



Article

Merja Kytö and Erik Smitterberg*

Clausal and phrasal coordination in recent American English

<https://doi.org/10.1515/cllt-2022-0035>

Received April 22, 2022; accepted November 3, 2022; published online November 25, 2022

Abstract: Several studies have shown that there is considerable cross-genre variation as regards what linguistic units tend to be coordinated by *and*. While literate, expository writing favors coordination of phrasal units such as noun phrases, coordinated units are more often clausal (e.g., main or subordinate clauses) in speech-related texts. This difference has been attested in studies that focus exclusively on coordination as well as in macro-level studies of co-variation among a large number of linguistic features. However, this register differentiation has increased over time: studies of Early and Late Modern English point to less pronounced differences among registers than those attested in the present-day language. This study fills a gap in research by considering data on coordination by *and* from the middle of the 20th century, a period that does not belong fully to either Late Modern or Present-Day English, and the late 20th and early 21st century, and thus ties diachronic and synchronic research on register variation in coordination together. We also examine language from films and television in order to complement historical findings for speech-related language with data on registers that arose in the 20th century.

Keywords: coordination; orality; recent English

1 Introduction: register variation and the study of *and*

The present study deals with an intriguing feature, namely variation and change in the use of the coordinating conjunction *and*. While there is a notable tradition in

*Corresponding author: Erik Smitterberg, Department of English, Uppsala University, P.O. Box 527, Uppsala, SE-751 20, Sweden, E-mail: erik.smitterberg@engelska.uu.se

Merja Kytö, Department of English, Uppsala University, Uppsala, Sweden, E-mail: merja.kyto@engelska.uu.se

research on *and* and its uses, there are still gaps in what previous studies have covered. Our study investigates the period between the 1950s and the 2000s, a period neglected so far in the study of *and*, and targets language from films and television in order to complement historical findings for speech-related language with data on registers that arose in the 20th century. To contrast such language use with a non-speech-like yet informal register, we consider data from Magazines. Also, in order to complement previous research, which has mostly focused on British English, we instead consider North American material.

The importance of register variation to the development of language use and individual linguistic features has been established in previous work, and evidence has also been presented on the patterned nature of such variation. The description of a register as such has been specified to cover “three major components: the situational context, the linguistic features, and the functional relationships between the first two components” (Biber and Conrad 2019: 6). The situational context pertains to factors such as the purpose of the text, whether it was spoken or written, and its audience. As linguistic features serve various functions in texts, “they are used to differing extents in different registers, conforming to the situational characteristics of those registers” (Biber and Gray 2016: 70).

The register-based approach is an exciting starting point for the study of language change and requires careful consideration as to the range of registers to be considered when tracing the trajectories of individual linguistic features. In their influential study of diachronic relationships among speech-based and written registers in English, Biber and Finegan (1997) showed that while popular registers aimed at “a wide, general readership” became more oral over the 19th and 20th centuries, “specialist expository registers ... have followed a consistent course towards ever more literate styles” (1997: 272–273) and emphasized the need to look into “texts at both extremes of the ‘oral-literate’ continuum” (1997: 254). The trend towards oral styles has also been in focus in work on colloquialization (e.g., Hundt and Mair 1999; Leech et al. 2009; Smitterberg 2021) and popularization (e.g., Biber and Gray 2012). Similarly, the drift towards literate styles has some affinities with what has been called densification (e.g., Leech et al. 2009; Smitterberg 2021), i.e., a change “that packs more information content into a given number of words” (Leech et al. 2009: 210), or economy (e.g., Biber and Gray 2012).

Biber and Finegan’s (1997) study was carried out within the multidimensional framework, which, despite the variety of periods and data sources considered, has presented cumulative evidence of the oral/literate opposition and its realization “as two fundamentally different ways of constructing discourse: clausal versus phrasal” (Biber and Gray 2016: 84; for details, see their Table 3.2 summarizing the oral/literate dimension in a selection of multidimensional studies of various discourse domains in English, pp. 85–86). In terms of linguistic features, the oral

direction of this dimension comprises, for instance, mental and communication verbs, present tense, progressive aspect, adverbs, stance adverbials, dependent clauses, and discourse markers. Conversely, the literate direction is characterized by nouns, nominalization, attributive adjectives, and prepositional phrases (Biber and Gray 2016: 84). At the oral end, many of the features are those typical of everyday, colloquial usage. As for *and*, the conjunction displays a Janus-faced behavior, its clausal uses aligning with oral features and its phrasal uses with literate features (see below; see also Biber 1988: 89). It thus presents an attractive object for a study of register variation from the perspective of oral versus literate norms for production. Combining synchronic and diachronic perspectives is especially valuable as there is evidence that this register differentiation has increased over time: studies of Early and Late Modern English point to less pronounced differences among registers than those attested in the present-day language.

The present study thus fills a gap in research by considering data on coordination by *and* from a period that begins immediately after the endpoint of Late Modern English chosen by Beal (2004), namely 1945, and thus ties together diachronic and synchronic research on register variation in coordination. In addition, we examine speech-purposed language, i.e., scripted language intended to be spoken, from films and television, to see whether it has been affected by the colloquialization of the norms for some written registers identified in studies of late-20th-century English such as Hundt and Mair (1999) and Leech et al. (2009); results for these registers are contrasted with the comparatively informal but largely non-speech-related English found in magazines.

2 Previous work on uses of *and*

Previous corpus-based studies have shown that there is considerable cross-register variation as regards what linguistic units tend to be coordinated by *and*. Coordination of clausal units, e.g., main or subordinate clauses, tends to be a characteristic of spoken or speech-related discourse, while in expository writing and other literary registers, *and* is typically used to conjoin phrasal units such as noun phrases; research also indicates that these tendencies have become more pronounced over time.

In his visionary study of uses of *and*, Chafe (1982) investigated the use of conjunctions in two “maximally differentiated” registers, dinner-table conversations and academic writing (Chafe 1982: 36), showing that *and* was not only used more than the other conjunctions (e.g., *but*, *so*, and *because*) included in the study but that it also occurred four times as often in spoken conversation as in

academic writing. Results corroborating the main trend of clausal *and* being a characteristic of spoken registers and phrasal *and* typical of writing were obtained in a follow-up study by Chafe and Danielewicz (1987) based on conversations, lectures, letters, and academic papers.

