Stukker and Sanders (2012) point to the absence of direct English equivalents of several highly specialized causal expressions in French, German, and Dutch (e.g., the objective set of *parceque/weil/omdat* vs. the subjective set of *car/denn/want*, all of which are covered by English *because*). In the study of the causal connectives in English and Norwegian, Meier (2002) emphasizes the lack of connective specialization arguing that both *because* and *since* are equally felicitous marking relations that contain discourse-given information in the CAUSE segment (Meier, 2002, p. 51):

(5) They cannot have been flesh and blood, *since* they lived God knows how long ago.

Yet, as his corpus investigation demonstrates, while both *since* and *because* are indeed able to cover the function of *since* in (5), it is *since* that is the preferred choice. This finding suggests that certain specialization of the English connectives cannot be excluded.

Several other studies preliminarily confirm Meier's (2002) results. In a cross-linguistic comparison of backward causal connectives translations in English and French, Zufferey and Cartoni (2012) demonstrate that while the phrases *because, since*, and *as* do not exhibit any significantly different distributions in the subjective/new[1] category, both *because* and *as*, in fact, do specialize in conveying objective (*because*) and subjective (*as*) relation types. This result is to some extent in line with Andersson's (2019) study of the English phrases *as a result* and *for this reason* in written discourse, which indicated that (although not fully barred from marking other relations) *as a result* is overwhelmingly more frequent marking non-volitional event types, whereas *for this reason* shows a vague specialization between epistemic and volitional relations. The latter finding is particularly interesting, as *for this reason* has been argued to be constrained to objective relations in the studies based on retrospection (e.g., Knott & Sanders, 1998). Nevertheless, what all these observations suggest is that

English connectives are not unlikely to specialize in signaling specific types of causality, described in terms of the subjective-objective distinction between event types. The unexplored questions are the scale of this specialization and the source of potential deviations from the "preferred" domain of causality.

The latter problem pertains to the phenomenon of "non-prototypical" instantiations of the connective use (Stukker & Sanders, 2012), commonly identified in the cross-linguistic empirical studies. Even in languages where the connective specialization is quite strong, the connectives have been consistently found in relations that differ from the contexts they most frequently occur. This is the case for the Dutch inferential connective *dus* ('so'), which has been confirmed to signal relations also in the relatively objective volitional domain (like (2) above; Stukker & Sanders, 2012), or the objective *parce que* and the subjective *car* ('because') in French, which have been attested in both experimental and corpus studies to be interchangeable in many objective and subjective contexts (e.g., Degand & Pander Maat, 2003; Zufferey et al., 2018). According to Stukker and Sanders (2012), such non-prototypical uses of connectives are possible only in discourse relations that allow for interpretation of more than one causality type, which in natural language context usually involves the presence of both "subjectivity indicators" (e.g., modality, first-person pronoun, etc.; Traugott & Dasher, 2002) and objectivity features (e.g., objective connectives). As a result, the "mismatched" connective is relevant in the intended relation interpretation.

Based on these observations, Stukker and Sanders (2012) hypothesize that instantiations of the subjective-objective distinction between causal categories may depend on discourse type, and are likely susceptible to register conventions/themes discussed in context (i.e., context-sensitive); hence the non-prototypical usages of connectives. While this is a plausible idea, to date, most corpus research in the field has been focused only on written genres. Several notable exceptions are studies on German and French (Breindl & Walter, 2009; Zufferey, 2012) and, more recently, Sanders and Spooren's (2015) article on Dutch. This particular investigation of several registers (including spoken and semi-spoken discourse) indicated that in Dutch the preferences identified in writing remain unchanged across the language and medium type; however, both a large-scale study on Chinese (Li, 2014), and the aforementioned investigations of French and German, have demonstrated that connectives usually exhibit at least some degree of context-sensitivity. These findings suggest that input from language genres other than writing is needed to describe the systematicity of the connective functions as signals of causality.

The current study therefore sets out to investigate the potential relationship between discourse type and functions of specific English connectives. This task comprises the question as to why dependencies may arise. In the light of what is already known about both English (Andersson, 2019) and other languages, we suggest that while discourse type may indeed determine the functions of at least some English connectives, the aforementioned "mismatches" in the connective context of occurrence will be primarily related to a specific rhetorical goal or the speaker's intention, commonly pertaining to intersubjective meaning negotiation. This is often signaled by the connective choice and may be paired with other discourse features. Consequently, a related explorative question concerning the potential relationship between modality and discourse connectives in the heuristics of subjectivity analysis is also addressed below. To sum up, our investigation should ultimately yield further insights into the question whether subjectivity marking across different domains of language use is a proof of its basic role in categorization of causal events in English or a matter of explicitness bound to specific purposes and contexts of use.

Aims and research questions

As mentioned, the question hardly addressed in the existing subjectivity studies is the extent to which the functions of English discourse connectives can be described in terms of the objective-subjective distinction, that is, whether or not specific connectives are consistently chosen by the language users in the context of specific discourse relations. Such preferences have been labeled "subjectivity profiles" of the connectives (e.g., Stukker & Sanders, 2012) and can be established with reference to several dimensions of subjectivity (e.g., a subjective SoC, discussed below). This aspect is tackled here in a systematic analysis of the context of causal relations signaled with four English connectives, the unambiguous phrases *as a result* and *for this reason,* and the multifunctional connectives *so* and *therefore*, in different domains of both written and spoken discourse. A related question, which naturally emerges from our choice of material, is whether the connectives preserve their subjectivity profiles in different domains of language use[2] (e.g., fiction, academic prose, business meetings), that is, how prominent the identified tendencies are, not only in formal/edited registers, but also in more natural communicative contexts.

To answer these questions reliably, the study adopts two complementary perspectives. First is the perspective most commonly encountered in corpus studies, concerned with the use of a connective as identified in the corpus sample. A common problem with this approach pertains to the impracticalities of manual corpus coding, which is why the analyses usually rely on small datasets (see e.g., Stukker and Sanders (2012) for a summary; but see Bestgen et al. (2006) for a large-scale investigation of Dutch connectives in newswire). For instance, while the current study ambitiously starts with reasonably large samples of 250 instances per connective, at the level of individual domains of language use, the samples of some connectives become quite small. Another problem of sample-based investigations is that such data do not directly allow inference about the choice between connectives in specific discourse contexts. The reason for that pertains to the fixed sample sizes, which do not reflect the overall differences in frequency of the different connectives. To tackle this problem, our study adopts an additional focus, which is the language user's choice of connective in a given discourse context. We quantify this aspect by calculating estimated probabilities for the connective choice (based on the structure of the British National Corpus [BNC] data[3]). As a consequence, we investigate the data both through the prism of questions related to "what's in the text", and through the 'predictive' perspective of the choice between different connectives most likely to be made by a language user in a given discourse context. To our best knowledge, this kind of analysis has not been commonly adopted in subjectivity studies or in linguistic research in general.[4] Consequently, the more specific research questions pursued are as follows:

RQ1 Sample analysis:

(1a) Are English discourse connectives specialized to mark the subjective-objective characteristics of different causal event types? To what extent are these patterns stable across different domains of language use?
(1b) Do the connectives distributions over modalized contexts suggest that modality figures in subjectivity marking in English?

RQ 2 Predictive analysis:

(2) What general preferences of language users can we extrapolate from the corpus samples to the whole populations of the four connectives? What predictions can we make about the most probable user choices to mark a specific discourse relation in a given context?

Methods

Corpus data

The current material consists of random samples extracted from the BNC (Aston & Burnard, 1998). The sampling and coding procedures were carried out by the first author and aimed at collecting a random sample of 250 target instances of each connective per written/spoken discourse type. The samples of *as a result* and *for this reason*, however, became smaller in spoken discourse due to the scarcity of these phrases in speech. Table S1 (supplementary material, section I) lists the total frequencies of the analyzed items, including the corresponding proportions of target cases.

The process of sample selection included both automated elimination of nontarget instances (e.g., *as a result of*) and manual discarding of non-connective uses based on the collocation search.[5] These comprise, for example, certain sentence-initial instances of *so* (e.g., prefacing independent discourse units: *So, what's up, guys?*) or *therefore* in the sentence-final position. Unclear cases (mostly in speech) were also disregarded.

Annotation criteria

Subjectivity variables

This section describes the main categories of the concept of subjectivity, primarily based on Sweetser's (1990) classification of causal events. Our study follows the idea that subjectivity is commonly linked to specific conceptual and linguistic features in the relation context.[6] To date, this idea has been adopted by all existing subjectivity studies (at least to some extent). Our analysis focuses on the following discourse variables: discourse relation type, identity of the SoC, and modality type, which are all briefly described below.

Discourse relation type. Following Sweetser's (1990) classification of causal relation subtypes, illustrated in (1)–(4) above, all relations in our samples have been categorized based on the presence or absence of an SoC as the source of acting in the physical world (volitional relations) or the source of reasoning, evaluation and judgment (see Sanders & Spooren, 2015). In the latter case, the CAUSE segment of the relation is not an actual cause of the following event but a reason/premise for making the utterance that follows. Both speech acts (understood here in a broadly Austinian sense; see (4) above) and epistemic relations (paraphrased as "X and therefore it is concluded that Y", see (3) above; Sanders & Spooren, 2015) involve a subjective SoC that is the source of reasoning; however, it is speech acts that have been described as the most subjective relation type (owing to their hearer-oriented character; Pander Maat & Degand, 2001), which is the idea we follow.

SoC type. Our approach to subjectivity of the discourse SoC is in line with the recent endeavors in the field (e.g., Sanders & Spooren, 2015), which assume that the utterances that have to be interpreted with a reference to the SoC's mental domain are subjective. Consequently, in keeping with the distinction between subjective and objective causality, the implicit speaker (i.e., Author) SoC is treated as a maximally subjective instantiation of an SoC presence in discourse, for instance:

(6) But funding has been significantly less than other programmes, dissemination of materials less effective and leadership less dynamic. *For this reason* and probably also because Social Studies is not an area where governments readily welcome international initiatives, support for the programme is distinctly lukewarm. (BNC: BLY 517)

This SoC type endows the relation with a subjective perspective without an explicit presence of a speaker/agent (instead conveyed by, for instance, epistemic and attitudinal stance elements, as underlined in (6)). In contrast, the second most subjective SoC, Current Speaker, is overtly signaled by the presence of a first-person pronoun. The remaining categories, both regarded as relatively objective, are the Character SoC (including third-person pronouns and noun phrases) and Blend. In the present study, the Blend category comprises relations that involve an SoC with a vague identity, based on the combination of different points of view.[7] One such example are passive constructions, which commonly merge several perspectives, and often appear neutral:

(7) The 12 [political prisoners] also refused to wear their prison uniform. *As a result*, they were transferred to different prisons. (BNC: A03 603)

While the intentionally acting SoC can be identified in the first argument in (7), the resultative event *per se* (i.e., the transfer reported on in the second argument) involves the perspective of a non-volitional participant (the prisoners) and a backgrounded decision-making entity. Therefore, (7) was categorized as a Blend.[8]

Modality type. While modality *per se* has received relatively little attention in the existing empirical studies on subjectivity,[9] computational research efforts demonstrate that modal auxiliaries are one of the most reliable predictors of the subjective versus objective uses of causal connectives (Levshina & Degand, 2017). However important to the methodological developments in the field, this result obviously does not mean that modality is a necessary signal of subjectivity in naturally produced language. Given that modal auxiliaries have been found to comprise a mere 10% to 15% of all finite verb phrases in all registers of the English language (Biber et al., 2002), the present study is concerned with the extent to which modality contributes to subjectivity marking in English and whether it relates to the connective choice.

Needless to say, in some contexts, modality conveys an axiom of objective reality (e.g., *Brain needs oxygen*), and so the mere presence of a modal verb cannot be regarded as a default signal of sense attenuation. Consequently, our corpus samples have been coded according to Lyon's (1982) idea of a meaning continuum between a confident inference by a subjective speaker (the "I-say-so" component) and an objective periphery meaning ("it-is-so" component) related to the factual state of affairs. In this view, the two standard categories of modality (i.e., deontic and epistemic) cannot be fully separated, and so the interpretations of the verb *must* in the sentence *You must be very careful,* is context-dependent and may look as follows (Lyons, 1982, p. 109):

(a) You are required to be very careful (deontic, weakly subjective)
(b) I require you to be very careful (deontic, strongly subjective)
(c) It is obvious from evidence that you are very careful (epistemic, weakly subjective)
(d) I conclude that you are very careful (epistemic, strongly subjective)

For the sake of the statistical analysis, however, our study merges the above categories into Modality Type 1, which includes the strongly subjective types (b) and (d), and Modality Type 2, which comprises the weakly subjective types (a) and (c). Both categories included in Modality Type 1 have a clear context modulating function, whereas there is a cline between the more factual (a)-type and the weakly subjective (c). Yet, since modality is often hard to disambiguate (e.g., *will* as a future tense marker or a signal of epistemic eventuality; Jaszczolt, 2003); merging of the above categories was deemed practical.

Domains[10] of language use

To tackle the under-researched question of the relation between discourse connectives distributions and the domain of language use, and to avoid the classic trap of the Yule–Simpson paradox,[11] our corpus samples were divided into four domains (i.e., poststratified). The domains adopted for our purposes follow the pre-existing BNC categories of the data, with two exceptions. One is a domain of written discourse, in the following labeled 'Non-Academic,' which is an amalgam of several smaller registers, considered roughly a semi-formal discourse (e.g., biographies, unpublished written material, etc.). Another exception comes from spoken language, and has been branded 'Leisure' in the BNC. This domain consists of text types primarily categorized as such in the BNC; however, due to the corpus design, our random samples of spoken language comprise also relations found in uncategorized[12] transcriptions of informal conversations recorded in different contexts. Since all these instances tackle casual conversational topics, we decided to include them in the Leisure domain. While not optimal (particularly in the case of the heterogeneous Non-Academic domain, which may be less straightforward to interpret), these choices were made for the purpose of reducing the number

Table 1. Domains of Language Use in the BNC

Genre	Domain			
Spoken	Public (Publ)	Educational (Educ)	Business (Busn)	Leisure (Leis)
Written	Academic (Acad)	Newspapers (Newsp)	Non-academic (NonAc)	Fiction (Fict)

of language domains in our statistical analysis. Table 1 provides a general overview of the language domains included in our study.

