

On the role of cumulative knowledge building and specific hypotheses: the case of grammatical complexity

Tove Larsson,¹ Douglas Biber¹ and Gregory R. Hancock²

Abstract

As corpus linguistics matures as a field, there is an increasing number of research areas in which we have accrued sufficient knowledge such that we can start to build knowledge in a cumulative manner by (a) synthesising findings and generalisations made by previous research and interpreting new findings in relations to those, and (b) formulating and testing increasingly specific predictions/hypotheses resulting from (a). This paper outlines what a move towards cumulative knowledge building may look like for the field and offers a case study on grammatical complexity as illustration. In building knowledge in a more systematic way, we can engage more deeply with the claimed generalisable findings from previous research and help move the field's state-of-the-art forward.

Keywords

confirmatory study designs, confirmatory techniques, cumulative knowledge building, grammatical complexity, hypothesis testing

I. Introduction

As corpus linguists, we tend to take pride in the fact that we are engaged in a scientific enterprise, often contrasted with the introspective methods of more traditional linguistic research (see, for example, the discussion in McEnery and Brezina [2022: Chapter 1]). Like science, corpus linguistics focusses on empirical investigations of naturally occurring data, usually collected using sophisticated computational tools and often analysed using statistical

¹ Northern Arizona University, Department of English, PO Box 6032, Flagstaff, AZ 86011–6032, USA.

² Quantitative Methodology: Measurement & Statistics (QMMS), Department of Human Development and Quantitative Methodology, 1230D Benjamin Building, 3942 Campus Drive, University of Maryland, College Park, MD 20742–1115, USA.

Correspondence to: Tove Larsson, *e-mail:* Tove.Larsson@nau.edu

techniques. However, we argue in this paper that corpus linguistic research often differs from other scientific research in one key respect: by its failure to adopt a cumulative perspective on research findings.

A cumulative approach to knowledge building is considered by some to be the hallmark of science:

Science is generally cumulative—meaning that as more data is collected and more discoveries are made, science builds toward a more complete and accurate understanding of the physical universe—the goal of science in general [...].

(Zeigler, 2012: 585)

However, the sciences did not begin by taking this approach. Early on, scientists started out merely reporting facts without any apparent intention to build cumulatively on previous findings or form hypotheses based on them. The development of sciences in the western tradition can be traced in documentation from the Royal Society of London. It is reported that in the early days, due perhaps to the vast numbers of new sightings to report and describe, many researchers focussed on reporting anything and everything they observed:

Until 1800, [...] [t]he most articles and pages were devoted to **observations and reports of natural events**, ranging from remarkable fetuses and earthquakes, though astronomical sightings, anatomical dissections, and microscopical observations.

(Bazerman, 1988: 65, emphasis added)

As noted by Gross *et al.* (2002: 21), this approach to science was coupled with the absence of hypotheses and theories, which presents challenges for distinguishing the signal from the noise: ‘In the absence of a hypothesis [...], there is no pressure on [a researcher] to exclude from his article any observation’. Even some scholars working in the seventeenth century noted and commented on the contrast between these two approaches, for example:

Leibniz criticized 17th century English science for its emphasis on **the bookkeeping of nature over the synthesis of factual information into a unified theory**: [...] ‘I should be astonished if Mr. [Robert] Boyle, has so many fine experiments, would not come to some theory of chemistry after meditating so long on them. Yet in his books, and for all the consequences that he draws from his observations, he concludes only what we all know, namely, that everything happens mechanically.

(Quoted in Wiener [1951: xxv], cited in Gross *et al.* [2002: 4], emphasis added)

In sum, as reported by Bazerman (1988: Chapter 3) in his study of articles from the influential journal *Philosophic Transactions* of the Royal Society of London between 1665 and 1800, the sciences have followed

a historical trajectory of research moving from observation of any-and-all natural phenomena, to noting generalisable patterns found in previous observations and investigating whether those patterns are maintained in new contexts, to actually testing specific predictions/hypotheses based on previous research.

Given the relatively short history of corpus linguistics and the wealth of data to which we tend to have access, it is perhaps not surprising that the main focus so far along our trajectory has been on exploration and description of (quantitative and qualitative) observations and findings. Exploratory work is foundational, and is, as such, a prerequisite for empirically based hypotheses. However, for research areas where we have accrued some knowledge, we would arguably want to make better use of previous findings to formulate and subsequently test increasingly specific hypotheses. Concretely, building on the trajectory of the sciences in their early days, we propose the following three stages³ along which we would want corpus linguistics to develop as it matures:

- (1) Report observations and findings;
- (2) Synthesise the findings from previous studies to identify apparent patterns, and interpret the findings of new studies relative to those previously found patterns; and,
- (3) Formulate and test increasingly specific predictions/hypotheses/theories, based on the syntheses of research findings from Stage 2.⁴

In this paper, we expand on this view. We elaborate on Stage 2 in Section 2 and on Stage 3 in Section 3. We also include a case study where we illustrate what a cumulative approach to knowledge building could look like in Section 4.

2. Cumulative knowledge building

For any given topic in any given field, we have to start somewhere. That is, if we, as a field, do not know anything about a topic, the natural first step is to start by collecting observations (i.e., to start with an exploratory design), in an attempt to be able to map out the territory (i.e., Stage 1). However, authors of subsequent studies on the same topic have a choice to make: either (a) they continue with the narrative of ‘no one has looked at [this unexplored corner of the topic]’, or (b) they can explicitly build on the findings of previous work (i.e., Stage 2). A graphical representation of these two types of study designs can be found in Figure 1.

³ It should be noted, however, that the stages are not necessarily sequential, nor are they fully discrete. For topics that have not yet received sufficient focus for a body of literature to have emerged, there is always going to be room for Stage 1.

⁴ We would regard theory building as a further extension of Stage 3.