The datasets investigated in these studies were relatively small compared with the large-scale study of the occurrence and distribution of *and* in various registers presented by Biber et al. in their *Longman Grammar of Spoken and Written English* (1999). The some 40-million-word corpus used for this investigation contained texts representative of spoken conversation, fiction, news, and academic prose. These registers displayed notable variation in the occurrence of *and*: *and* was used some 20 times per 1,000 words in spoken conversation, 20 times in news, 27 times in academic prose, and 28 times in fiction (Biber et al. 1999: 81 and p.c. by Geoffrey Leech; see Culpeper and Kytö 2010: 160). Contrary to the results obtained by Chafe (1982) and by Chafe and Danielewicz (1987), *and* was thus more frequent in academic prose and fiction than in spoken conversation; at the same time, in accordance with these previous studies, clause-level *and* proved to dominate in spoken conversation and phrase-level *and* in academic prose (the former in nearly 80% and the latter in nearly 70% of the instances recorded for each register; for discussion, see Culpeper and Kytö 2010: 160).

Further insights into uses of *and* have been presented in a number of factor analyses. In Biber's pioneering (1988) study, clausal uses of *and* loaded as one of the oral, "involved" features on Dimension 1 ("Involved vs. Informational Production") while phrase-level uses loaded as a "literate" feature and clause-level uses as an "oral" feature on the single dimension that separated "oral" and "literate" discourse in Biber (2003).

As for the diachronic perspective, even though we have no direct evidence of early speech (prior to the invention of the phonograph in 1887 and even much later), we have indirect access to spoken interaction via texts produced in authentic speech situations such as criminal or ecclesiastical trials (speech-based language) or via texts containing dialog constructed to mimic speech in, e.g., plays (speech-purposed language). Such documents naturally have to be approached with some caveats as regards, for instance, scribal interference and authorial choice; for instance, no one-to-one correspondence between an actual speech event and its written record in witness depositions or trial proceedings should be taken for granted. In their study of the uses of *and* in Early Modern English speech-related trials and drama texts and contemporary history and scientific writing, Culpeper and Kytö (2010: Ch. 7) showed that in overall incidence figures, *and* was more common in the Early Modern English data than in the *Longman Grammar* corpus data (29 and 22 tokens per 1,000 words, respectively) and that this difference was even more pronounced for clausal *and* (18.1 and 12.1 tokens per 1,000

words, respectively). As for register variation, the results obtained for Modern English were echoed in the distribution of clausal versus phrasal *and*: clausal *and* was used more than phrasal *and* in the speech-related registers (e.g., in some 70% of the instances in the trials), while the non-speech-based texts presented a less pronounced dominance of clausal uses (e.g., in some 55% of the instances in science). A decrease in the use of clause-level *and* was also attested in the data and was tentatively ascribed to “the increasing regulation of clauses” as of the Middle English period onward, other conjunctions assuming some of the earlier functions of *and* (Blake 1996: 226, for details, see Culpeper and Kytö 2010: 167).

A later study by Kytö and Smitterberg (2019) of the uses of *and* in speech-based Late Modern English targeted the Old Bailey Corpus (1720–1913). In addition to looking into general trends of development, i.e., stability and change, the aim was to find evidence for possible influence of the gender and social class parameters on the distributions of clausal, phrasal, and other uses. The general trend in the Old Bailey data was a clear increase in the use of clausal coordination and a decrease (albeit less pronounced) in the use of phrasal coordination (and ambiguous instances). This result was interpreted to indicate a “genre drift” (cf. Biber and Finegan 1997) whereby oral features start gaining ground in formal spoken registers, here in trial proceedings where the language is produced in a relatively formal speech situation such as the late modern courtroom. The subsets of data distinguished along the sociolinguistic parameters reflected, by and large, these overall trends of change, which was taken to point to “relative sociolinguistic homogeneity” in the use of *and*. There were also weak indications to warrant the suggestion that women and witnesses from higher classes may have led the change (Kytö and Smitterberg 2019: 246–248).

Further evidence of the trajectories of development of *and* related to processes of colloquialization and densification in Late Modern English was presented by Smitterberg (2014, 2021) in his studies of newspaper language and other registers. Smitterberg (2014) demonstrated that the proportion of clausal *and* increased significantly in British newspapers between the periods 1830–1850 and 1875–1895. In Smitterberg (2021), these results were contrasted with developments from the beginning to the end of the 19th century in the registers that make up A Corpus of Nineteenth-Century English (CONCE). These results demonstrated that register differences in clausal versus phrasal uses of *and* were in general less pronounced in the 1800s compared with the results for late-20th-century English presented in Biber et al. (1999). This makes the latter half of the 1900s an interesting period to cover, as today’s pronounced register differentiation may in part have developed during these decades. Only three samples in CONCE displayed significant change: clausal coordination became more prevalent in debates from the Houses of

Parliament and in private letters written by men, while women's private letters evinced the opposite tendency. In the debates, the change coincided with a shift in the predominant mode of speech representation from indirect to direct speech; this shift may have contributed to making the accounts more faithful representations of the underlying speech events, which may have favored clausal *and*. Alternatively, the spoken debates may themselves have become more oral over time, as indicated by Geisler's (2002) factor-score analysis of CONCE. The development in men's letters tallies with other tendencies towards orality attested in Late Modern English private letters. The opposite development in women's letters was attributed to the survival of an earlier type of cohesive strategy in these texts: dashes that were frequently reinforced by *and* were often used instead of sentence boundaries to structure the text of the letters from the early 1800s. In contrast, late-19th-century letters by women were sentence-based to a greater extent. This difference in the importance of the written sentence as an organizational device inflated the proportion of clausal *and* in the 1800–1830 sample.

