Article

Tove Larsson*, Luke Plonsky and Gregory R. Hancock

On the benefits of structural equation modeling for corpus linguists

https://doi.org/10.1515/cllt-2020-0051
Received August 22, 2020; accepted November 23, 2020;
published online December 10, 2020

Abstract: The present article aims to introduce structural equation modeling, in particular measured variable path models, and discuss their great potential for corpus linguists. Compared to other techniques commonly employed in the field such as multiple regression, path models are highly flexible and enable testing *a priori* hypotheses about causal relations between multiple independent and dependent variables. In addition to increased methodological versatility, this technique encourages big-picture, model-based reasoning, thus allowing corpus linguists to move away from the, at times, somewhat overly simplified mindset brought about by the more narrow null-hypothesis significance testing paradigm. The article also includes commentary on corpus linguistics and its trajectory, arguing in favor of increased cumulative knowledge building.

Keywords: corpus linguistic methodology, structural equation modeling, measured variable path models, model-based reasoning, null-hypothesis significance testing

1 Introduction

Studies in corpus linguistics studies have seen a steady increase in the use of sophisticated statistical methods in recent years. Answering the call from corpus methodologists for techniques better suited to the multivariate nature of corpus linguistic data (e.g., Gries 2015a), the field has gone from relying heavily on descriptive

***Corresponding author: Tove Larsson**, English Department, Uppsala University, Uppsala, Sweden; and Centre for English Corpus Linguistics, Université catholique de Louvain, Louvain-la-Neuve, Belgium, E-mail: tove.larsson@engelska.uu.se. https://orcid.org/0000-0002-0489-2697
Luke Plonsky, English Department, Northern Arizona University, Flagstaff, AZ, USA, E-mail: luke.plonsky@gmail.com
Gregory R. Hancock, Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, USA, E-mail: ghancock@umd.edu

statistics and monofactorial methods to using techniques such as different forms of regression analyses and classification trees (Larsson et al. under review). Despite these recent advancements, however, there are still questions pertaining to the complex nature of language that our current methods cannot easily address. Furthermore, the lack of familiarity with appropriate techniques may even constrain the types of questions researchers pose. In an effort to expand our analytic repertoire, this article seeks to introduce structural equation modeling (SEM) and discuss its great potential for corpus linguistic analysis in a nontechnical manner; specifically, we focus on measured variable path models, a fundamental building block of SEM (for an overview of other techniques within this framework relevant to applied linguists, see Hancock and Schoonen 2015). The article also includes commentary on the field of corpus linguistics with regard to cumulative knowledge building. No prior knowledge of SEM techniques is assumed; instead, we will build on readers' knowledge of a related technique, namely multiple regression.

SEM is a powerful analytical framework that encompasses a large array of statistical techniques (e.g., path analysis, confirmatory factor analysis). These techniques are commonly used in other social and behavioral sciences (including neighboring fields such as Second Language Acquisition) to investigate theories involving causal effects of one or more independent variables on one or more dependent variables, and even among those dependent variables themselves. Despite the many strengths and versatility of SEM, however, its techniques remain practically unknown in corpus linguistics (with the notable exception of Gries 2003; and studies using confirmatory factor analysis, e.g., Biber 2001; Hu et al. 2019). To be clear, our intent is not to introduce techniques that add unnecessary complexity (see the discussion of minimally sufficient statistical methods in Egbert et al. 2020). Rather, we aim to introduce tools that allow corpus linguists to answer research questions that are otherwise beyond reach given current statistical methods, thereby helping the field move forward.

The article is structured as follows: Section 2 offers a conceptual overview of benefits of measured variable path models compared to multiple regression. After having outlined the reasons why corpus linguists may want to add path models to their toolbox, we provide a more concrete introduction to measured variable path models in Section 3. Section 4 presents a worked example that serves to illustrate how this technique can be applied to corpus data; it also offers an outlook on two other techniques for readers interested in further expanding their SEM repertoire. Section 5 provides a concluding discussion.

2 Benefits of using path models

This section highlights some of the advantages of measured variable path models to show why the field would benefit from adopting them. Section 2.1 outlines

general advantages compared to multiple regression, whereas Section 2.2 discusses benefits specific to corpus linguists, and how these techniques may encourage certain paradigm changes in the field.