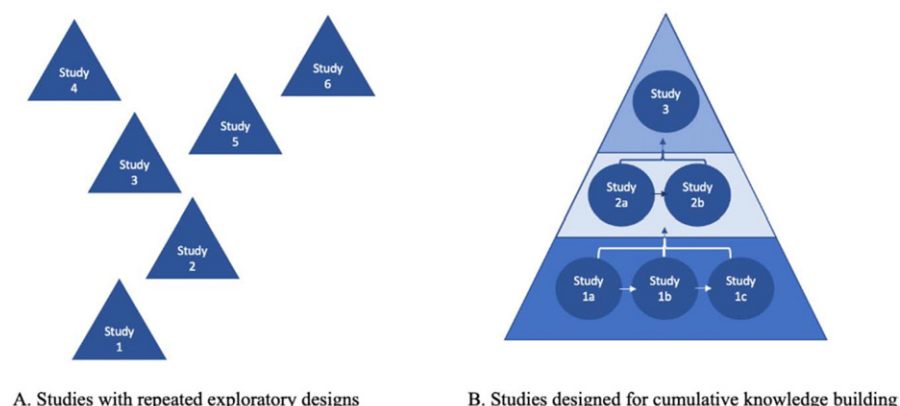


Figure 1: Repeated exploratory *versus* cumulative research designs.

There are two essential steps in Stage 2: (i) synthesising the specific findings and supposed generalisations made by previous research to identify apparent patterns, and then (ii) interpreting new findings in relation to those specific findings and generalisations.

It might seem obvious that the literature review and discussion sections of a published paper should set up a study to contribute cumulatively to knowledge building. Surprisingly, though, this does not seem to be the standard procedure in corpus linguistic studies. Instead, what we sometimes find is a simple catalogue of previous studies, listing what they looked at (and did not look at).

To illustrate this point, we present below two constructed – and much simplified – examples of literature reviews.⁵ Example A does a thorough job of identifying previous research, but it merely lists the topics of previous publications and does not synthesise (or even describe) the findings from those studies. Example B, by contrast, synthesises the findings of previous studies, and sets out to identify gaps in our existing knowledge that serve to inform the state-of-the art. An important distinction for our argument here is the notion of a ‘gap’ as any topic that has not been studied before (typical of repeated exploratory studies), *versus* a ‘gap’ as something that we do not know relative to a body of previous research findings (typical of studies designed for cumulative knowledge building).

A.

The question of which features are common in English as a Foreign Language writing has received some attention in recent years. Smith (2002) looked at the use of personal pronouns in L1-Swedish, L1-German, and L1-English groups. Jones (2007) looked at the use of emphatics in

⁵ Our intention is not to pick on any individual study, so we have chosen not to use an authentic example here, but instead draw from several sources to illustrate the point.

L1-Spanish groups compared to L1-French groups, and Adams (2019) compared the use of emphatics in L1 Norwegian and L1 English use. However, no study to date has compared the use of stance markers in L1-Italian, L1-Finnish and L1-Portuguese groups; therefore, this study sets out to do so.

B.

The question of which features of learner writing are shared among multiple English as a Foreign Language (EFL) populations has received some attention in recent years. These features are of interest, as they could be candidates for features of interlanguage that provide insight into the development of a second/foreign language. The combined findings from Smith (2002), Jones (2007) and Adams (2019) suggest that both personal pronouns and emphatics could meet the criteria, in that the features were infrequent in L1 English writing and very frequent across multiple EFL populations from different language families: specifically, both these features were more frequently used by L1 Spanish (Smith, 2002), L1 Arabic (Jones, 2007) and L1 Swedish (Adams, 2019) writers than by L1 English writers.

However, a cross-cutting line of research shows that personal pronouns and emphatics tend to be much more frequent in personal written registers than in informational written registers. For example, Jonson (2020) showed that these features were much more frequent in personal argumentative writing than in research reports, regardless of the L1 background. Lee (2022) similarly noted lower frequencies of personal pronouns and emphatics in journal articles than in argumentative writing.

Based on these results, we may expect that register may, at least in part, have affected the outcomes of the studies by Smith (2002), Jones (2007) and Adams (2019). The present study starts out from this prediction to further study the effect of register on personal pronouns and emphatics across multiple EFL population and, thus, to see to what extent they can be considered general developmental features.

In addition to the way in which we frame the introduction and literature review, a cumulative-knowledge-building approach will also result in a different framing of the discussion/conclusion section of a paper. That is, once the study has been carried out, the findings will subsequently be interpreted relative to the specific claimed patterns from the body of previous research findings, thus facilitating future studies wishing to build on the findings and refine the methods.⁶ Then, after we know enough about a topic

⁶ Whilst the focus of this paper is not replication in itself (at least not in its traditional application [see Sönning and Werner (2021)]), there are principles of replicability that apply in this context as well. Minimally, for both replication and cumulative knowledge building, it would be preferable for authors to use transparent and comprehensive reporting practices (Larson-Hall and Plonsky, 2015) and to make coding schemes/code, data, and materials available for follow-up studies to use whenever possible (Paquot and Callies, 2020). However, it should be noted that it is possible to replicate a study without building on its findings in a cumulative manner. For example, a common design for replication studies is to test whether the methods used in the original study yield the same (or at least comparable) results for a given population in a follow-up study, whereas a study that is designed for cumulative knowledge building would focus on the findings of both studies.

to start positing and testing increasingly specific hypotheses, we can move on to Stage 3. However, to be able to test such hypotheses, we need to use techniques that enable us to do so.

3. Confirmatory study designs: moving towards more specific hypotheses

Ideally, studies carried out in a cumulative framework would posit increasingly specific, empirically driven hypotheses based on the findings in previous research that could be tested in Stage 3 to enable us to make adjustments to our current, collective knowledge of a topic. In non-observational sciences, a common method for testing such hypotheses would be to design an experiment where certain variables of interest are studied whilst others are controlled for. Since corpus linguistics looks at naturally occurring observational data, we need alternative ways of testing hypotheses to see if they can be retained or if they should be rejected. We here focus on how we can go about this using statistical techniques.