As regards sentence-initial *and*, this feature became increasingly available over the 1800s but occurred predominantly in speech-based or speech-purposed registers such as cross-examinations in trials and drama comedy. As the present study contrasts two speech-purposed registers (TV and film scripts) with a non-speech-like though comparatively informal register (Magazines), special attention will be paid to the incidence and functions of sentence-initial *and* when we present our results (see Section 4.2). *And* at the beginning of a sentence has frequently been proscribed from the Late Modern English period on (see, for instance, Straaijer 2018: 24) but remains frequent in oral writing; it is thus a particularly good indicator of orality in written texts. The function of tokens of *and* also gradually shifts from (syntactic) coordinator towards (pragmatic/textual) connector along the phrasal–clausal–sentence-initial cline (cf. Dorgeloh 2004: 1762), which makes separating three categories in this respect worthwhile.

3 Method

3.1 Material

Of the registers included in Smitterberg's (2021) analysis of 19th-century English, the speech-purposed register Drama featured the second highest proportions of non-phrasal coordination as well as of sentence-initial *and*. (The only texts that came across as even more oral in this regard were the speech-based cross-examinations included in the Trials register.) As these results indicated that speech-purposed language is hospitable to clausal coordination, we were interested in

mapping developments for two of the most important additions to the range of speech-purposed registers in the 20th century: the cinema and television. To that end, we have drawn for data on the TV Corpus (henceforth *TV*) and the Movie Corpus (henceforth *Movies*) compiled by Mark Davies (2019).

For this study, we wished to contrast results for speech-purposed English with those for relatively informal written English, so that differences between registers would not be due solely to formality. We chose to use the Popular Magazines register from COHA, the Corpus of Historical American English, also compiled by Mark Davies (henceforth *Magazines*). As this subcorpus contains American texts only, we restricted the retrieval of data from *Movies* and *TV* to US and Canadian films and television programs, so that regional differences would not influence the results.¹

Since the earliest texts in *TV* are from the 1950s, that decade was a suitable starting point for analysis. To reduce the risk that one outlier decade might skew results for a register, we included three different time periods in the study. We originally aimed to look at roughly 30-year intervals, similar to those between members of the Brown family of corpora, since this type of “generational” gap has been shown to be sufficient to reveal diachronic developments (e.g., Hundt and Mair 1999). This would have meant sampling texts from the 1950s, 1980s, and 2010s. However, because *and* is a high-frequency word, some searches in the 2010s material could not be carried out, as the search hit the ceiling imposed by the interface regarding how many tokens can be returned by a search. For this reason, rather than restricting the subcorpora manually, which might have decreased representativity, we substituted the 2000s for the 2010s as the final decade covered by the study.

3.2 Data

The total sets of tokens in each of the nine register/period subcorpora (see Section 3.1) were reduced to 200 randomly selected tokens per subcorpus using the randomizer in the corpus interface. A 200-token random sample was considered likely to represent the subcorpus populations well while keeping the amount of work involved in the manual analysis of tokens manageable. The analysis was thus based on 1,800 tokens altogether.

¹ In a few cases, the regional classification may not reflect the actual variety that was predominantly used on a TV program or in a film; for instance, *Da Ali G Show* is included in the US/Canada sample although that TV show was hosted by a fictional British character played by the British actor Sacha Baron Cohen. It was beyond the scope of this study to examine each magazine, TV show, and film individually to ensure that the appropriate variety was represented.

A small number of these tokens ($n = 36$, or 2% of the data) were excluded from further processing because they were not considered part of the language created by the original writers. These tokens include song lyrics and proper names that include *and*. The remaining tokens ($n = 1,764$, or 98% of the tokens retrieved) received a classification and were included in our analyses. Any cases where doubt arose were checked by both authors to ensure consistency.

Our categorization was based on previous work as well as a corpus-driven approach to the tokens analyzed; that is, the number of categories included was subject to updates during the examination of the data in order to capture the range of variation in the material well. We finally arrived at a set of five categories. Tokens were assigned to these categories using primarily syntactic criteria, though semantics and orthography were also involved in some decisions.

The most important syntactic criterion was the linguistic status of the two constituents coordinated by *and*; following Quirk et al. (1985), we refer to these constituents as *conjoins*. To begin with, the phrasal category comprises those tokens whose conjoins were either phrases (excluding verb phrases) or parts of such phrases, such as the two nouns coordinated by the boldfaced *and* in (1):

- (1) *Nothing decent grows without enough food and water, **and** you ain't got enough for yourself or the soil.* (TV [*Alfred Hitchcock Presents*], 1950s)

In contrast, the clausal category contains tokens with two conjoins that contain at least part of a verb phrase – from a verb, such as the first conjoin in (2), to a main clause, as in (3) – as an immediate constituent.²

- (2) *And then, Mr. DeMartino asked me for the answer, so I stalled **and** said, "Hmm, let me see, Roosevelt's Big Deal, Roosevelt's Big Deal ..."* (TV [*Daria*], 2000s)
- (3) *One of the officers was outside most of the time, **and** I was very frightened.* (Movies [*The Prowler*], 1950s)

The verb is the most central clause element – for instance, an imperative main clause or a subordinate clause can consist only of a verb – so for the purposes of the present study it proved a suitable cut-off point for clausal status.

The remaining categories are to some extent subsets of the phrasal and/or clausal categories that were given special status. To begin with, a small number of tokens were ambiguous between phrasal and clausal status. Such ambiguity

² A phrasal conjoin such as a noun phrase can of course in turn contain clausal material (e.g., a relative clause), but that clausal material is then not an immediate constituent of the conjoin.

typically arose when one phrasal and one clausal conjoin are coordinated, as in (4):³

- (4) *There is a big difference, Mommy, between a school-girl fantasy **and** actually having an affair with your mother's husband.* (TV [*Dynasty*], 1980s)

We also distinguish sentential coordination, where *and* occurs at the beginning of a sentence, as a separate category. This category was identified primarily by orthographic means: if *and* occurred at the start of an orthographic sentence,⁴ as in (5) and (6), it was considered sentential, regardless of (i) whether its conjoins could be clearly identified and (ii) whether, if so, these conjoins were clausal or phrasal.