According to the BNC description of the corpus design, the domains listed in Table 1 are to a great extent context-governed (i.e., recorded in specific types of events for speech or representing specific type of writing), which means that they follow Biber et al.'s (1998, p. 154) scale of text register formality. Thus, in Table 1, the domains are ordered from the most to the least formal types.

Statistical analysis

The poststratification mentioned above implies that the domain sample sizes vary depending on the domain population sizes, the frequency of the connective in the different domains (see Table S1, supplementary material, section I), and random sampling effects. However, the statistical analysis was, following the common principle, made conditional on the domain sample sizes, that is, regarded them as given. Within each discourse domain, the distribution of the sample data was analyzed for subjectivity features (factors: Relation, SoC, and Modality). The sample data are summarized in tables of counts and proportions (see Supplementary material, sections I and II) and in mosaic plots (Baayen, 2008; Friendly, 1994). Further, log-linear models were fitted to these multinomial data, describing how Domain, Relation, and SoC type influence the frequency pattern of each connective and allowing for statistical tests of hypotheses (see Appendix 2 for more details).

This part of the analysis was a natural and convenient investigation for answering questions of RQ1 type, for instance, when comparing context types, where is a specific connective most commonly found in the data? Do the connectives differ significantly in their frequencies of subjective versus nonsubjective uses? Do their frequency patterns differ between the domains of language use? However, using additional information about the composition of the BNC enables us (albeit with somewhat greater uncertainty) to address questions pertaining to RQ2, that is, concerning the choice of connective made in a given context of speech or writing. Methodologically, answering such questions in the present form of a sampling study is more intricate, as it involves an application of Bayes' formula for inverse probability calculations, which we refer to as 'predictive analysis.' We describe the procedure below in terms of counts of the population instances of the connectives.

Once the total number of each connective per each BNC domain was obtained from the automatic sample extraction, the next step was to find the number of target instances. By reducing the crude total count by the proportion of target instances identified in the initial sampling, we arrived at an estimate of the total number of target instances for each connective in each domain. For any of these connectives, its estimated share among all four (i.e., the ratio of the number of target instances for this connective to the corresponding total over all four connectives) is the estimated probability that a language user has chosen this specific connective to signal the intended discourse relation.

However, since this analysis did not include subjectivity features (e.g., SoC type), it can be regarded as crude or provisional. A step toward answering RQ2 in more detail was therefore to divide the domain in parts, according to Relation, SoC, and Modality type. For a domain part satisfying a particular restriction on Relation/SoC, analogous calculations can be carried out. The only complication is that the size of such a domain part is not known; however, it can be estimated directly from the corresponding sample for each connective separately. The cost of specifying the instances of the connectives for Relation, SoC, and Modality type is an additional degree of uncertainty of the findings due to smaller sample sizes.

The predictive calculations described above can be represented (for each language domain separately) by the following formula for the probability of choice of a certain connective from the quartet studied here, exemplified by *so*:

> *Pr*(*so* | target and specified Rel&SoC) is proportional to the product
> *Pr*(Rel&SoC | target *so*) × *Pr*(target | *so*) × *Pr*(*so*).

The symbol (|) denotes conditioning, that is, *Pr*(B | A) is the conditional probability for event B, given event A. The proportionality constant is the same for all four connectives, so it need not be specified. The last factor is known for each BNC domain; the preceding two factors are estimated in the sample study.

All calculations and graphical illustrations were carried out in the program package R. Estimated proportions are often provided with their standard errors (± *s.e.*), from sampling uncertainty. For small samples and for proportions close to 0 or 1, when the standard error is not an adequate measure of uncertainty, it is replaced by a 70% confidence interval (two-sided or one-sided), approximately equivalent to the interval "± *s.e.*" for normal samples.

Results

The results of the sample study are presented as mosaic plots comprising three variables at the same time. In these plots, the tiles represent two factors horizontally and one factor vertically, with the tile area representing the corresponding proportion of the sample total (250 instances per connective, fixed by design) or of the domain (poststrata) totals.

Sample analysis in written discourse

Domain vs connective vs relation type. The poststrata sizes for the four connectives are shown in the mosaic plot of Figure 1, dividing the total sample size 4 × 250 = 1,000 in 4 × 4 = 16 tiles. The tiles represent horizontally the four connectives and vertically the four discourse domains (Acad –academic; Newsp –newspapers; NonAc –nonacademic; Fict –fiction), such that each tile area corresponds to the sample proportion for that combination of connective and domain. The general picture in Figure 1 is that all connectives are frequently found in the Non-Academic domain (between 53% and 63% of all instances). However, due to the heterogenous nature of this domain, this finding is somewhat less informative than the one-domain specific results (yet, it has to be pointed out that the domain includes semi-formal written registers). For the connective *so*, the next biggest domain of occurrence is Fiction (75/250, 30%), whereas the remaining three phrases are present mostly in the Academic domain (e.g., *therefore* 107/250, 43%) and quite rarely in Fiction. *As a result* differs from *therefore* and *for this reason* by a relatively high frequency in the Newspapers domain (12% vs. 2% of the latter two); a reason for that may be the factual nature of the themes discussed in news reports and hence the need for an explicit/objective connective. For the exact numbers, see Supplementary material, section II. The sampling standard errors in the Figure 1 percentages are ≤3 percentage units.

Figure 2 is an analogous 4 × 4 mosaic plot showing the relationship between Relation type versus connective (SpA –speech acts; Epi –epistemic; Vol–volitional; and NonV –non-volitional). The general picture that emerges from the tile areas of Figure 2 is that English connectives exhibit certain specializations in marking specific relation types: *so* and *therefore* are both most frequently found in the subjective (epistemic and speech act) relations (of *therefore* 199/250 = 80%; of *so* 151/250 = 60%). There are, however, differences between the connectives in their distributions over specific relation types: *so* is more common than the other phrases with speech acts, whereas *therefore* is found mostly in the epistemic category. *As a result*, in contrast, is most frequent in objective relations (only 21% in subjective contexts and absent in speech acts). Finally, *for this reason* is about equally distributed over

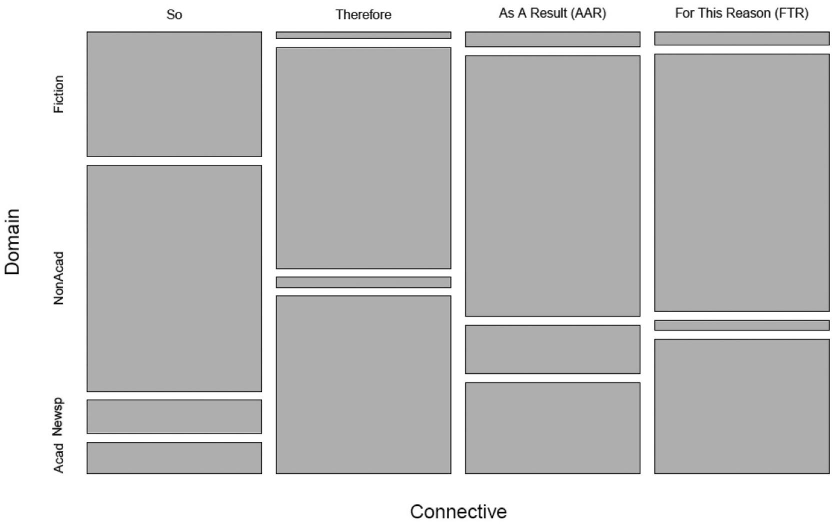

Figure 1. Domains per connective, written discourse, $n = 4 \times 250$.

subjective and objective relations, yet slightly pointing toward subjective contexts (57% vs. 43%; of the latter category 34% volitional and only 9% in non-volitional relations). Overall, we note that none of the connectives is constrained to just one prototypical context; however, we can talk about significant tendencies of occurrence in specific discourse domains. Like in Figure 1, the sampling standard errors in Figure 2 percentages (and all other fractions of 250) are ≤3 percentage units.

The question that Figure 2 does not address, however, is the influence of the domains of language use on the frequencies of the different types of SoC, which are likely to interplay with relation type. To understand the general picture better, the following analysis proceeds in two steps. In the first step, only non-volitional causal relations are considered, as they are intrinsically devoid of an SoC and thus would yield partly uninformative plots of specific SoC and Rel combinations. These relations are

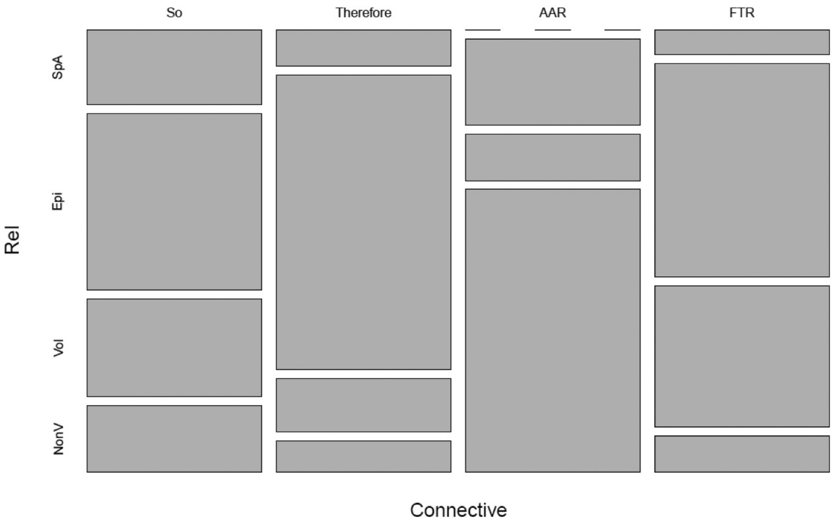

Figure 2. Rel versus connective, written discourse, $n = 4 \times 250$.

therefore excluded from further analysis so the plots become easier to interpret. Subsequently, the interplay of Relation and SoC type will be studied for each connective, with a focus on the extent to which this interplay may depend on language domain.

Non-volitional discourse relations. The current section discusses the written and spoken data all at once. Both genres are illustrated in Figure 3, which shows the frequencies of all four connectives in the context of non-volitional RESULT relations.

As we can see, there is a remarkable similarity between the bar heights for speech and writing. The main difference is that the spoken data have somewhat smaller proportions than the written material (recall also that in speech, both *as a result* and *for this reason* generated smaller datasets, due to exhaustive sampling of their small populations, which implies a lack of sampling errors). Overall, the non-volitional relations proportions are very high for *as a result* but not for the other connectives: in writing, about 70% (0.68 ± 0.03) of the instances of *as a result* mark an event without an SoC, consistently across all written discourse domains. In speech, *as a result* is also characterized by a much higher percentage of non-volitional relations (41%) than *so* and *therefore* (11% and 5%, respectively).

However, for the two latter connectives in non-volitional relations, a domain-dependency has been indicated. In written discourse, the non-volitional relations with *so* count as only 7% (5/75) in Fiction but 18% (25/136) in the NonAc domain (in the other two domains, *so* is rare). This difference between the domains is statistically significant at the 5% level (p = .02 by Fisher's exact test). Analogously, with *therefore*, the proportion of non-volitional relations is only 4% (4/107) in the Academic but 11% (14/133) in the NonAc domain (in the other two domains, *therefore* is infrequent). In this case the difference is significant at only just the 5% level (p = .047). The reason for these dependencies pertains likely to the register type, as it seems in the less formal register, particularly multifunctional connectives, may be used more flexibly. In contrast, in both biggest domains of the occurrence of *for this reason* (i.e., Acad and NonAc domains), the proportions of non-volitional relations were (only) around 9%, and so no difference between domains was found.

In speech the results are similar. For *so*, there is a statistically significant variation in non-volitional relations frequencies between the four domains (deviance test p = .01). More precisely, non-volitional

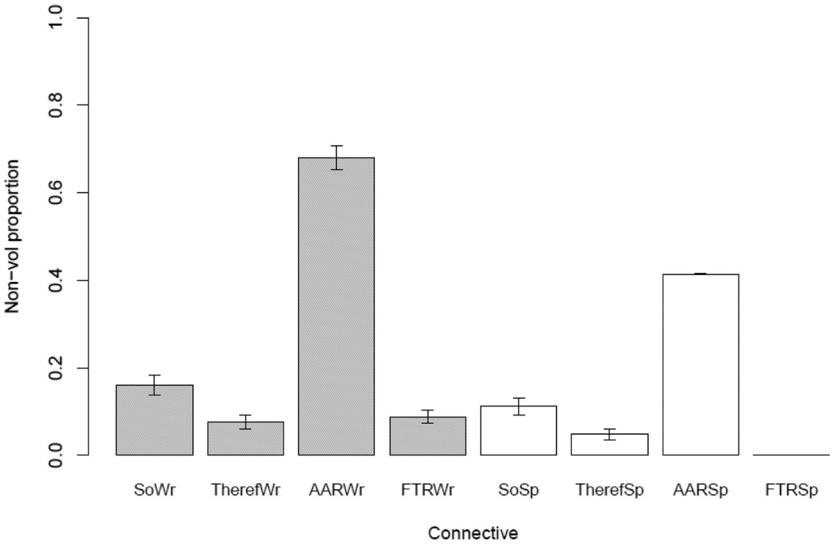

Figure 3. Non-volitional use of connectives, written and spoken data.

percentages are substantially larger in the Leisure and Education domains (≈15%) than in the Business and Public domains (4% and 0%). There were no domain differences indicated for either *therefore* or *as a result*. In the next step, the non-volitional relations are left out of the analyses so that the plots become easier to interpret.

Discourse relation type vs SoC type. The analysis presented in the current and the following sections discusses log-linear models (for each connective) for the counts of the 3 × 4 = 12 (Rel, SoC) combinations, with Domain as a third factor, excluding domains where the connective in question is rare. As already mentioned, non-volitional relations are excluded.