For topics where we have accrued a great deal of knowledge as a field, our hypotheses are likely to be ones of direction (more/less) and, eventually, ones of extent, rather than merely ones pertaining to existence (yes/no). As explained below, a traditional two-tailed test of statistical significance can only take us so far in this respect, so we will here propose refinements to that approach along with additional techniques that enable us to test highly specific hypotheses.

In the field, we most often use techniques from the null hypothesis significance testing (NHST) framework (see, for example, Larsson *et al.* [2022]). These are referred to as ‘inferential statistical techniques’ and can thus be contrasted with ‘exploratory statistical techniques’ such as cluster analysis and exploratory factor analysis. Nonetheless, although inferential techniques that test a null hypothesis may, at least to some degree, enable us to build on previous research, they fall short once our hypotheses start becoming specific. We elaborate on this in the following section by contrasting the null (specifically the ‘nil’)⁷ hypothesis and more specific hypotheses.

If we are interested in testing the existence of differences and relations, the null hypothesis and the accompanying alternative hypothesis in their common application are well suited. However, as outlined below, this approach is arguably far too broad to be of much use in a confirmatory framework – one where we wish to see if we should retain or reject a specific

⁷ Although not common in the field, we can also conduct tests targeting more specific, non-nil values – that is, instead of hypothesising that a difference or relation is zero (the prototypical application of a null hypothesis), we can have a null hypothesis of a specific difference (e.g., a mean difference of 7).

hypothesis resulting from findings of previous studies about direction and/or extent. That is, whilst we may be under the impression that we are in fact testing a hypothesis (it is called a null ‘hypothesis’, after all), that is not how we tend to use it. We will unpack this and delve deeper into what the null hypothesis and its opposite – the alternative hypothesis – actually help us do.

In its prototypical (and slightly simplified) form, we use the null hypothesis to test whether an observed difference or relation in our corpus sample is consistent with mere chance. We may, for instance, in a corpus sample, have found that there is a difference in terms of the frequency of attributive adjectives between academic writing and lectures in medical discourse when we look at the descriptive statistics, and we wish to test whether this difference is an artefact of this particular sample from the population (i.e., if the difference in frequency is merely due to chance), or if we may in fact expect to see a difference between these modes, had we had access to the full population. To do so, we formulate the following two hypotheses – a null hypothesis, and its opposite, the alternative hypothesis, such as:

- Null hypothesis (H_0): in the population, there is no difference in means.
- Alternative hypothesis (H_1): in the population, there is a non-zero difference in means.

We run an independent samples *t*-test, say, and if our obtained *p*-value is below our alpha level (typically 0.05), we reject the null hypothesis and thus conclude that our samples were drawn from populations with some non-zero mean difference (or relation).⁸

However, even if we can reject the null hypothesis, the conclusion that we are allowed to draw is very vague: ‘There is a non-zero mean difference between academic writing and lectures in medical discourse when it comes to attributive adjectives’. We are not in any way building upon previous findings on the topic – that is, our test is completely uninformed by prior research. Given the existence of decades of linguistics research, it would seem that we can (and indeed should) almost always formulate and test a more specific hypothesis than that.

To some extent, we can build in findings from previous research by using a one-tailed test (instead of the more standard two-tailed test, as exemplified above). In the case of attributive adjectives, as in the example above, we have a large body of research (e.g., Biber and Gray [2016] and the papers in Biber *et al.* [2022]) concluding that written discourse relies more

⁸ If our *p*-value is above our alpha, we have to retain the null hypothesis – though note that we cannot say that we have proven it correct (it is just that we do not currently have sufficient evidence to reject it).

heavily on attributive adjectives than spoken discourse. Building cumulatively on that body of research, we can use a one-tailed test, which would allow us to pose a directional hypothesis as follows:

- H_0 : in the population, there are not higher mean frequencies in the written discourse
- H_1 : in the population, there are higher mean frequencies in the written discourse

That is, instead of asking ‘is there a difference between groups: yes or no?’ over and over in our studies of attributive adjectives (*etc.*), we have here made at least some attempt at incorporating existing knowledge into our hypotheses. This kind of hypothesis can be tested with commonly used techniques in the field, such as one-tailed *t*-tests.

However, as we accumulate more and more knowledge on a given topic, we are likely to want to posit and test more specific hypotheses than that, ones that scale up to more advanced models, should we need them for our research questions. For example, as we will see in the case study below, we may want to test a hypothesis stating that spoken registers vary among a more homogenous pattern than written registers do with regard to features of grammatical complexity. To do so, we can use techniques from the structural equation modeling (SEM) framework (e.g., Hancock and Schoonen [2015]; see also Larsson *et al.* [2021] and Larsson *et al.* [2022] for introductions aimed at a corpus linguistics audience).⁹ Techniques under this methodological umbrella are well suited to testing specific hypotheses based on findings from previous research in that they are designed to enable us to (a) pose specific hypotheses and (b) assess the goodness-of-fit in relation to our data. Put differently, in this framework, we use models to test a specific hypothesis, and if the model corresponds sufficiently closely with our data, we retain our hypothesis; if not, we reject our hypothesis. Note that we are testing our specific hypothesis, whatever it may be, and we are not limited to testing the generally uninformed default null hypothesis.

In the next section, we formulate and test specific hypotheses in the context of a study on grammatical complexity. In an attempt to use minimally sufficient techniques (see Egbert *et al.* [2020: Chapter 6]), we use more simple techniques from the SEM framework, along with a version of a one-tailed test to illustrate what such a design may look like.