- (5) *To distrust [sic] myself, I stared at the headlight. **And** as I did, I felt so terrible I tried to pop a few antidizziness pills into my mouth.* (Magazines [*GoodHouse*], 1980s)
- (6) *Uh-huh. Aluminum, I think. **And** the edges have been sharpened.* (TV [*DuckTales*], 1980s)

Sentence-initial *and* is of considerable interest from the perspective of orality, since *and* frequently links potential sentences in conversation. In contrast, sentential coordination has traditionally been proscribed in writing (Smutterberg 2021: 178). For this reason, it was given special status in our categorization of the data.

Finally, some tokens where (parts of) verb phrases comprised the conjoins were separated from the clausal category on semantic grounds. These tokens had the common denominator that the two verb units appeared to refer to one single unit of action. This “unity of action” criterion was used to identify tokens such as (7), where *chokes* and *dies* were considered to describe one and the same event. We refer to such tokens as *coreferential verbs* in Sections 4 and 5.

- (7) *The whale that swallows the dolphin chokes **and** dies.* (Movies [*His Majesty O'Keefe*], 1950s)

Coreferential verbs have been shown to be of interest in previous work; for instance, Kytö and Smutterberg (2019) found that they became increasingly rare over time in trial transcripts from the Old Bailey Corpus.

³ In (4), *having* was analyzed as a verbal gerund as it takes a direct object (cf. Lyne 2011: 53); this analysis means that *having ... husband* was, in turn, considered a non-finite clause.

⁴ Interjections as well as a small number of adverbs (primarily *yes* and *no*) were allowed to intervene between the beginning of an orthographic sentence and *and* in sentential coordination.

4 Results

4.1 Quantitative findings

Table 1 in the Appendix presents raw frequencies and proportions of the five categories by decade and register. (As Table 1 shows, ambiguous tokens represent a mere 2% of the data; they will thus not be considered further.) These raw frequencies were used as input for a comparison of normalized frequencies.

Estimated raw frequencies of the four remaining categories in each decade sample of Magazines, Movies, and TV as a whole were calculated based on the proportion of each category in the random subsamples (see Table 1) and the total frequency of *and* in each decade sample (as verified in the original searches; see Schwarz 2017 for another example of this technique, applied to the frequency of passives). For instance, a total of 136,983 tokens of *and* were retrieved from Movies in the 1980s in the original search; out of the 200 tokens randomly selected from these data, 76 were relevant instances of clausal coordination. The estimated frequency of clausal *and* in Movies from the 1980s is then:

$$(76/200) \times 136\,983 = 52\,053.54$$

(This calculation has to be done individually for each subcorpus, as the proportions of different uses of *and* are not constant across the subcorpora.)

These estimated raw frequencies were then normalized to tokens per 1,000 words in what Biber et al. (2016) refer to as a *text-linguistic* analysis, with the aid of word counts available at english-corpora.org.⁵ For instance, as the word count for Movies from the 1980s is 10,739,129, the normalized frequency of clausal *and* in this sample is:

$$(52\,053.54/10\,739\,129) \times 1000 \approx 4.85$$

These normalized frequencies are given by register in Figures 1–3 (note that the scale on the value axis is different for Magazines) and in Table 2 in the Appendix.

As Table 1 and Figures 1–3 show, the phrasal and clausal categories together account for more than three quarters of the data. In the two speech-purposed registers, the clausal category predominates, while phrasal tokens are more frequent in Magazines. These findings support previous studies, which have shown that phrasal coordination is more prevalent in less oral writing. While the proportion of phrasal coordination in Magazines (55%; see Table 1) is quite similar to what Biber et al.

⁵ As Biber et al. (2016: 368) note, in text-linguistic analyses, the unit of observation can be either each text or each subcorpus. The label *text-linguistic* thus subsumes both research designs, the latter of which was used in the present study.

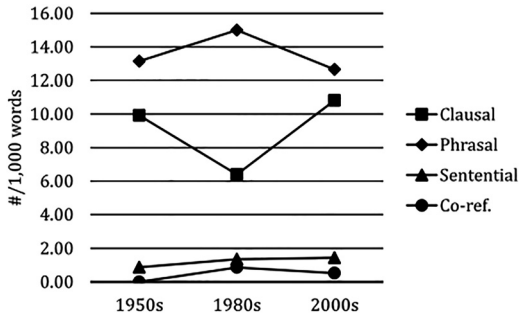


Figure 1: Coordination type by period in Magazines (frequencies per 1,000 words, ambiguous tokens excluded).

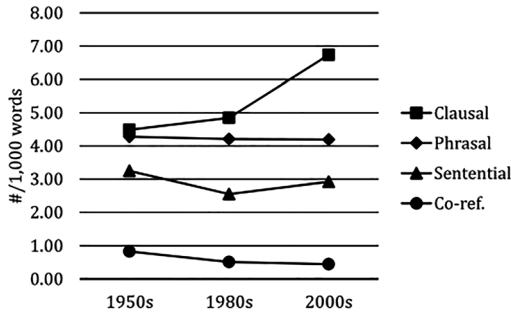


Figure 2: Coordination type by period in Movies (frequencies per 1,000 words, ambiguous tokens excluded).

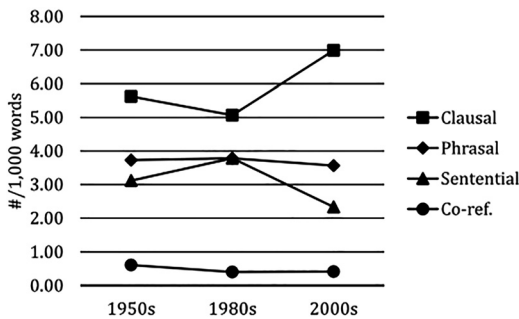


Figure 3: Coordination type by period in TV (frequencies per 1,000 words, ambiguous tokens excluded).

(1999: 81) found for present-day news texts, TV (28%) and Movies (32%) feature higher percentages of phrasal uses than conversation did in Biber et al.'s material.⁶ This is to be expected of registers with texts that were written to be spoken: they are likely to be the result of a mixture of oral and written norms.