To establish the subjectivity profiles of our connectives, we have to tackle the question of their distributions over Rel type, SoC category, and Domain. The simplest possible structure of the set of (expected) frequencies is that the distribution of the connective over any of these three factors is independent of the other two factors. For example, the distribution over Rel types (i.e., the probability/odds for any particular Rel type) would be the same, regardless of SoC type and Domain. Mathematically, such a total independence structure is expressed as a product of probabilities:

$$\textbf{Model 0}: Pr(\textbf{Domain}, \textbf{Rel}, \textbf{SoC}) = Pr(\textbf{Domain}) \times Pr(\textbf{Rel}) \times Pr(\textbf{SoC})$$

However, as is evident from the data, the total multiplicativity formulated in Model 0 is inadequate. The Rel and SoC factors are far from multiplicative, that is, the distribution over Rel types is quite different for different SoC types, and vice versa, and this is true for all connectives. In other words, for comparisons of the probability distributions for Rel or for SoC between connectives or between domains, the whole (Rel, SoC) frequency table must be considered, not just Rel or SoC alone.

The simplest model allowing non-multiplicative (interacting) Rel and SoC is the following partially multiplicative model (Model 1), which plays an important role in the forthcoming analyses. Model 1 assumes (for a particular connective considered) that the expected frequencies and the corresponding probabilities *Pr*(Domain, Rel, SoC), can be factorized as follows:

$$\textbf{Model 1}: Pr(\textbf{Domain}, \textbf{Rel}, \textbf{SoC}) = Pr(\textbf{Domain}) \times Pr(\textbf{Rel}, \textbf{SoC})$$

The interpretation of Model 1 is that the (conditional) probability table for (Rel, SoC), given Domain, is the same for all domains. In other words, per domain, the probability (or odds) for any particular (Rel, SoC) combination is independent of Domain.

To find the simplest fitting structure for our data, we used successive model simplification, starting from a saturated model, with no assumed structure (i.e., wholly unspecified *Pr*(Domain, Rel, SoC)). Except for the connective *so*, we arrived at Model 1 as a result (see Appendix 2 for an account of this inferential process for each connective). Below we concentrate on the results and first discuss the distribution structure of *so*.

For the connective *so* (Table A1, A.2.1[13] and mosaic plots Figures 4 and 5), successive model simplification, starting from the saturated model, did not lead to Model 1 but showed a substantial interaction (non-multiplicativity) between Domain and SoC type ($p < .001$). In other words, the differences in sampling probability between SoC types differed between domains. More specifically, the Author and Blend categories were much more frequent as SoC in the NonAc domain than in Fiction, and, correspondingly, the Current Speaker and Character categories were much more frequent as SoC types in Fiction than in the NonAc domain (see bottom line of Table A1 in A.2.1). However, the interaction between Rel and SoC types (on average over domains) is of even stronger magnitude ($p \ll .001$), as seen from the pattern in the left third of Table A1. The likely interpretation of this effect is that English discourse relations may be signaled by language users via additional contextual features, such as SoC types (please note the similar interaction for the other connectives in the following). A relevant example here is the absence of Author SoC in volitional relations with *so*;

this combination, while rare in general, can be realized via the passive voice (common with *therefore*). With *so*, however, volition is most commonly conveyed via prototypically agentive SoC types (i.e., Current Speaker and Character).

For the connectives *therefore* and *for this reason*, in contrast, we conclude there is no domain-dependency in their (Rel, SoC) tables of frequencies (A.2.2 and A.2.4), at least not between their larger domains of occurrence. For *as a result*, there is only one large domain of occurrence (after the exclusion of non-volitional relations), so statistical comparisons would not be meaningful, yet the available data do not indicate any domain-dependence. We therefore disregard the influence of Domain for all three connectives. In other words, we accept Model 1. Within this model, the lack of multiplicativity between Rel and SoC was statistically tested for *therefore* and *for this reason,* and was found to be strong ($p \ll$ .001); note, however, that Table 2 in Section A.2.2 and Table 4 in Section A.2.4 look to a large extent similar (Rel and SoC interact in a similar way for the two connectives). For *as a result*, the relevant table is A3, in A.2.3. Tables A1 to A4 are summarized in a mosaic plot below (Figure 4).

Figure 4 shows the Rel versus SoC frequency table (for all connectives), cross-domain, the non-volitional outcomes excluded. More precisely, the frequencies are represented by the tile areas in proportion to the whole area for the connective considered. Importantly, for *so*, the cross-domain plot is somewhat misleading because of the influence of Domain; however, this effect is corrected for in Figure 5, which shows the frequencies of all (Rel, SoC) combinations for *so* per domain, block wideness proportional to domain size.

To summarize the most important findings of the analysis so far (see also Section A.2.5):

(i) Despite significant tendencies, none of the connectives is constrained to a single prototypical context of occurrence.
(ii) The subjectivity profiles of the connectives are relatively stable across language domains with the exception of *so*, which exhibits substantial and statistically significant variation between domains (see below).

As to the more specific SoC and Relation combinations, *therefore, as a result*, and *for this reason* are mostly used in epistemic relations (proportions 0.77 ± 0.03, 0.65 ± 0.05, and

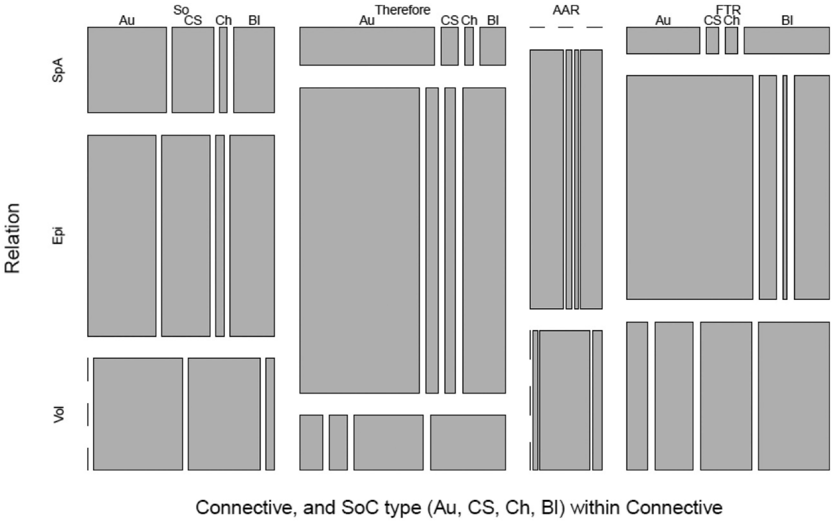

Figure 4. Rel versus connective, split by SoC, written, cross-domain, n = 749.

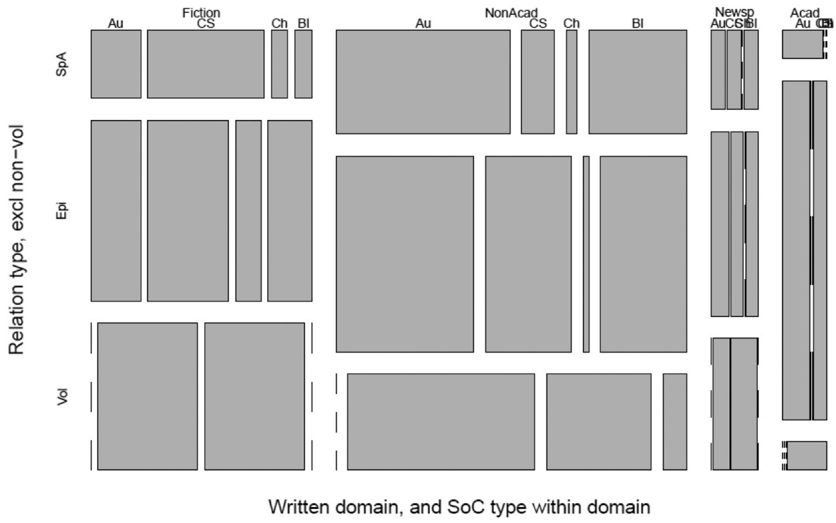

Figure 5. Relation versus SoC, per domain, for written SoC, *n* = 210.

0.56 ± 0.03), predominantly then in connection with the Author SoC. This likely pertains to the nature of epistemic contexts, which represent a prototypical setting of a subjective relation commonly issued by the most subjective SoC type. To some extent the same finding is true for *so*, yet the cross-domain frequency of epistemic relations with *so* is not as high (0.50 ± 0.03). Interestingly, the most apparent expression of the domain-dependency indicated for *so* is the combination of epistemic relations with Author SoC, where the frequency of this connective varies between the domains, from only 0.11 ± 0.04 in Fiction, via 0.22 ± 0.04 in the Non-Academic, and 0.20 ± 0.10 in the Newspaper domains to 0.57 ± 0.13 in the most formal, the Academic domain. Further, for all four connectives, the quite prototypical combination of SoC Character and volitional relation is common, from as high as (79 ± 8)% (22/28) for *as a result* to (28 ± 5)% (24/85) of instances of *for this reason*. However, the frequency of the volitional relation type is highest with *for this reason* (85/250, 34% of all its instances), which suggests that the connective *per se* implies a volitional action, and so a Character SoC may not be as necessary with *for this reason* as with the other connectives used for the same purpose.

Sample analysis: spoken discourse

Connective vs. domain vs relation type. Figure 6 shows the distributions of all investigated connectives over the four spoken discourse domains: Public (Publ), Education (Educ), Business (Busn), and Leisure (Leis). The total sample size is 2 × 250 = 500 (*so* and *therefore*), 1 × 70 (*as a result*), and 1 × 6 (*for this reason*) in 4 × 4 = 16 tiles.

The poststrata sizes for the four domains illustrated in the plot of Figure 6 indicate that the connective *so* is most frequent in the Leisure domain (114/250; 46% instances). In contrast, only 14% of instances of *therefore* occur in this domain, while the connective is most common in the Public (111/250; 44%) and Educational (63/250; 25%) domains, likely as a signal of inferential reasoning. However, the connective *as a result* is also found mostly in Educational domain (44/70; 63%), which is probably based on its factual nature. The phrase *for this reason* is left out of the below discussion due to its paucity in speech.

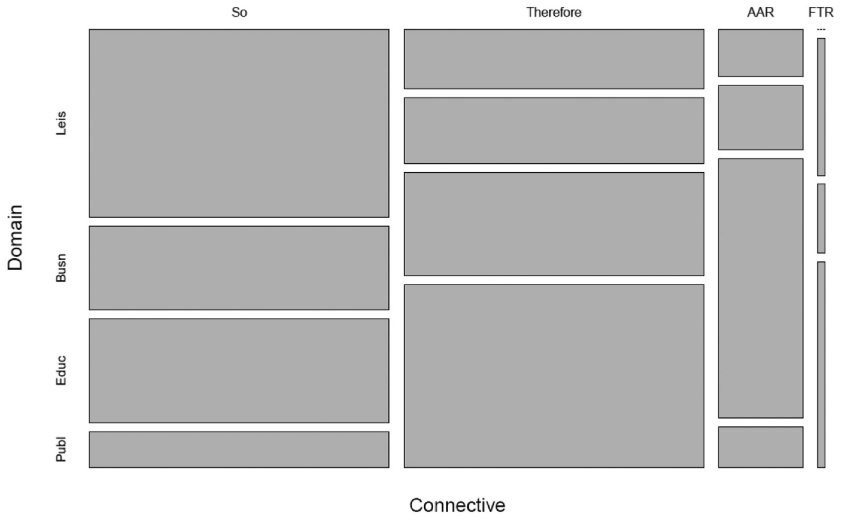

Figure 6. Domain versus connective, spoken discourse.

Discourse relation type versus SoC type. The analysis reported on in this section follows the procedure for the investigation of SoC types and discourse relations in written discourse introduced above (see A.2.6-2.9 for details). Recall that non-volitional outcomes are excluded. Our general findings resemble those obtained in written discourse in that *therefore* turns out to be stable across spoken domains, whereas *so* exhibits domain-dependent tendencies also in speech. *As a result* is essentially confined to one domain, and so testing its consistency over domains is counterintuitive.

In the log-linear modeling of *so*, the Public domain and the Blend SoC type were omitted as too infrequent with this connective. Both similarities and differences between the domains are demonstrated in Table A5 (A.2.6), where the fitted model is a domain-size weighted average over all three domains. The overall conclusion is that *so* in speech is most frequent in epistemic relations, with the Author or Current Speaker SoC (about 35% of instances in each of these categories). While this resembles the tendencies found earlier in writing, the distributions of SoC types in written discourse were spread out more evenly and included a substantial proportion of Blend SoC. A more formal comparison between written and spoken *so* is not undertaken here, as the domains are not the same (rendering the models incomparable).

As to the SoC and Rel combinations with *so* in speech, there is a statistically highly significant variation between the domains ($p < .01$). In the statistical analysis, it is possible to interpret this as an interaction either between Rel and Domain or between SoC and Domain. To put it briefly, comparing the least formal (Leisure) and the most formal (Education) domains, the frequencies of Author SoC (in speech acts and epistemic relations; absent in volitional contexts) are substantially higher in the Educational domain than in the Leisure domain (61 ± 7% vs. 35 ± 5%). Analogously, the opposite holds for the Relation type: volitional relations are much less frequent in the more formal domains (Education 4 ± 3% vs. Leisure 27 ± 5%).

With the connective *therefore*, since Model 1 fits the data reasonably well (see Table A6, A.2.7), it is motivated to use the same frequency table over Rel and SoC for all domains. The fitted table can be found as the left part of Table A6. The immediate observation emerging from the table is the prevalence of epistemic relations with *therefore* (75% of all instances), entirely dominated by high frequencies of Author and Current Speaker SoC (50% and 40% of the epistemic instances, respectively). The remaining SoC and Rel combinations are much less frequent or almost absent (e.g., Character in Speech acts, ≤1%). While this resembles the findings for spoken *so, therefore* does not show the instability of the subjectivity profile found with *so*.