⁹ It is difficult to give a precise definition of SEM as it covers an ever-widening array of techniques, but the characteristic shared by all of these techniques is that they assess the consistency of specific hypotheses motivated by previous research with the characteristics of our data (e.g., means, variances and correlations).

4. A case study: grammatical complexity in spoken and written registers

4.1 Introduction and rationale

Grammatical complexity is an example of a topic that, at least when it comes to studies of English, can be said to have moved along the trajectory from exploratory analyses (Stage 1) to systematic knowledge building (Stage 2). Existing studies have yielded specific hypotheses that are now ripe to be put to the test (Stage 3).¹⁰ The topic has been approached from many different perspectives and studied according to different frameworks (see Biber *et al.* [2022] for an extended discussion). In this study, we approach it at the level of the text to study the extent to which prose from a given register includes different types of complexity features, such as finite and non-finite relative clauses, attributive adjectives and pre-modifying nouns. Studies in this tradition, particularly from a register-functional perspective (e.g., Biber *et al.* [2022], Biber and Gray [2016] and Biber and Gray [2010]), have repeatedly found that it is a simplification to refer to ‘more or less complex’ in absolute terms, as there are different kinds of complexities. That is, numerous studies have noted the differing grammatical complexities of spoken and written registers. Specifically, these studies have shown that spoken registers rely on clausal complexity features, whereas written registers rely on phrasal complexity features (see, for example, Biber [1992] and Biber *et al.* [2022]).

One less-noticed finding from previous research, however, is that the spoken and written modes differ fundamentally in the extent to which the use of complexity features can be manipulated, in that the spoken registers ‘are produced and comprehended in real-time, setting a cognitive ceiling for the syntactic and lexical complexity typically found in these [registers]’ (Biber, 1988: 163). This fundamental difference is described in greater detail in Biber (1992: 159) where the findings of this study lead to the following conclusion:

[W]ritten registers differ widely among themselves in both the extent and kinds of discourse complexity, while spoken registers follow a single pattern with respect to their kinds of complexity, differing only with respect to extent.

Put differently, given the production constraints of spoken discourse, where only very limited pre-planning and post-editing is possible and where our

¹⁰ We fully acknowledge that neither this case study, nor the studies that it draws on and builds on, enable us to generalise beyond the English language in these contexts. As correctly pointed out by one of our reviewers, whilst the general principles discussed in the first sections of this paper hold cross-linguistically, we would need more research on other languages to be able to make claims that go beyond English in this case study. More generally, for cumulative knowledge building to be able to cross language boundaries, more research on other languages is sorely needed, albeit that this research falls outside the scope of this case study.

production typically needs to be processed in real time, we may expect there not to be much variation among spoken registers. By contrast, given the option of pre- and post-processing for written registers, producers of written prose, it would seem, are likely to be able to adapt to the situational characteristics of the context in which the text was written. For example, writers may more readily take the communicative purpose (informational, narrative, *etc.*) into consideration and adapt their writing accordingly than speakers are able to.

In this case study, we use these claimed fundamental differences between the discourse complexities of speech and writing as testable hypotheses. In particular, we test the two major claims based on the findings from Biber (1992):

- (1) All complexity features follow a single pattern across spoken registers, *versus* multiple patterns of variation across written registers; and,
- (2) For any individual complexity feature, there is more variability across registers in the written mode than in the spoken mode.

4.2 Data and complexity features

Our examination is based on analysis of multiple complexity features in a corpus of spoken and written registers. The texts included come from existing corpora, including the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWALL), the BNC 2014 Spoken Corpus; the lectures come from the British Academic Spoken English (BASE) corpus and the Yale open access lecture series. The corpora are matched pairwise across the modes in terms of level of interactivity, level of expertise of the audience and communicative purpose, as shown in Table 1. In total, we have data from 1,229 files, amounting to just over 6 million words (see Table 2 for an overview). All the texts were tagged with the Biber tagger (Biber, 2006) and the frequencies were normalised per 1,000 words. The specific statistical methods used will be explained in the respective results section.

As previous research has shown that certain linguistic features tend to be associated with spoken and written discourse, respectively (e.g., Biber *et al.*, 2022), we selected eight features that represent the different sections of the structure-syntactic function space, namely attributive adjectives, pre-modifying nouns, non-finite relative clauses, finite relative clauses, finite adverbial clauses, *that*-complement clauses, and *to*-complement clauses and adverbs, shown in Examples 1 to 4.

- (1) “How do we know *that we live in a four-dimensional universe*” she asked a crowd who filled the Hayden Planetarium on a stormy night last week.

(attributive adjectives = **bold**)

that-complement clauses = underlined and italics

finite relative clauses = underlined)

- (2) Detailed vision is required from the retinal periphery, because cell densities remain high over much of the retina.
 (pre-modifying nouns = **bold**
 finite adverbial clauses = underlined)
- (3) It might be worse because we don't **really** expect to have our views heard.
 (adverbs = **bold**
 to-complement clauses = underlined)
- (4) Others have business models based on scarcity.
 (non-finite relative clauses = underlined)

Mode		Situational characteristics		
Spoken registers	Written registers	Inter-activity	Level of expertise of the audience	Communicative purpose
Conversational opinion	Opinion blogs	High	Low	Personal/stance
Classroom teaching	Textbooks	Medium	Medium	Informational
Formal lectures	Research articles	Low	High	Informational
Conversational narratives	Fiction	Low	Low	Narrative

Table 1: Pairings of corpora for situational characteristics.

Register	No. of texts	Word count
Classroom teaching	145	103,8477
Conversational narrative	312	154,141
Conversational opinion	217	77,424
Lectures	94	592,822
Fiction	89	2,226,544
Opinion blogs	177	263,153
Research article	139	1,156,956
Textbooks	56	487,648
Total	1,229	5,997,165

Table 2: Overview of corpora used.