One unexpected result for the two speech-purposed genres is the low frequency of *and* overall in the material. While the total frequency of *and* in Magazines (24 tokens per 1,000 words; see Table 2) is comparable to Biber et al.'s (1999: 81) frequencies for news texts, the frequencies in Movies and TV (14 and 13, respectively) are lower than Biber et al.'s results for conversation. Given that the proportion of phrasal uses is also higher in Movies and TV than in conversation, it is possible that the connective use of clausal and sentential *and* in spoken narrative is not always transferred to speech-purposed writing, resulting in lower total frequencies of the coordinator.

Both TV and Movies feature a substantial increase in the frequency of clausal coordination between the 1980s and the 2000s (38–39%). In TV, a possible trade-off between sentential and clausal coordination over time may explain part of this development: when the frequency of sentential coordination rises, that of clausal coordination goes down and *vice versa*. This makes intuitive sense in that the distinction between clausal and sentential coordination is mainly a written feature in the speech-purposed registers. Consider (8):

- (8) *Yes, but only to those who have shale oil leases. **And** you know who's at the top of that list.* (TV [*Dynasty*], 1980s)

As Culpeper and Kytö (2010: 168) note, the sentence is a unit of limited relevance to spoken communication: we do not actually speak in sentences. The sentence-initial *and* in (8) could thus also have been rendered as clausal (e.g., ... *oil leases; and you ...* or ... *oil leases, and you ...*). The scriptwriter's decision to insert a period before *and* may have been motivated by prosodic patterns (tone units and pauses) of the imagined speech event as well as pragmatic considerations regarding the informational independence of the unit beginning with *and*. However, even such a hypothesized trade-off between clausal and sentential coordination would not account for all of the increase in clausal coordination in TV between the 1980s and 2000s, and in Movies there is no such relationship between the two categories that would explain the rise in frequency. It seems likely that the increase is at least partly an indication of a colloquialization of the norms for these speech-purposed registers.

⁶ Note that, for percentages, the total comprises only the relevant tokens of *and* in each subsample, while normalized frequencies were based on all 200 randomly selected tokens per subsample.

Sentence-initial *and* displays an expected pattern: it is more frequent in the speech-purposed registers (2.9 tokens per 1,000 words in Movies and 2.6 in TV, compared with 1.2 in Magazines; see Table 2) despite their lower total frequency of *and*. In terms of proportions, sentential coordination accounts for 22–23% of all relevant tokens of *and* in Movies and TV but for a mere 5% in Magazines (see Table 1). These differences between Magazines and the other registers hint at the possibility that sentence-initial *and* may have different functions in the three registers included. We will return to this issue in Section 4.2.

The most surprising finding for Magazines is arguably the lack of any clear diachronic pattern. Instead, regardless of whether the analysis is based on normalized frequencies (Figure 1; Table 2) or proportions (Table 1) of the categories of coordination, the 1980s stand out as markedly less speech-like, with higher levels of phrasal coordination than the 1950s and 2000s. However, the composition of the decade subcorpora may underlie this seeming lack of orality in the 1980s sample. When the actual sources of the Magazines tokens were examined, it became clear that one publication in particular predominates in some samples, namely *TIME Magazine*. Moreover, this magazine is particularly dominant in the 1950s and 1980s, when it accounts for roughly half of the randomly selected tokens (98 in the 1950s and 101 in the 1980s); the 2000s evince a more diverse pattern, with 51 tokens coming from *TIME Magazine*. Although the language of *TIME Magazine* has been described as colloquial (Schwarz 2017: 309), it comes across as literate with regard to coordination: 55 of the 98 tokens from the 1950s (56%) and no fewer than 71 of the 101 tokens from the 1980s (70%) are phrasal. The 2000s sample, again, is slightly less extreme, with 26 out of 51 (51%) tokens being phrasal. The unexpected dominance of phrasal coordination in 1980s Magazines is thus due largely to two factors: a bias towards *TIME Magazine* in the selection of 1980s data and a preference for phrasal coordination in 1980s texts from *TIME Magazine*. Although a random selection of data need not reflect the makeup of the underlying subcorpus, the dominance of *TIME Magazine* in the quantitative results means that caution needs to be exercised before it is assumed that results for Magazines are representative of the register as a whole.

The final category to be accounted for, co-referential verbs, is clearly a rare feature. It accounts for 3.6% of the data in TV and for 4.5% in Movies, but for a mere 1.9% in Magazines (see Table 1); however, the frequency of the feature is more similar across the three genres (0.42–0.52 tokens per 1,000 words; see Table 2). Raw frequencies are too low to allow safe conclusions regarding trends in the distribution of this category.

Overall, the analysis of normalized frequencies seems to indicate that register is a more important conditioning factor than time as regards the use of *and*. The main exception is the clear increase in the frequency of clausal coordination between the 1980s and the 2000s in Movies, which is not matched by a decrease in the frequency of the other categories; instead, the overall frequency of *and* rises in this register.

4.2 Focus on sentence-initial *and*

As shown in Section 4.1, over 20% of the relevant tokens of *and* in TV and Movies are sentence-initial, figures that are similar to what was attested for present-day conversation in Biber et al. (1999: 84). In contrast, fewer than 5% of the Magazines tokens occur at the start of a sentence, putting this register between (i) fiction and (ii) news and academic writing in Biber et al.'s (1999) analysis. Against this background, a closer look at sentence-initial *and* in the data is clearly warranted. In this partly qualitative examination, we focus on the Magazines and TV samples, as these registers feature the lowest and the highest proportions, respectively, of sentence-initial *and* (see Table 1).