As a final test for *therefore*, we carried out a statistical comparison between written and spoken discourse. The analysis (A.2.10) yielded a statistically significant interaction between these two genres and the SoC type ($p \ll .001$). The most striking difference is the high frequency of SoC = Current Speaker in the spoken use of this connective, relative to the written use (see Table A7, A.2.10). This finding is particularly pronounced in epistemic relations, with a frequency difference for Current Speaker of 0.25 (0.05 ± 0.01 vs. 0.31 ± 0.03), and likely pertains to self-mentioning, usually more frequent in speech than in writing

Finally, while *as a result* is not included in the log-linear modeling due to its low occurrence in speech (70 target instances found, further reduced to 41 by exclusion of the large proportion of non-volitional relations; recall Figure 3), we provide tentative data observations below. The majority (24/41) of instances come from the Education domain, with a remarkably high frequency of Author SoC type in epistemic relations (11/24; 0.46 ± 0.10). This finding seems to counter Stukker and Sanders's (2012) assumption that subjective relations marked with an objective connective contain subjective elements less often than those marked with a subjective connective, as the frequencies of the Author SoC and epistemic relation combination are similar *with therefore*; however, more corpus data would be needed to verify this reliably for *as a result*. Given the relative difficulty of linking the use of the factual *as a result* with an epistemic relation, we believe the subjective SoC type may be necessary to convey the intended level of subjectivity. The same is implied by the relatively common occurrence of Blend SoC with *as a result* (over domains) (23/41; 0.56 ± 0.08); this type usually occurs in factual objective relations that involve certain perspectivization (e.g., an evaluative adjective in an otherwise neutral context, "much needed oxygen") and may even involve epistemic interpretations. Importantly, all these observations resemble the earlier findings in written discourse.

To summarize the results in the speech section:

(iii) overall, the connectives *so* and *therefore* preserve their tendencies to occur in epistemic relation type earlier found in written discourse, and with Author and Current Speaker SoC. The latter type is generally more frequent in speech, likely because spoken discourse tends to focus on the current speaker (the "I"), which may be avoided in more formal registers. The somewhat surprising behavior is that of *as a result*, which exhibits quite a pronounced tendency to occur in epistemic relations with an Author SoC in Educational domain.

(iv) the match between the prototypical context of occurrence of the connective, that is, Rel and the most expected SoC type, seems less predictable in speech than in writing for *so,* and has been indicated to be domain-dependent.

Language user choice of connectives: predictive analysis

As mentioned before, corpus analyses confined to manually analyzed connective samples are limited in size and yield an approximate picture of the corpus composition per connective, at best. The predictive analysis, in contrast, is meant to provide a generalization to the composition of the whole population of the connectives under study. We shall therefore see how the tendencies found in the corpus samples correspond with predictions about the preferred uses for the whole population of the analyzed connectives. Please note that due to the scarcity of *as a result* and *for this reason* in spoken discourse (see Table S1, Supplementary material, section I), only written discourse results are included in the following discussion. More detailed results for both discourse types can be found in the Supplementary material. Figure 7 illustrates the distributions for the choice of connective for each of the four domains of the written discourse (see Supplementary material, section III, for specific findings on each domain separately). Uncertainties are discussed in Appendix 3.

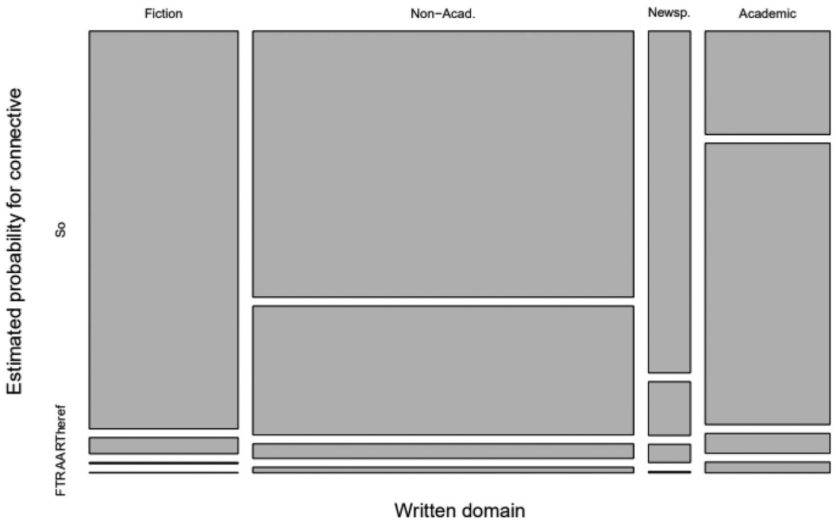

Figure 7. Choice of connective, per written domain.

Perhaps not unexpectedly, the connective most likely to be chosen by English language users in writing is the multifunctional *so* in all domains but Academic discourse, where *therefore* is clearly preferred. The predilection for marking causal relations with *so* is the strongest in Fiction and Newspapers, the two least formal discourse types, which are most likely to comprise elements of spoken discourse. Recall that the tendencies identified in the NonAc domain are very general due to its heterogenous nature. Yet, if we compare this domain with Fiction, the most conspicuous feature of the former is the pronounced presence of *therefore*. This observation may be related to the level of formality of the NonAc, which is composed of semi-formal registers. Finally, while the remaining two connectives are, overall, quite an infrequent choice, the presence of *as a result* is perceptible in all domains except Fiction.

Relation type and SoC. Here we discuss more detailed preferences for the connective use with specific Relation and SoC, yielded by the predictive analysis. All results can be found in section III in Supplementary material and are additionally illustrated in Appendix 1 (Figures A1–A4).

The Academic domain (Figure A1) has a very strong preference for *therefore* in the context of epistemic relations with an Author SoC (74 ± 5%) or Blend SoC (69 ± 11%). Given the multifunctional nature of the connective and, moreover, the multifunctional nature of the competing option, which is *so*, this is a pretty revealing finding. Another interesting observation concerns *as a result* and its strongly pronounced presence in non-volitional causality; while *so* can certainly cover such senses and is estimated to be chosen in about 50% of relations with no SoC, the 27% for *as a result* proves its strong functional relationship with such relations. Finally, an interesting observation (not shown in Figure A1) is the exclusive relationship of *for this reason* and Volitional RESULT with a Current Speaker SoC; while such relations are rare in this domain, *for this reason* is exclusively chosen in these cases.

Newspapers (Figure A2) is a very small domain, and hence not many reliable observations can be made. For all combinations of Rel and SoC shown, the probability of marking with the connective *so* is 0.7 to 0.9. In epistemic relations with Author and Current Speaker SoC, the most frequent alternative is *therefore*. Finally, *as a result* is chosen in 10% of all cases of Non-volitional RESULT (80% are marked with *so*).

In the Non-Academic domain (Figure A3) , *so* dominates (estimate ≥ 66%) only in two overlapping contexts: with the Current Speaker SoC, regardless of Relation type, and in speech acts, regardless of SoC type. *Therefore* is the most preferred (≥50%) connective in epistemic relations with Author and Current Speaker SoC, and in volitional relations with Blend SoC. Finally, *as a result* is chosen in 13% of non-volitional relations in this domain, while *for this reason* 5% of volitional relations (most frequently with a Blend SoC).

Fiction (Figure A4) is the least formal of the written domains and is therefore dominated by the connective *so*. Another pronounced preference relates to the choice of *therefore* in epistemic relations with a Character SoC (27%). This finding is likely related to the ability of *therefore* to occupy the clause-medial slot (excluded for *so*; yet, *so* is chosen in over 70% of epistemic relations with Character SoC) and signal an embedded conclusion, for instance:

(8) Jean-Claude was of the opinion that all Jews were rich, part of an international conspiracy and deserving, *therefore*, of whatever hideous fate was in store for them. (BNC: FAT 2664)

Finally, *as a result* is chosen in 4% of the cases of Non-volitional RESULT. Given its general paucity in this domain, and the multifunctional nature of *so*, which can cover also this sense, this is, again, quite a revealing finding.

Modality: sample and predictive analyses

As the sample analysis indicates, overall, in both written and spoken discourse, all four connectives are most frequently found in non-modalized contexts. This suggests that in most cases, either the connective itself is able to contribute the intended level of subjectivity, or the subjective perspective is signaled by other textual means. However, differences between the connectives distributions over modal contexts have been identified.

In written discourse (see Appendix 1, Figure A5), the connectives most strongly attracted to modal verbs are *therefore* and *for this reason*, while *so* and *as a result* are less common (respectively, 33%, 28%, 16%, and 16%; all ±3%). The percentage of modality is high particularly in epistemic relations, where modulation of the context is likely to figure in the subjective construal. The same is true of spoken discourse (see Figure A6), where the modalized contexts are mostly epistemic (with *therefore*) and speech act relations (with *so*). However, we have also discovered a general difference between written and spoken discourse, as the odds for the more subjective Type 1 versus Type 2 modality is higher in speech, and Type 1 is found to be more frequent than Type 2. This is likely related to the more subjective nature of the spoken genre and the tendency of both *so* and *therefore* to occur in the context of the Current Speaker SoC in speech. This highly subjective type of SoC may be related to more frequent subjective modals uses.

The predictive analysis for the connective choice in terms of the three modality categories (including a "no-modal" category)[14] indicated that the presence of a modal verb does essentially not affect the connective choice. Further, the estimates of the population composition in terms of modality (see Figure A7) yielded by the model, largely confirm the findings from the sample analysis and show that (regardless of the connective used) modality is not frequent. Of the two modality types, Type 2 is estimated to be more frequent than Type 1 across all domains of language use in both speech and writing, except from the Public domain of spoken discourse, where Type 1 prevails.

General discussion

The two main aims of the present study were, first, to establish whether English forward causal (CAUSE-RESULT) connectives exhibit identifiable subjectivity profiles in terms of their occurrences with specific relation, SoC, and modality types, and how/whether these profiles are affected by discourse type.

The second aim was to make predictions (based on the sample analysis) as to the most probable connective choice dependent on the discourse type and the other analyzed variables.

The general picture that emerges from our study is that despite an overwhelming preference for marking all types of causal events in English with *so* (confirmed by the predictive analysis), in the corpus, we can identify relatively stable tendencies for the connectives to signal the subjective-objective distinction, even though their functions are not always clear-cut. As we would like to argue, the connective specializations in English may not pertain entirely to conceptual distinctions between event types but rather to specific register and rhetorical purposes. In this respect, English resembles French (but see Blochowiak et al., 2020, on the most recent results on French) perhaps more than Dutch, where the difference between the two usages is quite sharp, yet non-prototypical instances of connective use are not uncommon (e.g., Sanders & Spooren, 2015). We therefore believe that our findings should have further-reaching effects on how the role of English connectives in marking different types of causality may be viewed.

Nevertheless, our results of the sample analysis differ between the connectives; while *therefore, as a result*, and *for this reason* are robust, *so* exhibits significant domain dependency in terms of the combinations of the most frequent SoC and relation types. While the connective is overall most frequent in subjective relations, the SoC type with *so* changes according to the language domain. This is likely related to the highly multifunctional nature of the phrase. In contrast, the factual *as a result*, while not barred from epistemic relations, needs the most subjective Author/Current Speaker SoC to signal subjectivity and is otherwise confined to non-agentive events. Further, the connective *so* is not only more frequent in speech acts, but in fact, the remaining three phrases are nearly absent from this relation type. Finally, given the low frequencies of *so* in objective contexts, we conclude that the phrase can be regarded as a marker of inferential/subjective construal. A similar conclusion can be drawn for *therefore*, which has a pronounced tendency to mark epistemic relations across all language domains, most commonly in the context of the most subjective Author/Current SoC types. However, this tendency is stable across both speech and writing, which can be said to be the crux of the difference between *therefore* and *so*—their functions are to some extent similar, yet it is *therefore* that can be regarded as a cue for epistemic relations (i.e., what is expected in its proximity, confirmed also in the predictive analysis as a preference for *therefore* in high-register inferential contexts), while *so* is simply associated with subjective settings. Finally, the two connectives have clearly different tendencies to occur in either more formal (*therefore*) or less formal language (*so*), both in speech and writing. Their usages are thus related to different register purposes.

As to *as a result* and *for this reason*, based on their unambiguous semantics, functional restrictions are expected. However, a particularly interesting case is that of *for this reason*, which has been claimed to be an objective connective, and while it indeed commonly signals volitional relations, we found it slightly prevailing in epistemic contexts. It is important to note, however, that even though volitional relations have been argued to be relatively objective (Stukker & Sanders, 2009), it is not uncommon that epistemic and volitional causality are signaled by the same connectives, based on the presence of an SoC, which enables a subjective construal (Pander Maat & Degand, 2001; Sanders, 2005). Finally, while *as a result* is not entirely barred from other relations, the overwhelmingly most frequent context of its occurrence are non-volitional events, which is a finding stable across discourse domains. Consequently, the phrase can be regarded as a cue for this specific relation in most cases, unless another textual feature signals a non-prototypical use; see (10) below). The main caveat to these findings is the scarcity of *as a result* and *for this reason* in spoken discourse; however, this observation is in line with previous hypotheses in the literature (e.g., Zufferey et al., 2018) that the tendencies in connective use can sometimes be explained based on different register purposes. This is at least partly the case for *as a result* and *for this reason* (the latter practically confined to high-register written academic prose) *versus* the multifunctional *so*.

Nevertheless, the results yielded by the predictive analysis confirm the tendencies found in the corpus samples for each of the connectives individually; however, as mentioned, the connective *so* has emerged as the most probable marker of English forward causality overall. While the exceptionally versatile nature of *so* can at least partly account for this finding, the more general conclusion is that subjectivity in the classic Sweetserian sense of distinction between event types may not play a basic role in the English speakers' categorizations of causality. A case in point is the PURPOSE relation (described in the literature as a type of causal relation), which is not analyzable in terms of Sweetser's (1990) subjective-objective distinction due to its intermediate nature between real-world and hypothetical events, yet the relation is frequently marked with the subjective *so* (Andersson & Spenader, 2014). Consequently, it can be concluded that the functions of other connectives in the contexts that could be marked with *so*, seem to be a matter of explicitness bound to specific purposes and/or contexts of use.