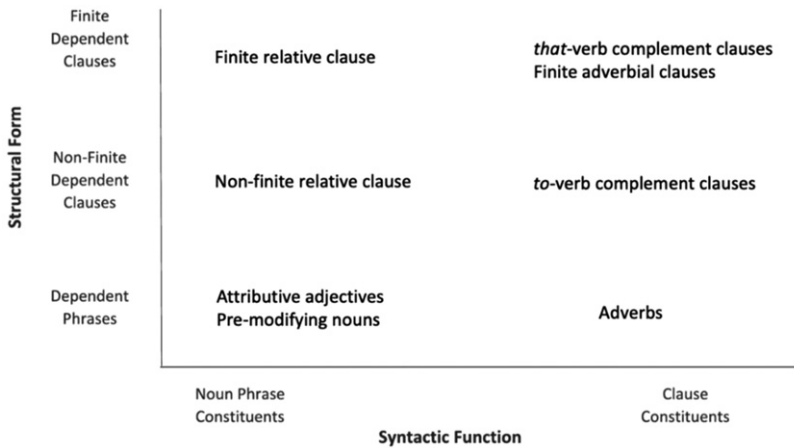


Figure 2: Taxonomy of selected complexity features according to structure and syntactic function (adapted from Biber and Gray [2010] and Biber *et al.* [2022]).

In more detail, we made use of the taxonomy of complexity features according to their structural form and syntactic function from Biber and Gray (2010) and Biber *et al.* (2022), as shown in Figure 2. In this framework, we can simultaneously describe each feature, such as a finite relative clause (e.g., *the woman **who showed up on time** is over there*), in terms of its form (as a finite dependent clause) and its function (a noun-phrase constituent). Specifically, the full space takes into consideration structural types (finite dependent clauses, non-finite dependent clauses and dependent phrases) as well as the cross-cutting syntactic functions (noun phrase constituent and clause constituent).

4.3 Results and discussion

In this section, we report on how we test our specific hypotheses resulting from previous studies of grammatical complexity: Hypothesis 1 in Section 4.3.1 and Hypothesis 2 in Section 4.3.2. Both hypotheses are operationalised through models, where we use fit indices to assess the adequacy of our hypothesised model in relation to our data. That is, the fit indices help us to understand the degree to which our hypothesised model is consistent with the reality of the data, thus enabling us to evaluate specific hypotheses based on previous research and contribute cumulatively to research on the topic.

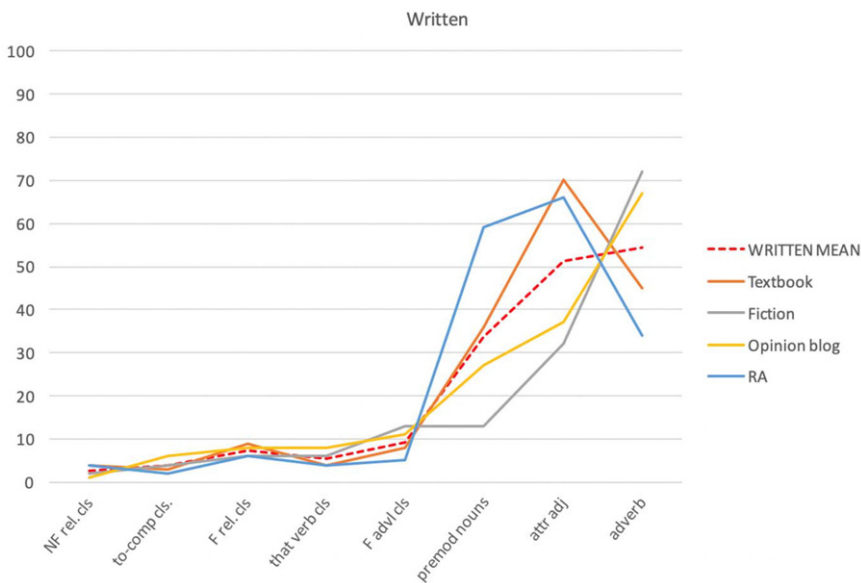


Figure 3: Mean per-text frequencies of each feature along with the mode mean for the written registers.

4.3.1 Hypothesis 1: all complexity features follow a single pattern across spoken registers, *versus* multiple patterns of variation across written registers

To test our first hypothesis, we followed the following steps: (i) operationalise ‘a single pattern’ *versus* ‘multiple patterns’, and (ii) assess statistically the adequacy of the hypothesis. For (i), we operationalised ‘a single/multiple pattern’ (or lack thereof) for a given feature using the variance of register means around the mode mean. We do not have any hypotheses related to within-group variance, so the focus here is exclusively on between-group variance. A graphical representation of the variability of the register means around the mode mean in our data can be found in Figures 3 and 4. Specifically, we looked at the aggregate variance of the register means around the mode mean. That is, for each feature in each register, we looked at how far each register mean is from the mode mean and then added these up. For example, for attributive adjectives, the mode mean for the written registers is 51 instances per 1,000 words. The registers included for this mode has the following means per 1,000 words: 70 (textbooks), 66 (research articles), 37 (opinion blogs) and 32 (fiction). Variance, which tells us how far our frequencies are spread out from their mean, is calculated by squaring each observation’s deviation from the mean and summing the squares, and then dividing this number by the number of observations minus one (i.e., the