The functional load of sentential *and* arguably differs along the register parameter. To begin with, when conjoiners are on the sentence level or above, the coordinator shades into a more general pragmatic and/or textual connector (e.g. Dorgeloh 2004: 1762); for instance, in narrative texts, sentence-initial *and* has been argued to increase coherence (Dorgeloh 2004: 1769–1770).⁷ As this expanded functional range is not formally marked (i.e., *and* has the same form regardless of its function), sentence-initial *and* has been claimed to exhibit *pragmatic ambiguity* in this respect (Sweetser 1990: 86–93). In addition, the prevalence of connector versus coordinator functions is arguably influenced by parameters like medium. In spoken registers, there are constraints on how large the passages that are connected can be: the ephemeral nature of the message conveyed and the capacity of hearers' short-term memory limit how much previous discourse they can remember and connect to an instance of *and*. Speech-purposed language is presumably produced with such constraints in mind. By contrast, in writing, it is typically possible to go back to and reinterpret a previous passage in the light of a sentential token. Proscription of sentence-initial *and*

⁷ This diversity of functions notwithstanding, we will continue to refer to the connected or coordinated units/passages as *conjoins* throughout this study, in the interest of simplicity.

(see Section 3.2) is also more likely to affect writing than speech, given the comparatively planned nature of written discourse.

The examination of sentence-initial *and* in Magazines and TV showed that three main semantic–pragmatic relationships between conjoins stand out in both corpora. First, as mentioned above, *and* can be used to further narration; the two conjoins then typically refer to events in chronological succession, and *and* helps to make the temporal order clear, as in (9):

- (9) *One time, some defensive lineman said to him, “Keep your hands off my face.” **And** Whitticker said to him, “You keep your face off my hands.”*
(COHA: Magazines [*Sporting News*], 2000s)

This function of sentence-initial *and* can be reinforced by the addition of *then*. Two of the four examples of narrative sentential *and* in Magazines as well as one of the 13 tokens in TV occur immediately after quoted speech, as in (9). It is possible that switches between quoted speech and spoken or written narrative make a link between these stretches of text especially useful.

More frequently, though, sentence-initial *and* has a less temporal function: the sentence that follows it presents additional information or elaborates on information that has just been given (22 tokens in Magazines, 107 in TV). Taken together, these functions account for 76% of tokens in Magazines and 79% in TV. As these two uses of *and* shade into each other without clear boundaries, presenting quantitative results on their relative proportions would involve too much subjectivity for those findings to be reproducible. However, a look at the relevant tokens indicates that there is more frequently an overt connection between conjoins, such that the second conjoin both elaborates on the first conjoin and includes some sort of reference to all or part of it, as in (10), in Magazines. In many of these tokens, *and* seems similar in function to an additive linking adverbial (Biber et al. 1999: 875–876) such as *in addition* or *furthermore*.

- (10) *There are other types of magical illnesses than soul loss such things as bewitchment, evil eye, corpse sickness. **And** every one of these ailments has several forms, each of which has its “signs” in the pulses.* (COHA: Magazines [*Atlantic Monthly*], 1950s)

In two cases, a resultative elaboration is reinforced by the combination of *and* and the conjunct *so*.

The TV tokens, in contrast, more often feature additional, less clearly linked information on the same overall topic in the second conjoin, as in (11):

- (11) *And i [sic] want all the copies of the prints for package “H”. **And** you keep this information completely to yourself, lieutenant, you understand me?* (TV [*The Closer*], 2000s)

One subtype of this function, illustrated in (12), occurs when there is a shift between speakers, and the first conjoin is really the same speaker’s previous utterance.

- (12) *Until about what time did you detain them there? Uh, it was about 20 minutes past 6. **And** would you tell the court please what transpired after they left police headquarters?* (TV [*Perry Mason*], 1950s)

This subtype is particularly common in courtroom settings, where conjoins are typically questions asked in cross-examinations (10 of the 21 tokens come from the show *Perry Mason*). As shown by Smitterberg (2021: 182; see also Culpeper and Kytö 2010: 171), these uses of *and*, through which a lawyer can control the course of a cross-examination as well as imply links between witnesses’ answers, are also a feature of real-life historical courtroom interaction. In contrast, in Kytö and Smitterberg’s (2019) study, which included only witness statements from the Old Bailey Corpus, only one sentence-initial token was attested.

One further function of sentence-initial *and* that is rarer but present in both corpora is the introduction of a new (sub)topic. Again, distinguishing this function from elaborating or providing more information on an established topic is necessarily a subjective endeavor, but the decision is easier when *and* co-occurs with other linguistic signals of a topic shift, e.g., *and you know what* or *and now for* in (13):

- (13) *Theirs was truly a wedding of December and December. It ended in a draw, with no one the winner except the insurance companies. **And** now for the epilogue of tonight’s story, after which I’ll scamper back.* (TV [*Alfred Hitchcock Presents*], 1950s)

In sum, although sentence-initial *and* is more frequent in speech-purposed writing than in Magazines, most of its functions occur in both types of register, though overt connections between conjoins are more common in Magazines. This relative similarity in distribution, together with the frequency difference, indicates that we are looking at spoken usage influencing comparatively informal writing.

5 Discussion

Our study of *and* has shown that there is a great deal of continuity across time as regards its distribution: clausal and phrasal uses predominate in oral and literate registers, respectively. For our 20th-century and early-21st-century data, register is also a more important parameter than time, i.e., register *variation* is more apparent in the material than language *change*. However, in the speech-purposed registers, clausal *and* increased noticeably in frequency between the 1980s and the 2000s, which indicates possible colloquialization in these scripts: the language may have become more similar to informal conversation around the turn of the millennium. This tendency is especially noticeable in Movies, where there is no corresponding decrease in the frequency of the other categories included. Like all living languages, English continues to vary and change, with register being one of the most important determinants of the nature of variation and the direction of change.

The unexpected result that the 1980s samples from TV and Magazines were less oral with regard to the frequency of clausal *and* than the 1950s and 2000s samples could be explained with reference to the relative similarity of clausal and sentential coordination in speech-purposed registers and to the make-up of the subcorpora from Magazines. In TV, conflating the clausal and sentential categories largely removes the outlier status of the 1980s subcorpora. For Magazines, the dominance of *TIME Magazine* in the 1980s data and the pronounced preference for phrasal coordination in this publication during that decade together account for much of the distribution. This finding also underscores the importance of considering the structure of one's dataset even when it derives from very large corpora such as COHA.