2.1 Why use measured variable path models?

Measured variable path models belong to the class of *covariance structure models*, which also includes *confirmatory factor analysis* and *latent variable path models*, among others (see Section 4.3). Measured variable path models (henceforth *path models*) are, like other covariance structure models, designed to help us understand why and how variables relate, that is, *covary*. As such, they aid in a task that is foundational to most linguistic inquiry: understanding what underlying mechanisms are associated with one or more given outcomes. Path models are proposed on theoretical grounds; that is, the overall design of the model and the direction of the causal relations specified in a model should be based on theory and/or findings of previous studies. Although there is no denying that SEM techniques, including path models, are more advanced and thus perhaps more time-consuming to learn than multiple regression, we will argue that this is a worthwhile investment, as compared to other, more commonly used techniques in corpus linguistics, path models have numerous advantages. We will focus here on two main ones: they (i) offer greater flexibility and (ii) enable researchers to move toward large-picture, model-based reasoning.

First, like multiple regression, path models offer regression-type coefficients, R^2 values, and allow us to carry out hypothesis testing procedures; indeed, multiple regression models are a special case of path models. Unlike multiple regression, however, path models are highly flexible in terms of the specification of hypothesized relations among variables and with regard to variable structure, thus enabling a broader range of research questions to be addressed, as outlined below.

In addition, even in cases where variables' relations could be investigated using multiple regression (several independent variables and one dependent variable; e.g., *gender* and *age* as they relate to the frequency of *hedges per clause*), path models offer a way to investigate these relations with greater control. For instance, while the independent variables in a multiple regression are free to covary whether or not we have reason to believe that they do based on theory, path models enable us to specify such relations among independent variables as we see fit to address specific research questions (e.g., they can be estimated freely, set to zero, constrained equal to each other). In the example above, we arguably have no reason to believe that the values for *gender* and *age* should be related in any way; within a path model we have the ability to more accurately represent the

hypothesized population, in this case, by constraining the covariance of *gender* and *age* to zero.

When it comes to overall model structure, path models (unlike traditional multiple regression) allow for inclusion of multiple dependent variables, and investigation of both mediating variables (*mediators*) and moderating variables (*moderators*). In brief, mediators help to explain the mechanism governing a causal relation between two other variables (Pearl 2012). To use an example from a published study, Fong and Ho (2017) looked at the extent to which factors such as *working memory* and *vocabulary skills* affect *listening comprehension*, hypothesizing based on theory and previous studies that *working memory* affects *vocabulary skills*, which in turn affects *listening comprehension*. Here, *vocabulary skills* is a mediator that is hypothesized to help explain the relation between the independent variable *working memory* and the dependent variable *listening comprehension*: *working memory → vocabulary skills → listening comprehension*.

Mediators are not to be confused with moderators. A moderator (often assessed through an *interaction term* formed from the independent variable and the moderator) explains *when* a relation holds between variables and the strength of that relation. For example, the difference in relative frequency of usage of *-ise* versus *-ize* (e.g., *organise* vs. *organize*) across regional background of the participants (British English vs. American English) might vary depending on speaker age. That is, without taking age into consideration, the regional background is likely to have a clear effect on the distribution of the spelling variants, such that overall, *-ise* is preferred in British English whereas users of American English prefer *-ize*. However, with age (the moderator) added, the picture might look different: For example, while older users of British English may still show a marked preference for the *-ise* spelling variant, the preference for this spelling variant for younger users of British English may be very minor or completely lacking.

In addition to the ability to include mediators and moderators, path models offer additional flexibility in that they enable researchers to compare paths within models, evaluate longitudinal relations including the mutual relations of variables on each other as they develop over time, and compare models across multiple groups. Moreover, path models can be expanded to accommodate *latent* variables (see Section 4.3) to counteract the attenuating effects of measurement error in observed variables (see, e.g., Hancock and Schoonen 2015).