On that note, two other strong tendencies that emerge from the predictive analysis are the pronounced presence of *as a result* in non-volitional relations in academic prose (27%) and the strong preference for *therefore* in the same domain (across all relations except for Non-volitional RESULT). While we know from the sample analysis that these particular combinations of Relation and connective are rather typical, given that *so* can cover most (if not all) of these functions, the predictive results are quite revealing as to the reliability of the corpus findings. Further, as mentioned above, all analyzed connectives do occur outside of their prototypical text environment. Since we did not identify any dependencies in terms of the language domain and the relation the connectives most commonly signal, in which case the non-prototypical uses could be bound to some specific register purposes,[15] we argue that the role of causal discourse connectives in English is not confined to precisely marking different types of causality but can be negotiated based on the speaker's specific rhetorical purposes (which may but do not have to be register related). The findings of the predictive analysis for *therefore* and *as a result* suggest that such purposes are often realized via a specific connective phrase, even as infrequent as *as a result*. This is, in fact, also the case for *for this reason*, which, presumably because of its paucity in discourse, does not exhibit any non-negligible preferences in any language domain. Yet, according to the cross-domain results (Figure 7 above), the phrase is relatively common in academic prose, particularly in epistemic and volitional relations.

One telling example of how a connective can be used outside its prototypical discourse environment is the domain of speech acts. As mentioned, except for *so*, the remaining phrases are not compatible with the environment of this subjective relation without additional textual features. Even so, their role marking speech acts is very limited. Consider :

(9) This is because it is much more difficult to recognize being too high than being a little on the low side. *For this reason*, I would like to request the presence of another Adjudicator. (BNC: A0H 1164)

As we see here, an overt performative is needed to convey the illocutionary force of a request. The bare connective would not contribute the requested level of subjectivity in this case, which probably is related to the strength of the involved illocutionary force. According to Sbisa (2001), illocutionary force is weaker for assertions[16] and stronger for imperatives or questions, where only *so* was found.

The interesting aspect, however, is the layer of meaning that the connective does contribute to the relation and hence the underlying purpose of its usage. The role of *for this reason* in (9) above matches its function in volitional causality, where it tends to point back to the very reason for the SoC's action, the difference being that in (9) the action takes place at the level of linguistic events. Similarly, *as a result* in the assertion below is used according to its fundamental discourse function of marking factual events, and thus can be said to emphasize the factual (constative) nature of the subjectively conveyed situation:

(10) I mean er our programme researcher Sophie er is er is not with us today and *as a result* we're a bit topsy and turvy (...) (BNC: KM2 167)

While these observations suggest that in subjective contexts objective connectives lose their role of causality operators, undergo subjectification, and signal the speaker's reasoning, the restricted occurrence of *as a result* and *for this reason* in such contexts confirms that they are deployed for very specific rhetorical purposes. As we would like to see it, the connectives are deployed for hearer-centered, intersubjective (Traugott, 2010; Verhagen, 2008), and argumentative purposes. As to (9), while the first clause is the premise for the speaker to pursue the request in the RESULT clause, which is a classic speech-act, and hence the relation could have been signaled with *so*, the rhetorical goal of *for this reason* is to convince the recipient that the belief/evidence in the first clause is true and constitutes the very reason for the ensuing request. Similarly, *as a result* in (10), can be seen to affirm the speaker's current reality and, consequently, aim at persuading the hearer that the state in the second clause is a tangible result (sic) of the situation in the first clause, as opposed to a mere expression of belief. Interestingly, this example comes from informal speech (Leisure domain), and so it cannot be argued that the high-register connective use was induced by the register convention .

These observations seem to be in line with Kamalski et al.'s (2008) findings on the stronger persuasive effect of the objective connectives in comparison to their subjective counterparts (induced by the hampering effect of the writer's intention explicit with subjective markers; Kamalski et al., 2008, p. 556). Several other studies have also demonstrated that shared knowledge has an influence on persuasion (e.g., Wiley, 2005), which seems to account for both (9) and (10), where the connective function may be geared at sanctioning the common ground. This is particularly true of *for this reason*, which has the anaphoric ability to point back to the (primarily real-world) cause in the preceding segment. While these hypotheses should be tested empirically, the idea that the objective connectives in English (and possibly in other languages) are used to strengthen the persuasive effect of subjective relations seems a viable theory of the reason for their occurrences in certain contexts.

Finally, the role of the subjective connectives in real-world causal relations could be governed, *mutatis mutandis*, by the same principle. Consider the following example:

(11) He was then released on bail and, as is the custom, the police required conditions of his bail that he should not go back to his girlfriend's address to prevent any possibility of any further offending. He *therefore* has been living with friends, sleeping on their floors (. . .) (BNC: HUU 167)

In the present annotation framework, (11) was coded as one of the very few instances of *therefore* in non-volitional relations,[17] assuming that the role of the connective in this factual context is unlikely to convey the speaker's subjective point of view. However, based on the inherently inferential nature of the phrase, its function could be interpreted as a hearer-oriented (intersubjective) modulation of the proposition meant to engage the addressee in negotiating the availability or interpretation of given evidence. Given the stable subjectivity profile of *therefore*, such a functional aptitude is quite remarkable, as it would mean that the connective itself can convey the intended level of subjectivity in an otherwise objective context. Experimental studies could be testing the scale and nature of this potential ability.

Nevertheless, the subsidiary research question on the relationship between discourse connectives and modality suggests that the presence of modal verbs does not influence the connective choice but rather pertains to the relation type and domain of language use. The frequencies of the two types of modality differed in speech and writing; however, as the predictive analysis shows, the weakly subjective Modality Type 2, is the most common choice across all language domains, except for the Public domain of spoken discourse. Given that most of the analyzed contexts are not modalized, and since modality Type 1 comprises subjective high-commitment expressions, in this formal high-stake domain, modality is likely to accompany the connective for specific rhetorical goals:

(12) They just couldn't organize anything <-|-> and *as a result* that's why we'll probably never get Kingmaker or the Ned's Atomic Dustbin. (BNC: HYM158)

Note that in (12) (as in (10) above), *as a result* contributes a factual layer to the context (in fact, signaled by the clearly inferential *that's why*), thus persuasively calibrating the meaning of the epistemic relation toward the speaker's objective reality, where A leads to B.

What our results therefore indicate is that despite the general tendency to mark forward causality in English with *so*, there are connectives that are quite strongly specialized signaling the distinction between objective and subjective causal event types. Moreover, as our sample analysis shows, despite its ambiguous character, the connective *so* is clearly associated with subjective contexts. The other phrases, however, may be even regarded as cues that the recipient will use to infer the presence of certain relations, to the extent that their occurrence in less prototypical discourse environments contributes an additional layer of meaning to the relation. As has been argued here, the underlying reason for such uses often pertains to the speaker's rhetorical goals.

The remaining question pertains therefore to more precise categorizations of the involved subtle distinctions between gradient senses and functional clines (as demonstrated in (9)–(12) above). A recent study of the French connectives *car* and *parce que* (Blochowiak et al., 2020) shares the belief pursued in the current discussion that the nature of the relationship between discourse connectives and subjectivity/objectivity requires refinements of Sweetser's (1990) categorizations of causality. The authors therefore propose a relevance-theoretic account[18] according to which the same causal relation can be considered objective at the level of its basic explicature (e.g., CAUSE) and subjective at the level of its higher order explicature (e.g., explanation of the CAUSE). However, the framework can tackle the problem of multiple layers of meaning only for some relations, and it does not yet capture more complex non-propositional effects, such as persuasive senses, discussed here. Also, the question of the role of the connective itself in relation interpretation is still open.

In summary, our study demonstrates that while English causal connectives do specialize in signaling specific discourse relations, their contribution to subjectivity construal is negotiable and may even be context-sensitive. Consequently, what we suggest is that the roles of connectives in discourse should be described not only in terms of their functions *per se* but rather in terms of their potential for cueing conceptual structures of causality, inferentiality, and argumentation. This kind of a function potential perspective should facilitate a richer and more explicatory description of the connectives, both their fundamental structures and their roles in attested language use.

Notes

1. That study distinguishes between the cases of the connective introducing "given" and "new" information, which is not relevant here, as forward causal connectives cannot be preposed: **So she stayed at home, Anna was ill*. The information flow is, therefore, unlikely to matter in subjectivity construal of forward causal relations.
2. Domains of language use could also be described as genres; however, we follow the categorizations and labels of the British National Corpus, which is where our samples come from.
3. Extension of results from the current corpus to larger or different populations is outside the scope of this study.
4. One example of a predictive perspective is Divjak and Arppe (2013) analysis based on polytomous models for predicting how prototypes for near-synonymous verbs are formed at different levels of abstraction. That methodology, however, assumes sampling methods different from those applied in our corpus investigation.
5. Since the automatic BNC tagger is not reliable for semantic distinctions (e.g., it treats multiword phrases such as *for this reason* as separate word tokens), a large part of the analysis was manual and led to the exclusion of sentence-medial instances, such as "Shakespeare is the most widely known and read of the classical playwrights and it is for this reason that a piece from one of his plays is nearly always obligatory at a drama school audition." (BNC: A06 243).
6. This idea has been systematically confirmed in empirical studies (e.g., Pander Maat & Degand, 2001; Sanders & Spooren, 2015; Scheibman, 2002; Traugott & Dasher, 2002).

7. Our study is loosely inspired by Sanders et al.'s (2012) idea of perspective blend. In fact, many existing analyses omit this category; however, we believe it may be interesting in certain registers.
8. For a more detailed discussion on perspective blending, see Sanders et al. (2012).
9. Cf. Stukker and Sanders (2009) investigation of Dutch newspaper texts, where modal auxiliaries are analyzed on a par with stance markers and attitudinal markers, under a broader category "Subjectivity elements".
10. The current study does not use the label "domain" with reference to the types of coherence relation types as distinguished by Sweetser (1990), which is quite common in the literature, but applies the label to the domains of use as specified in the BNC.
11. See Wikipedia "Simpson paradox" for examples.
12. The BNC corpus designers have not assigned these conversations to any domain. The admixture of the conversational component of the corpus contributes to the frequencies for *so* at 58%, and for *therefore* at 53%.
13. All results discussed in this section are to be found in Appendix 2, and so all abbreviations (e.g., A.2.1) concern the specific subsections of this appendix.
14. Detailed in Supplementary material (Sections S.III.4 and S.IV.4).
15. For instance, we could be finding that *as a result*, prototypically marking non-volitional events, in academic language changes its profile and signals mostly conclusions, based on the factual nature of the involved reasoning. This was, of course, not the case.
16. Understood here as a speech act where the speaker claims the situation holds.
17. This is how the relation would be interpreted without the connective, that is, A and as a result B.
18. The study was published 1 month before final submission of the current article and so the proposed framework, albeit relevant, cannot be discussed here in detail. Interestingly, this very article mentions English as a language with no connective specializations, based on the multifunctional properties of *because*.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Andersson, M. (2019). Subjectivity of English connectives. A corpus and experimental investigation of forward causality signals in written language. In I. Recio & O. Loureda (Eds.), *Methodological approaches to discourse markers*. John Benjamins Publishing Company

Andersson, M., & Spenader, J. (2014). Result and purpose relations with and without 'so'. *Lingua*, *148*, 1–27. https://doi.org/10.1016/j.lingua.2014.05.001

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British national corpus with SARA*. Edinburgh University Press.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge University Press.

Benamara, F., Taboada, M., & Mathieu, Y. (2017). Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, *43*(1), 201–264. https://doi.org/10.1162/COLI_a_00278

Bestgen, Y., Degand, L., & Spooren, W. (2006). Toward automatic determination of the semantics of connectives in large newspaper corpora. *Discourse Processes*, *41*(2), 175–193. https://doi.org/10.1207/s15326950dp4102_4

Biber, D., Conrad, S., & Repen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., Johansson, S., Conrad, S., Finegan, E., & Leech, G. (2002). *Longman grammar of spoken and written English*. Longman.

Blochowiak, J., Grisot, C., & Degand, L. (2020). What type of subjectivity lies behind French causal connectives? A corpus-based comparative investigation of car and parce que. *Glossa: A Journal of General Linguistics*, *5*(1), 50, 1–36. https://doi.org/10.5334/gjgl.1077

Breindl, E., & Walter, M. (2009). *Der Ausdruck von Kausalität im Deutschen: Eine korpusbasierte Studie zum Zusammenspiel von Konnektoren, Kontextmerkmalen und Diskursrelationen (Arbeiten und Materialien zur deutschen Sprache)*. Institut für Deutsche Sprache.

Degand, L., & Fagard, B. (2012). Competing connectives in the causal domain: French car and parce que. *Journal of Pragmatics*, *44*(2), 154–168. https://doi.org/10.1016/j.pragma.2011.12.009

Degand, L., & Pander Maat, H. (2003). A contrastive study of Dutch and French causal connectives on the speaker involvement scale. In A. Verhagen & J. van de Weijer (Eds.), *Usage-based approaches to Dutch* (pp. 175–199). LOT 2003.

Divjak, D., & Arppe, A. (2013). Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive Linguistics*, *24*(2), 221–274. https://doi.org/10.1515/cog-2013-0008

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistic Association*, *89* (425), 190–200.https://doi.org/10.2307/2291215

Jaszczolt, K. M. (2003). The modality of the future: A default-semantics account. In P. Dekker & R. van Rooy (Eds.), *Proceedings of the 14th Amsterdam colloquium* (pp. 43–48). ILLC, University of Amsterdam.

Kamalski, J., Lentz, L., Sanders, T., & Zwaan, R. A. (2008). The forewarning effect of coherence markers in persuasive discourse: Evidence from persuasion and processing. *Discourse Processes*, *45*(6), 545–579. https://doi.org/10.1080/01638530802069983

Knott, A., & Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, *30*(2), 135–175. https://doi.org/10.1016/S0378-2166(98)00023-X

Levshina, N., & Degand, L. (2017). Just because: In search of objective criteria of subjectivity expressed by causal connectives. *Dialogue & Discourse*, *8*(1), 132–150. https://doi.org/10.5087/dad.2017.105

Li, F. (2014). *Subjectivity in Mandarin Chinese. The meaning and use of causal connectives in written discourse* [Unpublished doctoral dissertation]. University of Utrecht.