Sentential coordination is of particular interest in a study of orality in writing, as this feature has frequently been proscribed, yet remains frequent in speech. Our results show that sentence-initial *and* is still firmly associated with speech-related language, where the influence of spoken norms for textual organization is strong, but also that it occurs in all three registers examined. Similarly, the various functions of this pragmatically ambiguous linker were represented in both TV and Magazines, with some tentative register differences emerging from a close reading of individual tokens. In particular, there were more often overt connections such as anaphora between parts of conjoins in Magazines, creating an effect of the second conjoin elaborating on the

information contained in the first one. This may be due to the ease with which the reader can go back and forth between even quite long conjoints in written text in order to interpret such connections.

One area of investigation that stands out as particularly relevant for further research on the development of the functions of *and* is the early 20th century. This period remains understudied despite valuable initiatives such as the extension of the Brown family of corpora backwards in time and their inclusion in factor score analyses like Biber and Finegan (1997). Registers that were established in these decades, such as movies with sound (originally referred to as “talkies”), can be studied from their emergence on, to see what types of variation existed as register-specific norms were being established. But several existing registers that may have been important for the establishment of the present-day use of *and* also underwent important developments during these decades. We know too little about the linguistic characteristics of, for instance, private letters that were written in the early 1900s, shortly after the achievement of near-universal literacy among first-language speakers of English, or of registers that addressed the public with a new level of respect – for political reasons (e.g., the language of elections after the advent of universal suffrage) or owing to financial pressure (e.g., popular newspapers such as the *Daily Mail* after its foundation in 1896). To what extent did orality characterize communication in these registers, and how did the patterns found in them influence the English of the 20th century and beyond? The study of levels of coordination in such texts can add to our knowledge of the diachronic interplay between spoken and written norms for production.

Another area where more research is needed concerns collocations and phraseology. *And* forms part of a number of more or less set phrases, including general extenders like *and stuff* (see Overstreet 1999) as well as binomial phrases such as *ladies and gentlemen* (see Biber et al. 1999: 1030–1036). Some of these set phrases are so frequent that they may be stored as formulaic units in speakers’ mental lexicons (indeed, our category of co-referential verbs partly overlaps with that of verb-*and*-verb binomials). To the extent that these set patterns are treated as single units by speakers, the status of *and* as a coordinator in such tokens is doubtful.

Finally, like any case of linguistic variation, the proportions of clausal, phrasal, etc. uses of *and* are affected by a large number of parameters. These include register and time, which have been in focus in this study, but also features that pertain to the writer/speaker. The simultaneous, independent influence of these parameters on the incidence of different types of coordination should ideally

be studied within the framework of a multifactorial analysis. We will return to this issue in our future work.

Our findings also have a couple of wider applications for corpus-linguistic studies in synchrony as well as diachrony. First, recent years have witnessed rapid increases in the scale of corpus compilation initiatives as well as the sophistication of statistical analyses of corpus data. However, these increases notwithstanding, qualitative, sometimes subjective analysis remains necessary as a complement to quantitative accounts in many cases. The identification of co-referential verbs and of different functions of sentence-initial *and* requires an amount of time devoted to each token that would be incompatible with too large a dataset. Moreover, even very large corpora can retain some bias regarding what sources were sampled, as became clear when the 1980s Magazines sub-corpus was examined in more detail. While very large corpora and statistical rigor have been indispensable in making corpus linguistics one of the most prominent methodologies in the study of language, linguistics by and large remains an interpretive field of scholarship founded on the individual scholar's or scholars' experience and expertise. As Egbert et al. (2020: 71) note, "the 'data' from quantitative corpus analysis requires linguistic interpretation at every stage in order to qualify as meaningful 'information' about language structure and use. In other words, statistical analysis can provide us with data, but that data must be interpreted if it is to be useful for linguistic description".

Secondly, as Biber (1988: 61–63) established almost 35 years ago, macro-level perspectives on register variation such as factor analysis and micro-level studies of individual features are in a mutually informative – ideally, symbiotic – relationship to one another. Micro-level studies inform macro-level analyses mainly in terms of what linguistic features are relevant to include in the latter. In turn, the positive and negative correlations established between the features included in those macro-level analyses can identify new research questions for micro-level studies – and the results of those studies can then help to inform macro-level perspectives, and so on. In the case of the present study, the main takeaway may be the need to consider clausal and sentential coordination separately, at least when variation in non-speech-related writing is examined, and to be aware of the multi-functionality of sentence-initial *and* when interpreting the outcome of the analysis. It is hoped that the next 35 years will continue to bring new insights into linguistic register variation based on this interplay between the two levels of analysis.

Appendix

Table 1: Coordination type by period and register (raw frequencies and row percentages).

Magazines	Phrasal		Clausal		Sentential		Co-ref.		Ambiguous		Total
	#	%	#	%	#	%	#	%	#	%	
1950s	106	53.0	80	40.0	7	3.5	0	0.0	7	3.5	200
1980s	122	62.6	52	26.7	11	5.6	7	3.6	3	1.5	195
2000s	97	49.0	83	41.9	11	5.6	4	2.0	3	1.5	198
Total	325	54.8	215	36.3	29	4.9	11	1.9	13	2.2	593
Movies	Phrasal		Clausal		Sentential		Co-ref.		Ambiguous		Total
#	%	#	%	#	%	#	%	#	%		
1950s	62	32.8	65	34.4	47	24.9	12	6.3	3	1.6	189
1980s	66	33.3	76	38.4	40	20.2	8	4.0	8	4.0	198
2000s	56	29.0	90	46.6	39	20.2	6	3.1	2	1.0	193
Total	184	31.7	231	39.8	126	21.7	26	4.5	13	2.2	580
TV	Phrasal		Clausal		Sentential		Co-ref.		Ambiguous		Total
#	%	#	%	#	%	#	%	#	%		
1950s	55	27.8	83	41.9	46	23.2	9	4.5	5	2.5	198
1980s	56	28.4	75	38.1	56	28.4	6	3.0	4	2.0	197
2000s	52	26.5	102	52.0	34	17.3	6	3.1	2	1.0	196
Total	163	27.6	260	44.0	136	23.0	21	3.6	11	1.9	591

Table 1: (continued)

All registers	Phrasal		Clausal		Sentential		Co-ref.		Ambiguous		Total
	#	%	#	%	#	%	#	%	#	%	
1950s	223	38.0	228	38.8	100	17.0	21	3.6	15	2.6	587
1980s	244	41.4	203	34.4	107	18.1	21	3.6	15	2.5	590
2000s	205	34.9	275	46.8	84	14.3	16	2.7	7	1.2	587
Total	672	38.1	706	40.0	291	16.5	58	3.3	37	2.1	1764

Table 2: Coordination type by period and register (frequencies per 1,000 words, ambiguous tokens excluded).