Second, path models provide a framework for moving away from overreliance on the somewhat simplified mindset associated with null-hypothesis significance testing (NHST), toward *model-based reasoning* (i.e., a mindset where relations between variables are viewed as part of a larger explanatory system, one which is evaluated as such). As will be familiar to readers, in the NHST epistemological system, differences or relations found in a sample are compared with a null

hypothesis stating that in the population there is some specific (typically zero) difference or relation. If there is sufficient evidence to reject this null hypothesis, usually operationalized by some p-value below a target α level (typically $p < 0.05$), the difference or relation is deemed 'statistically significant' and a difference/relation other than that specified in the null hypothesis (specifically lower or higher) is assumed to exist in the population from which the sample was drawn; otherwise the null hypothesized value is retained. As a result, for example, $p = 0.049$ leads to a different inference about a population than $p = 0.051$, although differing in probability by a mere two thousandths.

NHST has been criticized on several accounts, including its overly rigid dichotomy between the two inferential outcomes, and some of its critics have even suggested that we should stop relying on this framework altogether (see, e.g., Koplenig 2019; Plonsky 2015). With the rise of a new generation of accessible software packages, Bayesian statistics has gained ground as an alternative approach, one that instead allows us to express a *degree* of belief in an event while incorporating prior knowledge (see, e.g., Lee 2012; Levy and Mislevy 2016; Wallis 2020 for an introduction). While SEM models can be analyzed using a Bayesian framework (see Levy and Choi 2013), we will here limit our discussion to SEM, and specifically path models, as carried out within the far more customary frequentist framework, which includes NHST as pertains to specific model parameters. Such tests, however, are embedded within a broader path model, thus having the defining advantage of being part of a framework of model-based reasoning for understanding a system *as a whole*. Although able to complement each other, traditional NHST and modern model-based reasoning are based on very different epistemologies:

> To fully appreciate the import of this shift [toward model-based reasoning], we must recognize that the epistemological focal point has completely shifted. Within the epistemological tradition emerging from NHST, the focus is the null hypothesis, H_0, which we assume until it can be rejected in favor of a reasonable (but broad and relatively unspecified) alternative. […] The postrevolution focal point is no longer the null hypothesis; it is the current model. This is exactly where the researcher—the scientist—should be focusing his or her concern. […] The null hypothesis has always been a creation of the statistician. But for the scientist–researcher, his or her own research hypothesis is the natural focus. (Rodgers 2010: 4–5)

Put differently, instead of hoping that we can reject the null in favor of a crudely articulated alternative hypothesis (e.g., the population correlation $\rho \neq 0$), path models enable us to pose and assess the fit of a very specific but generally more comprehensive hypothesis (the system that we have built based on theory and/or previous research, *viz.* our model). If our proposed model does not exhibit a sufficiently good fit (meaning there is insufficient consonance between that specific

model and our data), then the model—as a whole or in part—can be rejected. That is, assuming beliefs about relations among variables are proposed on theoretical grounds, path models enable those relations to be evaluated, providing information about the sign and magnitude of the connections hypothesized within a larger interdependent causal system.

2.2 Benefits of path models for corpus linguists specifically

In the same way path models have helped other fields improve and expand their respective analytical repertoire, corpus linguistics stands to benefit greatly from the advantages of these methods as well. The possibility of investigating multiple dependent variables and mediating variables (and latent variables in the broader scope of SEM) would lead to more accurate descriptions of linguistic and contextual features and how they relate to one another. For example, we might be interested in the variables *months spent in an English-speaking country*, *frequency of filled pauses*, and *test scores on an oral proficiency test*, hypothesizing that *frequency of filled pauses* serves as a mediator between the other two variables (i.e., *months spent in an English-speaking country* could be hypothesized to influence *frequency of filled pauses*, which in turn might influence *test scores on an oral proficiency test*). Indeed, for corpus linguists specifically, we would submit that the advantages of SEM techniques go beyond the advantages outlined in Section 2.1: If we take these methods onboard, it could encourage additional, more far-reaching changes for the field at large.