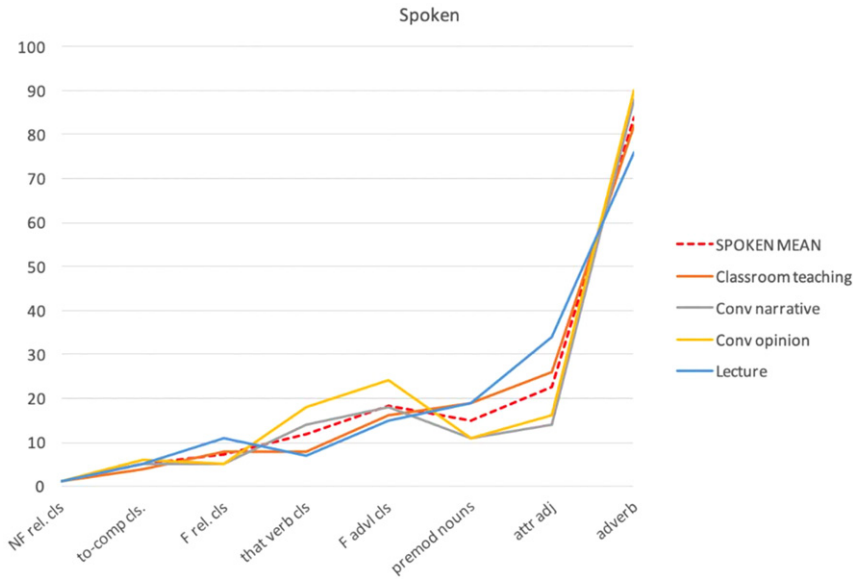


Figure 4: Mean per-text frequencies of each feature along with the mode mean for the spoken registers.

degrees of freedom). The variance for attributive adjectives in the written mode is calculated as follows:

$$\begin{aligned}
 \text{Variance}_{\text{attr.adj}} &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\
 &= \frac{(70-51)^2}{3} + \frac{(66-51)^2}{3} + \frac{(37-51)^2}{3} + \frac{(32-51)^2}{3} \\
 &= \frac{361}{3} + \frac{225}{3} + \frac{196}{3} + \frac{361}{3} \\
 &= 1012.333
 \end{aligned}$$

For (ii), we totalled up the eight feature-specific means' variances across all features in a register to be able to assess statistically the claim that there is more variability in the written mode than in the spoken mode. We used a one-sided design, thus building on the idea introduced in Section 3, to assess fit of a hypothesised model in relation to our data. Note, however, that unlike the one-tailed test example from Section 3.1, we assess the viability of a specific hypothesis, rather than the viability of a null hypothesis. We elaborate on this below.

The findings from previous research outlined in Section 4.1 and the first hypothesis that follows from them would have us predict that the aggregate variance would be higher in the written mode. That is, if the hypothesis is correct, we should expect to find that people are able to adapt their writing to the situational characteristics of the different registers to a

greater extent than the people whose spoken production is included in our data, meaning that we would expect more variability in the written mode. This hypothesis is tested in a first model (Model 1a). We also compare the fit of this model to a competing model stating that the aggregate variance is the same across the two modes (Model 1b).¹¹

We fitted the two competing models in Mplus (Muthén and Muthén, 2022) using robust maximum-likelihood (ML) estimation. Much simplified, ML sets out to find potential population values for, in our case, the variances across modes, that are consistent with pattern suggested by our hypotheses. The extent to which these values correspond to what we see in the actual data is assessed through model fit. Assuming that the first hypothesis best corresponds to what our data show, then Model 1a should have better fit than the competing model, Model 1b.

That is, to be able to know which of the two competing models best corresponds to our data, we need a way to assess model fit. The notion of model fit is not new to the field and has been introduced through techniques such as multiple regression, where we use measures such as R^2 and the Akaike Information Criterion (AIC). We will also use the AIC here as a way to help us select the model with the best fit. In more technical terms, AIC is an estimator of model replicability that enables us to assess the combined quality and parsimony of a model relative to other models, and as such provides a means for model selection (see, for example, Akaike [1973]). A lower AIC indicates better relative fit, where better fit means that our hypothesised model more accurately represents the patterns in the data.

When we ran the two models, the results show that Model 1a has an AIC of 59,657.2 and Model 1b has an AIC of 59,835.0, meaning that the Δ AIC (i.e., the difference in absolute terms between the models) is 177.8, in favour of our hypothesis. Had the Δ AIC difference between the models been, say, below 10, there is a considerable risk that sampling error may have caused us to mistakenly select the wrong model (Burnham and Anderson, 1998: Section 2.6), but our Δ AIC is well above that. This means that we have, indeed, gathered evidence in support of the first hypothesis that there is more variability across registers in the written mode for all features taken together.

4.3.2 Hypothesis 2: for any individual complexity feature, there is more variability across registers in the written mode than in the spoken mode

We will now dig deeper and look at each feature separately to test our second hypothesis. Here, we have to (i) operationalise ‘variability’, and (ii) assess statistically the adequacy of the hypothesis. To test the first hypothesis above,

¹¹ Note that we cannot test whether two groups have the same variance using the null hypothesis framework: retaining the null only allows us to conclude that we do not currently have sufficient evidence to reject it. In a model-based framework, by contrast, we can test the viability of a model stating that two groups have the same variance.

we looked at variance from the mode mean to get at the notion of a ‘pattern’. For this second hypothesis, by contrast, we use the means of each register for each feature without taking the mode mean into consideration, as explained below.

For (i), we operationalised variability with the help of its opposite: lack of variability. That is, for us to be able to provide support for our second hypothesis, then models constraining all register means per feature to be equal in the spoken data should have better fit than the corresponding, competing models constraining the means in the written data.¹² For example, we expect a model where the means for non-finite relative clauses is the same across all registers in the spoken data to have better fit than the corresponding model in the written data. To be able to constrain all the register means per feature to be identical and thus attain (ii), we used a technique from the SEM family, namely ‘mean structure models’. Models of this kind enable tests of hypotheses related to similarities (or differences) in means across groups in that we can, for example, force means to be equal and see to what extent such a model fits our data. More generally, like other SEM techniques, the model output provides information about the fit of our model in terms of our observed data. Like the analysis in Section 4.3.1, the less discrepancy there is between our hypothesised models and the data, the better the fit. However, we now extend the notion of fit to incorporate both absolute and relative fit: one may characterise the discrepancy between a model and the data in terms of ‘absolute fit’, where a model fits the data to a degree that the field considers acceptable, and ‘relative fit’, where a model fits the data better than a competing model.