Lyons, J. (1982). Deixis and subjectivity. Loquor, ergo sum? In R. J. Jarvella & W. Klein (Eds.), *Speech, place and action: Studies in deixis and related topics* (pp. 101–124). John Wiley and Sons Ltd.

Meier, E. (2002). Causal subordination in English and Norwegian. *Nordic Journal of English Studies*, *1*(1), 33–64. https://doi.org/10.35360/njes.90

Pander Maat, H., & Degand, L. (2001). Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics*, *12*(3), 211–245. https://doi.org/10.1515/cogl.2002.002

Pit, M. (2006). Determining subjectivity in text: The case of causal backward connectives in Dutch. *Discourse Processes*, *41*(2), 151–174. https://doi.org/10.1207/s15326950dp4102_3

Sanders, J., Sanders, T., & Sweetser, E. (2012). Responsible subjects and discourse causality. How mental spaces and perspective help identifying subjectivity in Dutch backward causal connectives. *Journal of Pragmatics*, *44*(2), 191–213. https://doi.org/10.1016/j.pragma.2011.09.013

Sanders, T. (2005). Coherence, causality and cognitive complexity in discourse. In M. Aurnague & M. Bras (Eds.), *Proceedings of the first international symposium on the exploration and modelling of meaning* (pp. 31–46). Universite de Toulouse-le-Mirail.

Sanders, T., & Spooren, W. (2015). Causality and subjectivity in discourse - the meaning and use of causal connectives in spontaneous conversation, chat interactions and written text. *Linguistics*, *53*(1), 53–92. https://doi.org/10.1515/ling-2014-0034

Sanders, T. J. M., Spooren, W., & Noordman, L. (1992). Towards a taxonomy of coherence relations. *Discourse Processes*, *15*(1), 1–35. https://doi.org/10.1080/01638539209544800

Santana, A., Nieuwenhuijsen, D., Spooren, W., & Sanders, T. (2017). *Causality and subjectivity in Spanish connectives: Exploring the use of automatic subjectivity analyses in various text types*. Discours Revue de Linguistique, Psycholinguistique et Informatique.

Sbisa, M. (2001). Illocutionary force and degrees of strength in language use. *Journal of Pragmatics*, *33*(12), 1791–1814. https://doi.org/10.1016/S0378-2166(00)00060-6

Scheibman, J. (2002). *Point of view and grammar. Structural patterns of subjectivity in American English conversation*. Benjamins.

Stukker, N., & Sanders, T. (2009). Another('s) perspective on subjectivity in causal connectives: A usage-based analysis of volitional causal relations. *Discours, Linearization and Segmentation in Discourse*, *4*(Special issue), 1–33. https://doi.org/10.4000/discours.7260v

Stukker, N., & Sanders, T. (2012). Subjectivity and prototype structure in causal connectives: A cross-linguistic perspective. *Journal of Pragmatics*, *44*(2), 169–190. https://doi.org/10.1016/j.pragma.2011.06.011

Sweetser, E. (1990). *From etymology to pragmatics*. Cambridge University Press.

Traugott, E. (2010). Revisiting subjectification and intersubjectification. In K. Davidse, L. Vandelanotte, & H. Cuyckens (Eds.), *Subjectification, intersubjectification and grammaticalization* (pp. 29–70). De Gruyter Mouton.

Traugott, E., & Dasher, R. (2002). *Regularity in semantic change*. Cambridge University Press.

Verhagen, A. (2008). Intersubjectivity and the architecture of language system. In J. Zlatev, P. T. Racine, C. Sinha, & E. Itkonen (Eds.), *The shared mind: Perspectives on intersubjectivity* (pp. 307–331). John Benjamins Publishing Company.

Wiley, J. (2005). A fair and balanced view at the news: What affects the memory for controversial arguments? *Journal of Memory and Language*, *53*(1), 95–109. https://doi.org/10.1016/j.jml.2005.02.001

Zufferey, S. (2012). "Car, parce que, puisque" revisited: Three empirical studies on French causal connectives. *Journal of Pragmatics*, *44*(2), 138–153. https://doi.org/10.1016/j.pragma.2011.09.018

Zufferey, S., & Cartoni, B. (2012). English and French causal connectives in contrast. *Languages in Contrast*, *12*(2), 232–250. https://doi.org/10.1075/lic.12.2.06zuf

Zufferey, S., Mak, W., Verbrugge, S., & Sanders, T. (2018). Usage and processing of the French causal connectives 'car'and 'parce que'. *Journal of French Language Studies*, *28*(1), 85–112. https://doi.org/10.1017/S0959269517000084

Appendix 1. Figures

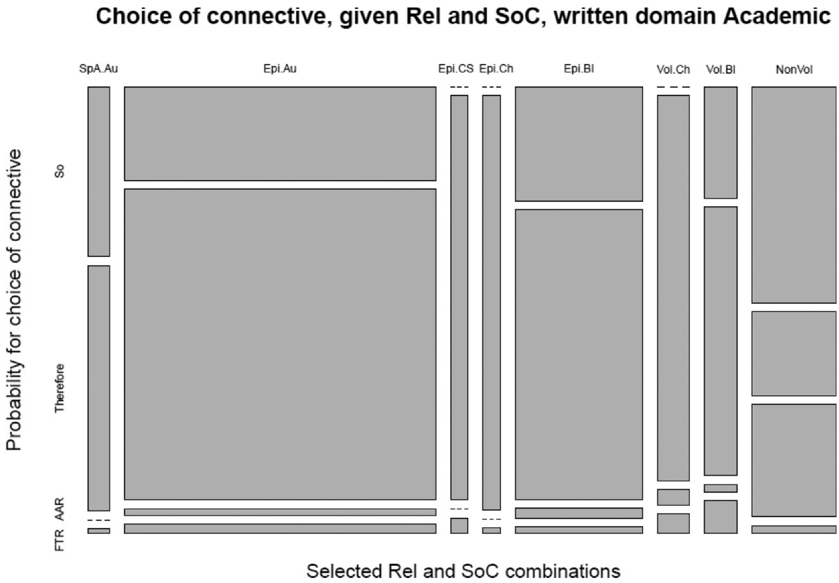

Figure A1. Choice of connective, given Relation and SoC, Academic domain, written discourse.

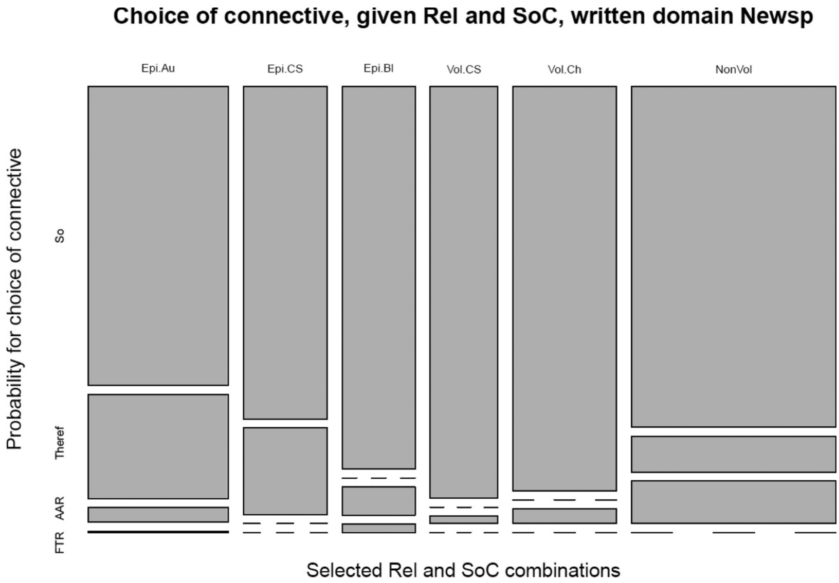

Figure A2. Choice of connective, given Relation and SoC, Newspaper domain, written discourse.

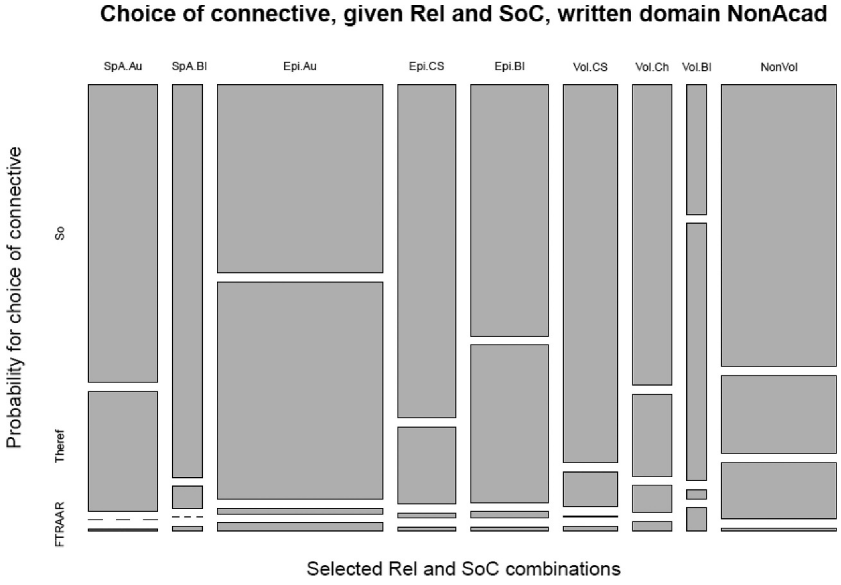

Figure A3. Choice of connective, given Relation and SoC, Non-academic domain, written discourse.

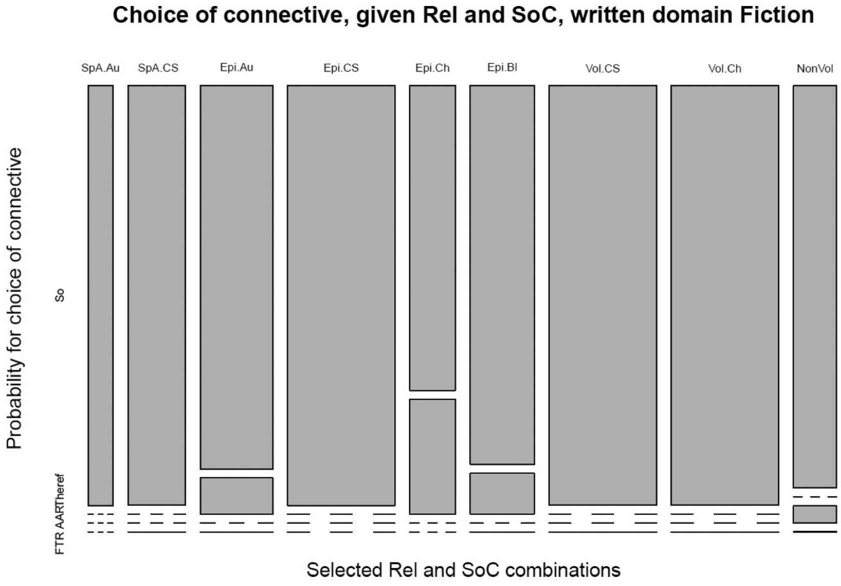

Figure A4. Choice of connective, given Relation and SoC, Fiction domain, written discourse.

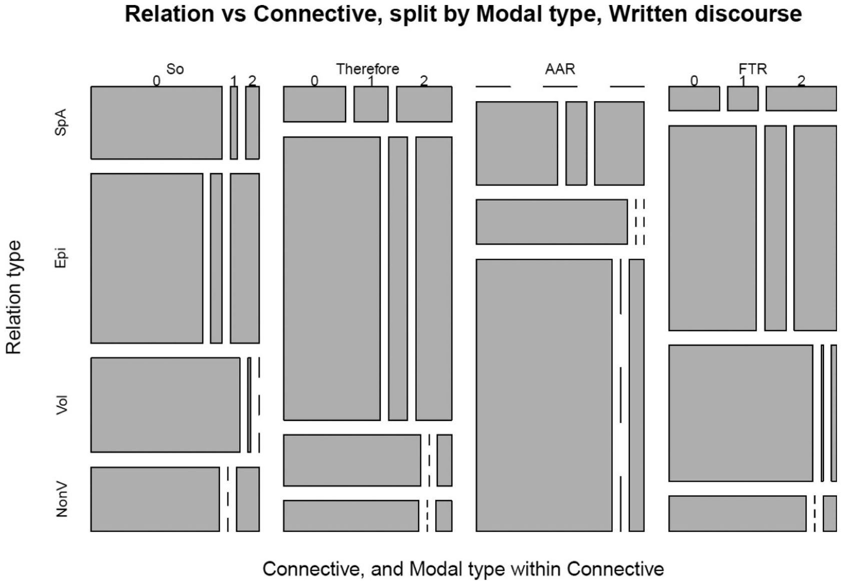

Figure A5. Relation vs connective, divided by modality type, written discourse.

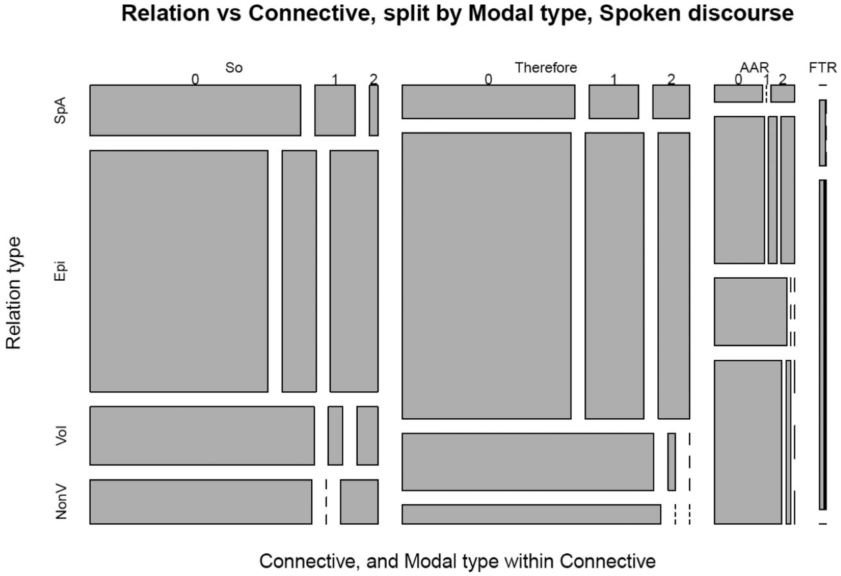

Figure A6. Relation vs connective, divided by modality type, spoken discourse.