For example, at a general level, wishing to apply these methods could serve as an incentive to move more toward theory and hypothesis-driven research, thus leading to an increased focus on cumulative knowledge building in the field.[1] As mentioned in Section 2.1, the direction of the causal relations within path models have to be supported by theory and/or findings of previous studies. Path models are not intended to be an exploratory data analysis procedure.[2] While several competing models can be compared in what may seem an exploratory manner, the

1 It should be noted right from the start, however, that our intention with this article is not to introduce new techniques and encourage their "tail wagging the dog"-type of use. On the contrary, we believe that moving toward cumulative knowledge building and the big-picture thinking that goes along with model-based reasoning would greatly benefit the field. The fact that doing so would also allow us to use SEM techniques to answer questions that commonly used methods in corpus linguistics have rendered unanswerable is merely an added bonus.

2 There are, nonetheless, techniques that fall under the SEM umbrella that are often practiced in a more exploratory manner, such as mixture models in which numbers of latent classes might be explored.

competing models are always proposed on theoretical grounds. What follows from this is that in order for us to be able to use path models on our data, our study has to explicitly follow on other, similar studies in the field and/or relevant theory. This means, for example, that if we want to model a causal relation between *register* and *frequency of extraposed clauses per text* in a path model, we must have reason to believe based on strong theory and/or previous studies that there is such a mechanism between the two such that register (e.g., academic writing vs. news) affects the frequency of extraposed clauses per text.

Due perhaps to the relatively short history of our field and recent technical advancements that have made massive amounts of data available to us, a large proportion of corpus linguistics research has been exploratory in nature (Biber 2020), at times without particularly strong ties to theory or findings of previous studies. While exploratory work serves several important functions, it is our conviction that we have accumulated enough knowledge on several topics that we can move further toward theory and hypothesis-driven research and cumulative knowledge building where we use the findings of previous studies to form testable, detailed hypotheses about our topic.[3]

Additionally, in encouraging broad, model-based reasoning, SEM techniques allow corpus linguists to be less dependent on the more narrow NHST paradigm. As discussed in the previous subsection, the NHST epistemological tradition has been heavily criticized on general accounts. For corpus data specifically, if *p*-values for individual variables are our primary tool for model selection, a quest for statistical significance is especially problematic given the fact that corpus linguists tend to work with samples that often are large enough to be *overpowered*, meaning that almost any measurable difference will come out as statistically significant, even if the effect size is small enough that the difference or relation completely lacks practical significance. Put differently, given a non-null difference/relation, a large enough sample will always yield statistically significant results (Kilgarriff 2005; see also Gries 2005 for a discussion of effect size in corpus linguistics). Thus, if we rely solely on statistical significance to help us decide which variables are useful predictors as part of the model-selection process for a regression model, a larger or smaller data set would likely result in different conclusions being drawn based on our results.

3 However, it is worth noting that in order to be able to move in this direction, some general, additional changes might be needed, including rigorous and transparent reporting practices (e.g., Paquot and Plonsky 2017) and more unity with regard to definitions and operationalizations of key concepts and constructs (e.g., Gries 2008), both of which are necessary if we are to be able to build on other researchers' work in a systematic manner. Increasing options for coding schemes, code, and even data that have been coded for linguistic features to be made available online through initiatives such as the IRIS database (www.iris-database.org) are likely to facilitate this process.

While path models do not solve the problem of overpowered samples *per se*, the problem is greatly ameliorated in that they enable us to move away from heavy reliance on the NHST paradigm toward more global, model-based reasoning.[4] That is, instead of focusing on rejecting the *null hypothesis*, which we know we are likely to be able to do given a large enough sample and which we almost always know to be false, we focus on the model as a system, and we assess whether or not to reject or retain *this particular model*—as a whole or in part.

With this conceptual groundwork in place, and given the advantages of path models in general and to corpus linguists specifically, we will move on to a more detailed introduction to path models.

3 An introduction to path models

Path models are related to multiple regression models in that they are both linear models relating independent and dependent variables. Nonetheless, there are crucial differences; we will here give a more in-depth treatment to two of these key differences, by way of introducing path models and their use. First, unlike multiple regression, path models allow for investigations of hypothesized causal links[5] within complex variable structures in models with multiple dependent variables; second, under the right circumstances, path models are rejectable (Hancock and Schoonen 2015: 163). These features of path models will be addressed and illustrated in turn below. Before doing so, however, some general concepts and terminology will be introduced.