More concretely, the absolute fit of the mean structure models in this section evaluates the hypothesis that the limiting production circumstances of the spoken mode would lead spoken registers to exhibit no variation; good fit in absolute terms would lead us to consider this second hypothesis to be supported. However, even in also comparing the fit of such models for the spoken registers to those of the written registers, we allow for a slightly less strong version of our hypotheses to be tested—namely that there is more limited variation in the spoken registers than in the written registers for all the features investigated; better fit in relative terms for the spoken mode than the written mode would lead us to retain this version of the hypothesis.

We ran these models in MPlus with robust maximum-likelihood estimation using the per-text means for each feature as our input (sample code can be found in Appendix A). To assess relative fit, we once again used AIC. To assess absolute fit for the individual models, we will turn to a commonly used

¹² Note that we cannot run a test of this kind using the NHST framework, as we are here formally testing the equivalent of a nil hypothesis (a null hypothesis of no difference). As mentioned above, in the NHST framework, we can never say that we have proven a nil hypothesis right, only that we do not currently have sufficient evidence to reject it; in a model-based framework, by contrast, a hypothesis of ‘no difference’ is possible to assess and we can then see to what extent it exhibits misfit, given the data.

	adverb	attr adj	pmod N	F advl cls	that cls	F rel cls	to cls	NF rel cls
M1: No variab. SPO								
AIC	10182.9	9480.0	9380.3	7730.6	7616.4	6933.3	5956.0	4427.4
M2: No variab. WR								
AIC	10633.1	9517.7	9585.8	7944.9	7537.2	6810.2	6161.6	4513.9
[Delta AIC]	450.2	37.7	205.5	214.3	79.2	123.0	205.5	86.5

Table 3: AIC and Δ AIC for all the mean structure models fitted (shading marks the best-fitting model).

fit index in the SEM framework, the root mean-square error of approximation (RMSEA).¹³ Whilst cut-off values have been offered (see, for example, Hu and Bentler [1999]), the methodological community generally believes that there can be no universal (i.e., non-model-specific) standards (see, for example, Hancock and Mueller [2011] and McNeish *et al.* [2017]). Still, values closer to zero indicate better fit, with values under 0.08 historically having been deemed desirable.

The results showed that in terms of individual model fit in absolute terms, no model reached the above fit guideline: of all the models in the spoken and written data, the best-fitting one, pre-modifying nouns, had an RMSEA of 0.10 for the model with constrained spoken register variances. We can thus draw the conclusion that there is, in fact, register variation in both modes. That is, we cannot retain the strong version of our second hypothesis that the limiting production circumstances of the spoken mode would lead spoken registers to exhibit no variation.

However, even if the absolute fit did not reach historically recommended levels, the relative fit will still enable us to evaluate the weaker version of our second hypothesis that there is more limited variation in the spoken registers across features by looking at the relative fit across the modes and features. That is, we compare the models for the spoken mode to those of the written mode to see whether the fit is better overall in the spoken data. Indeed, as shown in Table 3 where the AIC values and the Δ AIC for each model is displayed, for six of eight features, the models assessing the hypothesis that there is no variability among the register means in the spoken data had better fit than that of the written mean model counterpart. Only *that*-complement clauses and finite relative clauses had less variability among the written registers than the spoken ones. There is thus generally more variability across registers in the written mode than in the spoken mode, as hypothesised.

Based on the outcomes from both analyses, we see that the results mostly support the prediction from previous research that (a) all complexity features follow a single pattern across spoken registers, *versus* multiple patterns of variation across written registers, and (b) that there is less variability across the registers in the spoken mode than in the written mode.

¹³ Although other indices exist, such as the standardised root mean-square residual (SRMR) and the comparative fit index (CFI), these are not appropriate for the current models as those indices' focus is on modelling relations amongst variables.

The exceptions (*that*-complement clauses and finite relative clauses) are clausal features, most common in conversation.

It thus seems that we, based on our data, have support for the claim that the written mode, more than the spoken mode, enables language users to adapt their language to the level of interactivity, expertise of the audience, and communicative purpose, as evident by there being more variation among the written registers that differ in these respects. We see this in particular for the high-frequency features. Nonetheless, the fact that we do not see a perfect split between the spoken and written mode in terms of the variability of individual features suggests a slightly more complex picture than was hypothesised in Biber (1992). It seems that we, in future studies, should also take into consideration the frequency of the feature and the extent to which it is associated with the written or the spoken mode.

5. Conclusion

We would like to encourage authors in the field to identify explicitly the specific findings and supposed generalisations made by previous research and then compare those specific generalisations across studies to describe the claimed state of our knowledge. As a subsequent step, these findings can inform studies where specific hypotheses are tested. For many (if not most) topics for which we use corpus techniques to study language use, we arguably have enough knowledge to be able to posit more specific hypotheses than what we tend to do. We can use a wide array of different statistical techniques to test such specific hypotheses.

A cumulative approach to knowledge building would enable us to engage more deeply with the claimed generalisable findings from previous research in a way that, we believe, would help move the field's state-of-the-art forward. If we also work towards adding to knowledge where there is already a solid foundation, we will be able to reach much further in terms of our collective understanding of language and language use.

References

- Akaike, H. 1973. 'Information theory and an extension of the maximum likelihood principle' in B.N. Petrov and F. Csáki (eds) 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, 2–8 September 1971, pp. 267–81. Budapest: Akadémiai Kiadó, Republished in S. Kotz and N.L. Johnson (eds). 1992. *Breakthroughs in Statistics*, vol. I, pp. 610–24. Berlin: Springer-Verlag.
- Atkinson, D. 1999. *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. Mahwah, New Jersey: Lawrence Erlbaum.