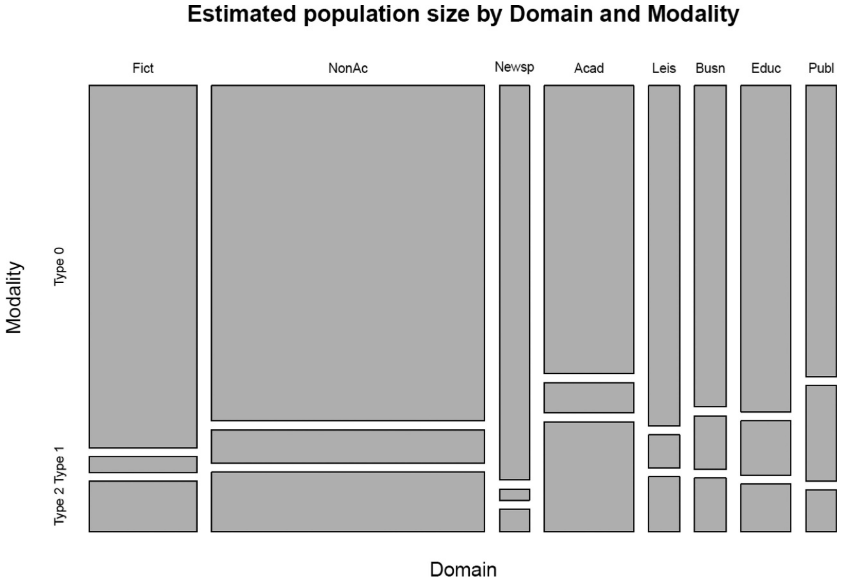

Figure A7. Estimated population size by domain and modality type.

Appendix 2. Log-linear models for connective frequencies

Log-linear models can describe how different factors (Connective, Domain, Rel, SoC) influence the counts (or frequencies) for the combinations of factor levels possible. As described at the beginning of the Results section in the manuscript, our interest is to see to what extent the expected counts can have a *multiplicative structure* –as expressed in the factors listed above. If factors are not multiplicative, we say that they *interact*. A partially multiplicative model, labelled Model 1, was also introduced in the Results section, and will be in focus below. According to this model, for a particular connective considered, the expected counts and the corresponding probabilities *Pr*(Domain, Rel, SoC) can be factorized as:

Model 1: $Pr(\text{Domain, Rel, SoC}) = Pr(\text{Domain}) \times Pr(\text{Rel, SoC})$,

This means that *the (conditional) probability table for (Rel, SoC) within domain is the same for all domains*, or in other words, the probability for (Rel, SoC) is independent of domain.

In order to avoid too many combinations with zero frequency, and not (on average) very low counts, an analysis was carried out for each connective separately, but only for the largest domains of occurrence for the connective in question. Some combinations, however, yielded very low counts, which are involved in the analysis. This implies that model test *p*-values are approximate and should be regarded as crude.

Complete data can be found in the Supplementary material. Below, the connective *as a result* has been abbreviated to AAR, while *for this reason* – to FTR.

2.1. So, written discourse

The data analyzed represent Fiction and NonAc domains and all 3×4 (Rel, SoC) combinations (Non-Vol excluded). Table A1 below shows the corresponding data tables together with a model-fitted table representing Model 1.

Successive model simplification, starting from the saturated model (i.e. no assumed structure), showed first no indication of any three-factor interaction, and next no indication of any interaction between Domain and Rel (both *p*-values 0.5). However, Domain interacted strongly with SoC ($p<0.001$). This implies that we should be able to see deviations from Model 1 in Table A1, and they are easily discernible in the summary line of Table A1, as the four frequencies are compared between domains. The Author and Blend categories are substantially more frequent in the NonAc domain than in the Fiction domain (together 0.60 vs 0.27), and correspondingly less frequent for Current Speaker and Character (0.41 in NonAc vs 0.73 in Fiction). This is the feature behind the statistically significant interaction between Domain and SoC for *so*.

Table A1. Fitted *Pr*(Rel, SoC) according to Model 1, and the corresponding data per domain, representing connective *so*, written discourse, Fiction and Non-Academic domains (sample sizes 70 and 111).

So, written	Fitted, Model 1				Data, domain Fict				Data, domain NonAc			
Rel \ SoC	Au	CS	Ch	Bl	Au	CS	Ch	Bl	Au	CS	Ch	Bl
SpA	0.10	0.06	0.01	0.06	0.04	0.10	0.01	0.01	0.14	0.03	0.01	0.08
Epi	0.18	0.15	0.03	0.12	0.11	0.19	0.06	0.10	0.22	0.14	0.01	0.14
Vol	0	0.16	0.12	0.01	0	0.19	0.19	0	0	0.14	0.08	0.02
Sum	0.28	0.37	0.16	0.19	0.16	0.47	0.26	0.11	0.36	0.31	0.10	0.24

The interaction between Rel and SoC (on average over domains) is of an even much stronger magnitude (p<<0.001). One glance at the structure within any of the three parts of Table 1 suffices to see that the probabilities are very far from multiplicative over Rel and Soc. The most obvious feature is the absence of the (Vol, Au) combination, where, given the other parts of the Au column and the Vol row, we would have expected a frequency of 0.1 or more.

2.2. Therefore, written discourse

Table A2 for *therefore* is analogous to Table A1 for *so*, but the two 3×4 tables of data now represent the two domains NonAc and Acad and all (Rel, SoC) combinations except the Non-Vol. The left part is the fitted Model 1, which is a form of average over the two tables for NonAc and Acad.

Table A2. Fitted *Pr*(Rel, SoC) according to Model A.1, and the corresponding data per domain, representing *therefore*, written discourse, Non-Academic and Academic domains (sample sizes 119 and 103).

Theref, wr.	Fitted, Model 1				Data, domain NonAc				Data, domain Acad			
Rel \ SoC	Au	CS	Ch	Bl	Au	CS	Ch	Bl	Au	CS	Ch	Bl
SpA	0.07	0.00	0.00	0.01	0.11	0.01	0.01	0.01	0.03	0	0	0.02
Epi	0.50	0.05	0.03	0.18	0.47	0.06	0.03	0.16	0.53	0.04	0.04	0.20
Vol	0.02	0.01	0.05	0.06	0.02	0.03	0.04	0.07	0.02	0	0.07	0.05

Data indicated no three-factor interaction (p-value 0.5), so this feature was first omitted from the model. In the successive model testing, the interaction between Domain and SoC could be deleted (p-value 0.3). The model was further simplified to the form of Model 1, by neglecting the interaction between Domain and Rel (p-value 0.07). Table A2 shows that the fitted model agrees quite well with the data for both domains. As in the case of *so*, the interaction between Rel and SoC is strong here (p<<0.001). Most notable is that when Relation category is SpA or Epi, Author dominates for SoC, whereas for Rel category Vol, SoC categories Character and Blend dominate together.

2.3. As a result, written discourse

For *as a result*, only NonAc domain is large enough to be analyzed. This is because all Non-Vol instances are left out of the analysis. For comparison with Tables A1 and A2 for *so* and *therefore*, the corresponding Table A3 for *as a result*, including all non-empty domains, is presented below (even though the Newsp and Acad domains are too small for reliable statistical comparisons with NonAc). Comparing Table A3 below with Table A2, we see clear similarities, in particular the high frequencies of (Eps, Au) and (Epi, Bl), but also striking differences, for example the total lack of Rel=SpA for *as a result*, and the high frequency of (Vol, Ch).

Table A3. Data per domain for connective *as a result*, written discourse, Non-Academic, Newspaper and Academic domains (sample sizes 52, 12 and 16, respectively).

AAR, written	Data, domain NonAc				Data, domain Newsp				Data, domain Acad			
Rel \ SoC	Au	CS	Ch	Bl	Au	CS	Ch	Bl	Au	CS	Ch	Bl
SpA	0	0	0	0	0	0	0	0	0	0	0	0
Epi	0.29	0.10	0.06	0.15	0.33	0	0	0.33	0.50	0	0	0.31
Vol	0	0.02	0.33	0.06	0	0.08	0.25	0	0	0	0.13	0.06

2.4. For this reason, written

Table A4 for *for this reason* is analogous to Table A2 for *therefore*, with the same two domains NonAc and Acad. The left part is the fitted Model 1, which is a form of average over the two tables for NonAc and Acad. Results for *for this reason* are quite similar to those for *therefore* above. The saturated model could be simplified to Model 1 (p-value 0.9), in which the interaction between Rel and SoC is strong (p <<0.001). Notable is that this interaction feature is similar to what we saw for *therefore*: when Relation category is Epi, Author dominates as SoC, whereas for Rel category Vol, Author is the least frequent SoC category.

Table A4. Fitted *Pr*(Rel, SoC) according to Model 1, and the corresponding data per domain, representing *for this reason*, written discourse, Non-Academic and Academic domains (sample sizes 141 and 74).

FTR, written	Fitted, Model 1				Data, domain NonAc				Data, domain Acad			
Rel \ SoC	Au	CS	Ch	Bl	Au	CS	Ch	Bl	Au	CS	Ch	Bl
SpA	0.03	0.00	0.00	0.03	0.04	0	0.01	0.04	0.01	0	0	0.01
Epi	0.40	0.06	0.01	0.10	0.38	0.06	0.01	0.09	0.43	0.04	0.01	0.12
Vol	0.05	0.07	0.10	0.14	0.05	0.08	0.11	0.13	0.04	0.07	0.09	0.16

2.5. General conclusions about domains, for written discourse

For *for this reason* and *therefore*, considering their larger domains NonAc and Acad, we conclude that the (Rel, SoC) probability tables are quite stable across domains. For *as a result*, the scarcity of data (except for non-volitional relations) makes similar comparisons between domains meaningless, but the data look stable across domains. For *so*, some differences between domains Fict and NonAc (largest) are established; however, with a coarser measure, the domains are similar also for *so*.

2.6. So, spoken discourse

Only three SoC categories are considered below, because the Blend category is overall almost absent. Three domains are compared, Leisure (95 instances), Business (48), Education (54). Table A5 shows the frequencies in the three domains, together with a fitted version of Model 1, i.e. allowing interaction between Rel and SoC, but no interaction involving Domain.

Successive model simplification from the saturated model showed no indication (p-value 0.5) of any three-factor interaction. We could next delete the interaction between Domain and Rel or the one between Domain and SoC as insignificant (p-values both 0.2), but the remaining interaction was in both cases statistically significant (p=0.003 and p=0.004, respectively). Thus, there is a statistically certain deviation from Model 1, to be considered below. The interaction between Rel and SoC, allowed in Model 1, is also significant, and of even stronger magnitude also in this case (p<<0.001).

Table A5. Fitted *Pr*(Rel, SoC) according to Model 1, and the corresponding data per domain, representing connective *so*, spoken discourse, domains Leisure, Business and Education.

So, spoken	Fitted, Model 1			Data, Leisure			Data, Business			Data, Education		
Rel \ SoC	Au	CS	Ch	Au	CS	Ch	Au	CS	Ch	Au	CS	Ch
SpA	0.12	0.04	0.01	0.08	0.04	0.01	0.12	0.02	0	0.17	0.04	0
Epi	0.32	0.32	0.03	0.26	0.31	0.02	0.29	0.38	0.04	0.44	0.30	0.02
Vol	0	0.11	0.07	0	0.15	0.13	0	0.15	0	0	0.02	0.02
Sum	0.44	0.46	0.10	0.35	0.49	0.16	0.42	0.54	0.04	0.61	0.35	0.04

The differences between domains seen in Table A5 can be ascribed to variation in either the distribution over Rel or in the distribution over SoC, and because of this indeterminacy, we must look at the specific combinations of Rel and SoC to understand the source of the substantial differences:

Leisure domain includes much more of (Vol, Ch) than Business and Education;
Leisure domain includes much less of (SpA, Au) than Education has, Business in-between;
Education includes much more of (Epi, Au) than Leisure and Business;
Education includes much less of (Vol, CS) than Leisure and Business.

2.7. Therefore, spoken discourse

All 3×4 = 12 category combinations of Rel and Soc are compared, and all 4 domains (smallest sizes 34 and 36 for Leisure and Business, size 61 for Education, and Public of size 107, total 238). As with *therefore* in writing, Model 1 fits the data reasonably well (p = 0.11 for the simplification from the saturated model to Model 1). Further, the hypothesis of multiplicativity between Rel and SoC must clearly be rejected (p<<0.001) also in speech. Table A6 shows the fitted model and the two largest domains for all (Rel, SoC) combinations. In Table A7 further below, *therefore* spoken is compared with *therefore* written discourse. (not shown in the table)

Table A6. Fitted *Pr*(Rel, SoC) according to Model 1, and the corresponding data per domain, representing connective *therefore*, spoken discourse, Education and Public domains.

Theref, sp.	Fitted, Model1				Data, domain Education				Data, domain Public			
Rel \ SoC	Au	CS	Ch	Bl	Au	CS	Ch	Bl	Au	CS	Ch	Bl
SpA	0.05	0.03	0	0.00	0.05	0	0	0	0.07	0.04	0	0
Epi	0.38	0.31	0.01	0.06	0.48	0.26	0	0.10	0.37	0.31	0.01	0.07
Vol	0	0.08	0.04	0.03	0	0.07	0.05	0	0	0.10	0.01	0.02

2.8. As a result and For this reason, spoken

These connectives are omitted from this analysis because of their scarcity in speech.

2.9. General features, for all connectives in speech and writing

For all connectives, there is a strong interaction between the Rel and SoC factors. Most common (Rel, SoC) combinations are (Epi, Au) and (Epi, CS). SoC = Character (Ch) is rare, and (Vol, Au) is usually absent. These observations appear to represent general features of the language, whereas other (Rel, SoC) combinations may be frequent for one connective and rare for another.