To enable us to represent causal links among variables, path diagrams are used. These serve as a pictorial depiction of the theoretical part of the model. Some standard annotation conventions used for path diagrams can be found in Figure 1. As is shown, observed variables are enclosed by rectangles, while latent variables (i.e., constructs that cannot be measured directly) are enclosed by ovals; latent variables will be discussed in Section 4.3. Direct effects (i.e., the direct influences of one variable on

4 In fact, in path models, larger samples can even be advantageous in that they not only enable us to model and detect very fine-grained relations between variables that would not have been noticeable in smaller samples, but also allow us carry out more complex moderation and multigroup models.

5 Regression technically is about prediction irrespective of causal mechanism, specifying a conditional mean of the outcome based on the predictors. Path models, by contrast, are rooted in beliefs about causal processes, specifically that each independent variable has a direct causal bearing on the outcome. Regression can be, and has been, used for such models, but it is typically restricted to only single-dependent-variable models (see Breiman and Friedman 1997; Variath and Brobbey 2020 for exceptions).

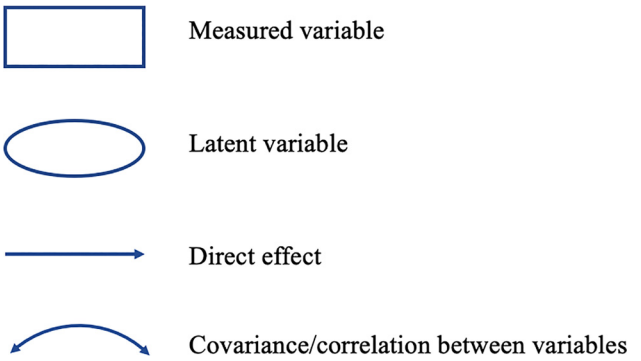

Figure 1: Examples of path model diagram symbols.

another) are indicated by a unidirectional arrow, where the direction of the arrow specifies the direction of the hypothesized causality. Double-headed arrows (often curved) denote covariance/correlation between two independent variables[6] typically arising due to one or more mutual causes outside the scope of the model.

A term often used for the process of postulating hypothesized links among variables is *model specification*. The process, which is an essential part of SEM modeling, involves the application of previous research and theory to justify the variables and their connections in a theoretical path model. As pointed out by Schumacker and Lomax (2016: 71), "[p]ath analysis does not provide a way to specify the model, but rather estimates the effects of the variables once the model has been specified by the researcher on the basis of theoretical considerations". That is, path modeling is not in and of itself a causal modeling technique; rather, it is a method used to test "theoretical models that depict relations amongst variables" (Schumacker and Lomax 2016: 69). In order to be able to draw conclusions about causality, the following four conditions have to be met by the variables and their relations specified in the model: (i) there is temporal ordering of variables; (ii) there is covariation or correlation among variables; (iii) other possible causes have been controlled for; and (iv) if X is manipulated, it causes a change in Y (Schumacker and Lomax 2016: 69).

By way of illustration and as a way of bridging the conceptual gap between multiple regression and path models, Figure 2 shows what a multiple regression model would look like in a path diagram, assuming the aforementioned criteria for causality are in fact met. The variables come from Kyle and Crossley's (2018) study of writing quality as operationalized by the dependent variable *holistic TOEFL*

6 Or between two dependent variables' residual/error terms, each of which represents external influences on a dependent variable other than those explicitly contained within the model.

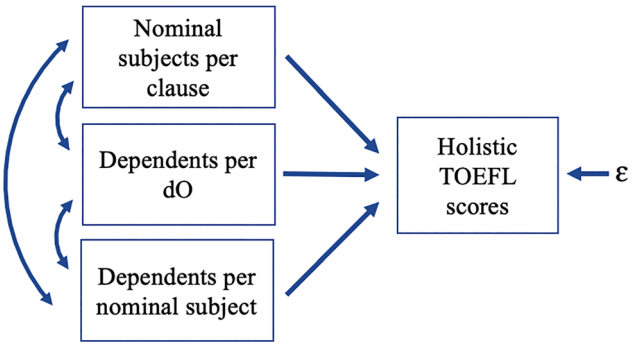

Figure 2: Multiple regression as a path diagram (variables from Kyle and Crossley 2018).

scores; the authors consider more independent variables in their regression model than what are listed here, but the ones included here are some of the ones found to be of key importance, namely *nominal subjects per clause*, *dependents per direct object*, and *dependents per nominal subject*. The epsilon term to the right of the model collectively represents all other elements having a causal bearing on the dependent variable that are independent of the three depicted causal variables.