- Bazerman, C. 1988. *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Madison: University of Wisconsin Press.
- Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D. 1992. 'On the complexity of discourse complexity: a multidimensional analysis', *Discourse Processes* 15, pp. 133–63.
- Biber, D. and B. Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge: Cambridge University Press.
- Biber, D. and B. Gray. 2010. 'Challenging stereotypes about academic writing: complexity, elaboration, explicitness', *Journal of English for Academic Purposes* 9 (1), pp. 2–20.
- Biber, D., B. Gray, S. Staples and J. Egbert. 2022. *The Register-Functional Approach to Grammatical Complexity: Theoretical Foundation, Descriptive Research Findings, Application*. Oxfordshire: Routledge.
- Egbert, J., T. Larsson and D. Biber. 2020. *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Cambridge: Cambridge University Press.
- Gross, A.G., J.E. Harmon and M. Reidy. 2002. *Communicating Science: The Scientific Article from the 17th Century to the Present*. New York: Oxford University Press.
- Hancock, G.R. and R.O. Mueller. 2011. 'The reliability paradox in assessing structural relations within covariance structure models', *Educational and Psychological Measurement* 71, pp. 306–24.
- Hancock, G.R. and R. Schoonen. 2015. 'Structural equation modeling: possibilities for language learning researchers', *Language Learning* 65, pp. 160–84.
- Hu, L.-T. and P.M. Bentler. 1999. 'Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives', *Structural Equation Modeling: A Multidisciplinary Journal* 6 (1), pp. 1–55.
- Larsson, T., J. Egbert and D. Biber. 2022. 'On the status of statistical reporting versus linguistic description in corpus linguistics: a ten-year perspective', *Corpora* 17 (1), pp. 137–57.
- Larsson, T., L. Plonsky and G.R. Hancock. 2021. 'On the benefits of structural equation modeling for corpus linguists', *Corpus Linguistics and Linguistic Theory* 17 (3), pp. 683–714.
- Larsson, T., L. Plonsky and G.R. Hancock. 2022. 'On learner characteristics and why we should model them as latent variables', *International Journal of Learner Corpus Research* 8 (2), pp. 237–60.
- Larson-Hall, J. and L. Plonsky. 2015. 'Reporting and interpreting quantitative research findings: what gets reported and recommendations for the field', *Language Learning* 65 (1), pp. 127–59.
- McEnery, T. and V. Brezina. 2022. *Fundamental Principles of Corpus Linguistics*. Cambridge: Cambridge University Press.
- McNeish, D.M., J. An and G.R. Hancock. 2017. 'The thorny relation between measurement quality and fit index cutoffs in latent variable models', *Journal of Personality Assessment* 100, pp. 43–52.

- Muthén, L.K. and B.O. Muthén. 2022. *Mplus: Statistical Analysis with Latent Variables: User's Guide (Version 8)*.
- Paquot, M. and M. Callies. 2020. 'Promoting methodological expertise, transparency, replication, and cumulative learning: introducing new manuscript types in the International Journal of Learner Corpus Research', *International Journal of Learner Corpus Research* 6 (2), pp. 121–4.
- Sønning, L. and V. Werner. 2021. 'The replication crisis, scientific revolutions, and linguistics', *Linguistics* 59 (5), pp. 1179–206.
- Zeigler, D. 2012. 'Evolution and the cumulative nature of science', *Evo Edu Outreach* 5, pp. 585–8.

Appendix A: Sample code to illustrate how the models were set up can be found below. Any annotation is preceded by an exclamation point.

TITLE:

Mean constraint model, using ADVERB

DATA:

FILE IS JEL_demo_data.csv;

VARIABLE:

NAMES ARE

REGISTER

ATTRADJ ADVERB NFRELCLS FADVLCLS

FRELCLS PMODN TOCLS THATCLS;

USEVARIABLES ARE ADVERB;

GROUPING IS Register (1=RAW1 2=OpinBlogW2 3=LecS1 4=FicW4
5=ConvOpinS2 6=ConvNarrS4 7=ClassTeachS3 8=TextW3);

!RA = research article

!OpinBlog = opinion blogs

!Lec = lecture

!Fic = fiction

!ConvOpin = conversational opinion

!ConvNarr = conversational narrative

!ClassTeach = classroom teaching

!Text = textbooks

!suffix W = written (measures W1-W4)

!suffix S = spoken (measures S1-S4)

ANALYSIS:

ESTIMATOR IS MLR;

!This is a robust correction to maximum likelihood

MODEL:

!This sets up the model to be estimated within each group

ADVERB; !This represents the variance

[ADVERB]; !This represents the mean

MODEL RAW1:

!This sets up the model to be estimated within the RA group

ADVERB;

[ADVERB] (meanW1); !The bracketed expression assigns a
!unique name to the group mean

MODEL OpinBlogW2:
 ADVERB;
 [ADVERB] (meanW2);

MODEL LecS1:
 ADVERB;
 [ADVERB] (meanS1);

MODEL FicW4:
 ADVERB;
 [ADVERB] (meanW4);

MODEL ConvOpinS2:
 ADVERB;
 [ADVERB] (meanS2);

MODEL ConvNarrS4:
 ADVERB;
 [ADVERB] (meanS4);

MODEL ClassTeachS3:
 ADVERB;
 [ADVERB] (meanS3);

MODEL TextW3:
 ADVERB;
 [ADVERB] (meanW3);

MODEL CONSTRAINTS:
 !The following imposes equality constraints on
 !all four spoken means, which will induce
 !some degree of badness of fit.
 !The written means are left unconstrained.
 meanS1 = meanS2;
 meanS2 = meanS3;
 meanS3 = meanS4;

OUTPUT:
 sampstat; !This provides descriptive statistics by group