2.10. Comparison of *therefore* in written and spoken discourse, based on Model 1:

For *therefore* in both written and spoken discourse, the observed data fit models of type Model 1. Hence, their (Rel, SoC) tables may be compared, even though the domains are not the same. This is possible by simply neglecting the Domain factor. In Table A7, the corresponding (Rel, SoC) frequency tables are shown, together with their difference. The sample sizes are 231 and 238, respectively, and the fitted table under the hypothesis of no difference is almost precisely the simple average of the tables for written and spoken discourse. (not shown in the table)

Table A7. Cross-domain (Rel, SoC) frequency tables for connective *therefore*, written and spoken discourse, and their difference.

Therefore	Therefore written				Therefore spoken				Difference Wr–Sp			
Rel \ SoC	Au	CS	Ch	Bl	Au	CS	Ch	Bl	Au	CS	Ch	Bl
SpA	0.07	0.00	0.00	0.01	0.05	0.03	0	0.00	0.02	-0.02	0.00	0.01
Epi	0.50	0.05	0.03	0.18	0.38	0.31	0.01	0.06	0.11	-0.25	0.03	0.11
Vol	0.02	0.01	0.05	0.06	0	0.08	0.04	0.03	0.02	-0.07	0.01	0.03

A model with the three factors Rel, SoC and discourse Type (written vs spoken) showed no significant 3-factor interaction (p=0.2), nor an interaction between Rel and Type (p=0.7), but a highly significant interaction between SoC and Type (p<<0.001). In other words, the distribution between SoC types certainly differs between written and spoken discourse.

The striking difference between written and spoken discourse is the high frequency of SoC = Current Speaker in the spoken use of this connective, relative to the written use, in particular when Rel type = Epistemic.

A corresponding comparison between written and spoken *so* is not carried out, because Model 1 was rejected in both cases, implying that the table depends on the domain, and the domains are different in written and spoken discourse.

Appendix 3

Statistical uncertainties related to RQ1 and RQ2

A.3.1. Uncertainty in domain sizes

For each connective, 'crude' domain sizes, counting all instances, are known. Domain sizes for target instances are not known, however, but must be estimated. This is done by reducing the crude number by the estimated target proportion for the connective in question (and sometimes for the domain in question). The data for the estimated target proportions are given in the Supplementary material, Table SA1. They are more or less uncertain. Some adjustments were made when they appeared to differ between domains (in particular for connective *so*, see below). In the predictive approach (inverse analyses, RQ2), the target proportions play an important role, since they are different for different connectives, whereas they are not at all important for the sample studies undertaken per connective (RQ1). Here are two examples:

(1) *So* Spoken, 250 target instances of 540; 250/540 = 0.46, or 46%;

(2) *Therefore* Written, 250 target instances of 276; 250/276 = 0.91, or 91%.

The uncertainty in these percentages, from only sampling randomness, when expressed as standard errors, *s.e.*, is in both cases 2 percentage units, and we write 0.46 ± 0.02 and 0.91 ± 0.02. The corresponding relative errors in the estimated target proportions are ±4% and ±2% of the proportions, respectively. This is not much, and can often be neglected in comparison with other sources of error and uncertainty.

We can often think of the standard error as a measure of confidence, such that with about 95% confidence the true value is in the interval 'estimate ± 2 × *s.e.*'. This comes from the normal distribution. The corresponding interval with one *s.e.* has then about 70% confidence.

For *so* Written, it was realized that the target proportion no doubt varied between domains. This is probably not exclusively the case for *so* Written. Such effects generate an additional error. Substantial such effects for a connective will be observed when comparing what we should expect, if the target proportion did not vary, with the actually observed domain strata sizes from the sampling, but it is not possible to compensate for them with precision, due to the sampling uncertainty. The likely presence of some such systematic errors should be kept in mind when judging the results of the inverse analyses. Here is an example:

(3) For *for this reason* (FTR), Written, we expect that (only) 4 items of the sample of 250 should fall in the Fiction domain, as judged from the crude domain sizes and assuming the same target proportion in all domains. The actual sample size for the Fiction domain was 8, and the probability of getting such a large sample size as 8 (or larger), when 4 was expected, is only about 0.05 = 5%. This small probability is an indication that there might be reason to doubt the assumption of domain-independent target proportion for FTR in the population. However, since FTR is generally quite an infrequent connective, and particularly so in Fiction, it was judged unnecessary to revise the assumption for FTR. On the other hand, it was revised for *so*.

A.3.2. Sampling uncertainty, per connective

This uncertainty directly affects the answers to questions of type RQ1, but also indirectly RQ2 analyses (see later sections).

Sample sizes of 250 are not small, but when we found it necessary to partition the samples between domains, in order to achieve more relevant answers to the questions, we got much smaller samples. It is a pity that the generally largest of the four written domains, NonAc, is of less interest, since it makes the samples from the other domains smaller, but this must be accepted. Here are some examples of Written sample sizes:

(4) *So* Written Fiction 75, *therefore* Written Academic 107

(5) *So* Written: Newsp 20 and Academic 19
(6) *Therefore* Written: Fiction 4 and Newsp 6

The uncertainty not only depends on these sample sizes but also on the population frequency of the characteristic studied, which could for example be the frequency of a specific SoC characteristic or a combination of Rel and SoC. The standard error is highest when the frequency is 1/2 (= 50%), when the standard error is $\frac{1}{2\sqrt{n}}$, where n is the domain sample size. In the examples above, the standard errors in this 'worst case' scenario are
(4) 0.05 – 0.06 (5–6 percentage units);
(5) ≈ 0.12;
(6) 0.20–0.25.

In case **(6)** the standard error is a too simple measure of uncertainty, how- ever, but it is anyhow clear that the uncertainty is huge. When the frequency in question is 0.10 (10%) or 0.90, instead of $\frac{1}{2}$, the corresponding standard errors are smaller by a factor of 0.6:
(4) 0.03 – 0.04;
(5) ≈ 0.07;
(6) 0.12–0.15.
Here are a couple of illustrations:

(7) *Therefore* Written Academic, n=107: SoC=Author, 60 instances of 107
=> proportion 0.56 ± 0.05 (more precisely ±0.048)

(8) *Therefore* Written Academic, n=107: Modal (type 1 or 2) 8+27=35
instances => proportion 0.33 ± 0.05 (±0.046);

Note: Nonmodal is complementary, thus 0.67 ± 0.05 (same *s.e.*)

Many combinations of Rel and SoC has a zero number of observed in- stances. The population proportion may of course be zero (and in some cases it may be natural to assume it is zero), but at the other end, how large can the population proportion realistically be? This depends on the sample size, and let us consider a couple of examples. First, we return to example **(8)** above, with $n = 107$. Suppose a zero is observed (e.g. the combination *{SpA, Ch}*), and that such a zero was not a very unlikely event to happen, more precisely that the probability of a zero was at least 5%. The demand on the population for this to happen is a population proportion of at most 0.012, i.e 1.2%. Thus we can feel pretty sure the *Therefore* Academic cor- pus domain did not have much more than 1% instances of the combination

{SpA, Ch}. A general formula for small proportions is ln(20)/*n*, where ln is the so-called natural logarithm function.

Let us now go to example **(6)**, the small *Therefore* Newspaper domain, which has also a zero for *{SpA, Ch}*, but with a domain sample size of only $n = 6$. Then the domain could of course have much more of *{SpA, Ch}*) without any of them seen in the small sample. In this case the demand on the domain to make this plausible is that the domain population proportion of *{SpA, Ch}* be at most about 40% (that is, it could be quite large).

Note that the 'small' samples for AAR and FTR Spoken are not small as explorations of the corpus. They comprise the whole corpus populations of these connectives. On the other hand we know from these population sizes that AAR and FTR are quite rare in Spoken language, and as a consequence that the BNC is insufficient, if we are interested in the (rare) use of them.

A.3.3. Uncertainty in the 'predictive approach', whole domains

The simplest situation for the predictive approach is when we consider the choice of connective within a certain domain, without specifying Rel, SoC or Modal. For each connective, the domain population sizes are known, as soon as we know the target proportions, because the crude domain population sizes are known (without statistical uncertainty). if we can calculate the domain size for each connective, we immediately also know the probability that a random choice of instance would result in a specific connective. It is just to sum the total number of target instances over the four connectives, and see how large proportion of this sum was taken by each connective. Thus, except for the need for the target proportions, discussed in Section A.3.1, there is no statistics involved. As remarked in Section A.3.1, the sampling errors in the target proportions are small, although for connective *so* there is a variation between domains to consider and adjust for.

A.3.4. Uncertainty in the 'predictive approach', domain parts

When we further specify a category of Rel, SoC, or Modal, or a combination of such categories, the situation gets more complicated. The domain population size of that category must be estimated from the sample data, for each connective separately. We consider two examples jointly, the (Rel, SoC) combinations *{SpA, Au}* and *{Epi, Au}* in the Written Academic domain (cf. Example **(7)** above. We neglect connectives AAR and FTR, which are rarely chosen by the language user and therefore have little *uncertainty* effect on *so* and *therefore*. Also, as we will see, it is enough complicated to make statements of uncertainty about the choice between two connectives. For *so*, the domain sample size is n=19, one instance of *{SpA, Au}* and 8 of *{Epi, Au}*. For *therefore*, the corresponding domain sample size is $n = 107$, with 3 instances of *{SpA, Au}* and 55 of *{Epi, Au}*. In the *{SpA, Au}* category, the probability for choice of *so* is estimated to be 40%, and for choice of *therefore* 59%. In the *{Epi, Au}* category, the corresponding probabilities are 22% and 74%. After elimination of AAR and FTR this probability increases to 23% and 77%, respectively.

As in the previous section, if we had known the total number of in- stances of *so* and *therefore* in the corpus that satisfy the joint specification of domain, Rel and SoC, it would just be to count the proportion of *so* and *therefore* respectively in their total (sum over *so* and *therefore*). The total number per connective satisfying the specification is not known, but it is naturally estimated, by multiplying the domain total with the domain sample relative frequency of instances satisfying the specification of Rel and Soc. From the previous paragraph we have for example the relative frequen- cies 1/19 and 8/ 19 for *so* (*{SpA, Au}*) and *{Epi, Au}*, respectively) and the relative frequencies 3/107 and 55/107 for *therefore*. We see again the problems discussed in Section A.3.1. The sampling uncertainty for *therefore* is small, but for *so* it is quite large, in this case.

Instead of treating the proportions of *so* and *therefore* (the ratios to their sum), that is, their respective estimated probabilities of choice, as described above, we go over to their odds, which has mathematical/statistical advantages. More precisely. the odds for *so* is the ratio of the probabilities of choice of *so* and of 'not-so', (that is of *therefore*). From the odds, we can calculate the probabilities (and vice versa): add 1 to the odds, invert, and subtract from 1 (so odds = 1 corresponds to equal chances, probability = 1/2). The odds for category *{SpA, Au}* is about 4 to 6, more precisely estimated as 0.69. For *{Epi, Au}* it is estimated to be 0.30 The next step is to see if one of the two contributions to the uncertainty dominates over the other, that is if the uncertainty coming from *so* dominates over that coming from *therefore*, or vice versa, or if they contribute equally. Their relative contributions of variance is estimated by $1/y - 1/n$, where y is the number of instances observed in the sample of size n (specified Rel, SoC, domain, connective; formula derivation excluded). In the two examples we get:

{SpA, Au} Estimated variance contribution from *so* = 1 − 1/19 = 0.95; from *therefore* it is only 1/3 − 1/107 = 0.32;

{Epi, Au} Estimated variance contribution from *so* = 1/8 − 1/19 = 0.07; from *therefore* it is only 1/55 − 1/107 = 0.009.

We note that for each (Rel, SoC) combination the contribution from the small domain (for *so*) is substantially higher than that from the relatively large domain (for *therefore*). This is not surprising. We first consider the second (Rel, SoC) combination.

For the *{Epi, Au}* category of the Academic domain, with odds ratio 0.30, we have seen that the contribution to uncertainty from *so* dominates.

The relative variance corresponds to a relative standard error of $\sqrt{0.07} \approx 0.3$. This yields an interval around odds 0.3 by adding and subtracting about 30% of the odds ratio value itself, which yields the odds ratio interval 0.2 to 0.4 . Transforming back to probabilities of choice we get the interval 0.18 to 0.28 around the estimate of 0.23 for choice of *so*, and the complementary values for *therefore*. This is a quite reasonable precision to draw conclusions from.

For the *{SpA, Au}* category of the Academic domain, the situation is more difficult, and the uncertainty in the odds ratio will be much higher. The standard error calculus is not sufficient, cf. the calculation for zero instances toward the end of Section A.3.1, related to example **(6)**.

The statistical error in *so* dominates over the error in *therefore*, so we only consider the former. One instance of *{SpA, Au}* in a domain sample of 19 tells that the probability for choice of *so* can be neither very close to zero nor very high, so we can construct a two-sided confidence interval for the probability by exact calculations in the binomial distribution. We have chosen the 70% confidence level in order to match the ± *s.e.* interval. The confidence interval in this case was (0.0085, 0.167). This was used to get an interval for the odds of *so* versus *therefore*, which was finally transformed to an interval for the probability of *so* (or of *therefore*): 0.10 ≤ Pr(*so*) ≤ 0.68.

With a zero number of instances of *so* (or *therefore*), as for example in Academic domain *{Epi, CS}*, a one-sided confidence interval would have been necessary (and natural). In this case, more care is needed in the calculation of an interval for the odds, since when zero instances are observed, we are comparing with zero (or infinite) estimated odds. Details are omitted.

The uncertainty in choice probabilities is to a large extent controlled by the number of instances of *so* for the discourse context in question. If this number is not very small, a large number of *so* is expected in the corpus, and the probability of choice of *so* cannot be small. When the number of instances of *so* is small, or even zero, the probability of choice of *so* is likely to be quite uncertain. On the other hand, for AAR or FTR the probability of choice will be quite small for almost all discourse contexts, and so will the uncertainty in absolute terms for AAR and FTR, whereas their relative uncertainty is likely to be large.