With these concepts introduced, the first difference between path models and multiple regression that we will discuss here is that path models allow for different configurations of multiple dependent variables, as they include several regression equations (Schumacker and Lomax 2016: 69), thus providing more flexibility in terms of hypothesized variable relations than multiple regression. For example, Figure 3 displays models with three variables in different configurations.

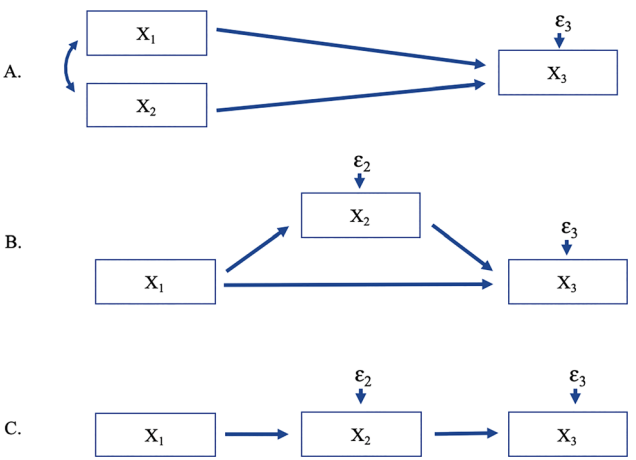

Figure 3: Some possible three-variable models (adapted from Schumacker and Lomax 2016: 72).

In model A, X_1 and X_2 are hypothesized to covary (for reasons outside the model) and to each have a direct effect on X_3. In model B, X_1 has a total effect on X_3 that includes both a direct effect and an indirect effect, with the latter mediated by X_2. In model C, however, the effect of X_1 on X_3 is hypothesized to be completely mediated by X_2; there is thus no additional direct effect of X_1 on X_3. We can also note that path models A and B are *just-identified*, meaning that the variables are all linked to one another with either single- or double-headed arrows, thus leaving a model with zero degrees of freedom. Such models will always have perfect fit, that is, be able to find best-fit values for its parameters that perfectly reproduce the variances and covariances observed in the sample of data. Model C, by contrast, is not just-identified, as there is no direct path from X_1 to X_3. This is an *over-identified* model, which means that it is likely the case that the best-fit parameters will not perfectly reproduce the sample information. For these kinds of models, one or more restrictions have been imposed, meaning that they enable us to test a particular hypothesis resulting from previous research and/or theory. That is, in imposing a restriction, we can assess the model's fit to the data, in order to determine whether we should retain or reject this model as a reasonable explanatory system for these variables.

This leads us to the second key difference between path models and multiple regression to be discussed here, namely that (over-identified) path models are rejectable: "[I]f the pattern of relations among the measured variables is sufficiently inconsistent with the hypothesized connections specified in the model, then the model as a whole may be rejected as an explanatory system" (Hancock and Schoonen 2015: 163). That is, if the data are sufficiently inconsistent with the model, the model as a whole can be deemed implausible enough that it should not be retained. For corpus linguists, this would enable us to (also) view models as systematic wholes.[7]

To know whether to accept or reject a hypothesized model, data-model fit indices are examined; these indices should also be reported in the study. As will be discussed further in Section 4.2 below, it is generally recommended that researchers check and report indices of different types: absolute (e.g., standardized root mean-square residual, SRMR), parsimonious (e.g., root mean-square error of approximation, RMSEA), and incremental (e.g., comparative fit index, CFI; see, e.g., Hu and Bentler 1999; Kline 2005 for a more detailed account of model fit). Many researchers also use and report a chi-square value for the model as a whole; however, this test is very conservative (Hancock and Schoonen 2015: 176) and

7 Nonetheless, assuming the path models are deemed acceptable, they detail the coefficients for each variable, which allows for conclusions about the relative importance/impact of individual variables to be drawn.

perhaps more useful in order to compare models differing in one or more parameters (e.g., models B and C in Figure 3) (e.g., Schoonen et al. 2011). An example of how these indices can be used to assess the model fit is provided in the next section, where we will use empirical data in a worked example in an attempt to move away from the more abstract realm toward the concrete.