Magazines	Phrasal	Clausal	Sentential	Co-ref.
1950s	13.14	9.92	0.87	0.00
1980s	15.00	6.40	1.35	0.86
2000s	12.65	10.82	1.43	0.52
Total	13.51	9.21	1.24	0.47
Movies	Phrasal	Clausal	Sentential	Co-ref.
1950s	4.28	4.49	3.25	0.83
1980s	4.21	4.85	2.55	0.51
2000s	4.19	6.74	2.92	0.45
Total	4.21	6.05	2.90	0.52
TV	Phrasal	Clausal	Sentential	Co-ref.
1950s	3.73	5.63	3.12	0.61
1980s	3.78	5.06	3.78	0.41
2000s	3.56	6.99	2.33	0.41
Total	3.60	6.69	2.55	0.42

References

- Beal, Joan C. 2004. *English in modern times: 1700–1945*. London: Arnold.
- Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, Douglas. 2003. Variation among university spoken and written registers: A new multi-dimensional analysis. In Pepi Leistyna & Charles Meyer (eds.), *Corpus analysis: Language structure and language use* (Language and Computers 46), 47–70. Amsterdam: Rodopi.
- Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style*. Cambridge: Cambridge University Press.
- Biber, Douglas, Jesse Egbert, Bethany Gray, Rahel Oppliger & Benedikt Szmrecsanyi. 2016. Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In Merja Kytö & Päivi Pahta (eds.), *The Cambridge handbook of English historical linguistics*, 351–375. Cambridge: Cambridge University Press.
- Biber, Douglas & Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In Terttu Nevalainen & Leena Kahlas-Tarkka (eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen* (Mémoires de la Société Néophilologique de Helsinki 52), 253–275. Helsinki: Société Néophilologique.
- Biber, Douglas & Bethany Gray. 2012. The competing demands of popularization vs. economy: Written language in the age of mass literacy. In Terttu Nevalainen & Elizabeth Closs Traugott

- (eds.), *The Oxford handbook of the history of English*, 314–328. Oxford & New York: Oxford University Press.
- Biber, Douglas & Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge: Cambridge University Press.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson.
- Blake, Norman. 1996. *A history of the English language*. London: Macmillan.
- Chafe, Wallace L. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen (ed.), *Spoken and written language: Exploring orality and literacy*, 35–53. Norwood, NJ: Ablex.
- Chafe, Wallace & Jane Danielewicz. 1987. Properties of spoken and written language. In Rosalind Horowitz & S. Jay Samuels (eds.), *Comprehending oral and written language*, 83–113. San Diego: Academic Press.
- COHA = The Corpus of Historical American English. 2010. Compiled by Mark Davies. Available at: <https://www.english-corpora.org/coha/>.
- Culpeper, Jonathan & Merja Kytö. 2010. *Early Modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.
- Dorgeloh, Heidrun. 2004. Conjunction in sentence and discourse: Sentence-initial *and* and discourse structure. *Journal of Pragmatics* 36(10). 1761–1779.
- Egbert, Jesse, Tove Larsson & Douglas Biber. 2020. *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge: Cambridge University Press.
- Geisler, Christer. 2002. Investigating register variation in nineteenth-century English: A multi-dimensional comparison. In Randi Reppen, Susan M. Fitzmaurice & Douglas Biber (eds.), *Using corpora to explore linguistic variation* (Studies in Corpus Linguistics 9), 249–271. Amsterdam & Philadelphia: John Benjamins.
- Hundt, Marianne & Christian Mair. 1999. “Agile” and “uptight” genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4(2). 221–242.
- Kytö, Merja & Erik Smitterberg. 2019. The conjunction *and* in phrasal and clausal structures in the *Old Bailey Corpus*. In Nuria Yáñez-Bouza, Emma Moore, Linda van Bergen & Willem B. Hollmann (eds.), *Categories, constructions, and change in English syntax*, 234–250. Cambridge: Cambridge University Press.
- Leech, Geoffrey, Marianne Hundt, Christian Mair & Nicholas Smith. 2009. *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.
- Lyne, Susanna. 2011. *The subject of the verbal gerund: A study of variation in English*. Uppsala: Uppsala University PhD dissertation.
- Movies = The Movie Corpus. 2019. Compiled by Mark Davies. Available at: <https://www.english-corpora.org/movies/>.
- Overstreet, Maryann. 1999. *Whales, candlelight, and stuff like that: General extenders in English discourse*. Oxford: Oxford University Press.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London & New York: Longman.
- Schwarz, Sarah. 2017. “Like getting nibbled to death by a duck”: Grammaticalization of the GET-passive in the TIME Magazine Corpus. *English World-Wide* 38(3). 305–335.

- Smutterberg, Erik. 2014. Syntactic stability and change in nineteenth-century newspaper language. In Marianne Hundt (ed.), *Late Modern English syntax*, 311–330. Cambridge: Cambridge University Press.
- Smutterberg, Erik. 2021. *Syntactic change in Late Modern English: Studies on colloquialization and densification*. Cambridge: Cambridge University Press.
- Straaijer, Robin. 2018. The usage guide: Evolution of a genre. In Ingrid Tieken-Boon van Ostade (ed.), *English usage guides: History, advice, attitudes*, 11–30. Oxford: Oxford University Press.
- Sweetser, Eve. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.
- TV = The TV Corpus. 2019. Compiled by Mark Davies. Available at: <https://www.english-corpora.org/tv/>.