4 A worked example using empirical data

As mentioned above, in order to be able to make causal inferences from the path model, we have to make sure that it is specified based on previous research and/or theory. At times, previous research may offer somewhat different—and even contradictory—structures that could each be modeled. A common next step in the SEM framework for such cases is to fit competing models and then compare them using fit indices; this is illustrated in Section 4.1. Section 4.2 illustrates how previous research may lead us to specify such competing models, the fit of which will subsequently be tested and discussed. Finally, Section 4.3 broadens the scope and offers some suggestions for other, related techniques that readers interested in expanding their repertoire might want to look into: confirmatory factor analysis and latent variable path analysis.

4.1 Specifying the models

The current section presents the results of a reanalysis of some data from a study of syntactic complexity (Larsson and Kaatari 2020) to illustrate, through a worked example, how path models can be used for corpus linguistic inquiries. As discussed above, while path models can be used to model fairly large systems of variables, we will start here with some relatively simple models in order to hit some key points without adding unnecessary complication to this introductory account of path models.

Syntactic complexity (i.e., the grammatical sophistication exhibited in language production) has been studied extensively in recent years, using other methods such as multiple regression, multidimensional analysis, and random forests. For the purpose of the present analysis, we will use path models to take a closer look at noun phrase (NP) complexity. Specifically, we are interested in the hypothesized effect that register (in our case, academic prose vs. popular science) and disciplinary group (here, social sciences vs. natural sciences) have on different measures of NP complexity. The corpus data come from BNC-15 (Kaatari 2017), which is a carefully sampled subcorpus of the *British National Corpus* (BNC)

(Burnard 2007). The subset used for the present study comprises 20 texts from each register and disciplinary group; each text is sampled to be approximately 10,000 words, making the total size of our sample 80 texts and just over 800,000 words.[8] The three measures of syntactic complexity were extracted using the *Tool for the Automatic Analysis of Syntactic Sophistication and Complexity* (TAASSC; Kyle 2016), and the output is provided in mean frequency of each measure per NP.

We know from previous studies that certain syntactic features vary with register (e.g., Biber and Gray 2016; Biber et al. 2020). That is, we have good reason to believe that there is a causal relation between register and syntactic complexity, such that the situational circumstances under which a text was produced (e.g., expected expertise of the audience) affect the written output. In particular, measures of NP complexity have been found to clearly differ across register (Biber and Gray 2010; Biber et al. 2020; Larsson and Kaatari 2020). Biber et al. (2020, based on Biber and Gray 2010) highlight three types of NP modification that have all been found to be more frequent in formal registers: adjectival, prepositional, and nominal modification. We will therefore focus on these three types in our study. Example (1) illustrates adjectival and prepositional modification; the NPs are in square brackets, the attributive adjectives are bolded, and the (nested) prepositional phrases are underlined. Nominal modification is exemplified in Example (2) (bolded and underlined along with the head noun); the postmodifying prepositional phrase is underlined.

(1) [the **main** difficulty] is [the **complete** lack of [any defence of [the conclusion]] (ACAD_SS.CMN.sampled.txt)
(2) [The function of [this behaviour]] is probably [a **defence system**] (ACAD_NS.FU0.sampled.txt)

Based on the aforementioned findings, we can hypothesize that register has a causal effect on all three of these types of modification.

However, in addition to register, differences regarding the patterning of syntactic features have also been found across disciplinary groups (e.g., social sciences vs. natural sciences; Staples et al. 2016). Writing differs across discipline in terms of both style and conventions. Therefore, it is not unreasonable to expect that the disciplinary group also might have an impact on syntactic complexity, such that the conventions of the field in which the text was produced affect the written output of its scholars. In particular, the use of nominal and adjectival

8 The structure of these data is hierarchical/nested, as is that of data used in virtually all corpus linguistics studies (Gries 2015b). Although design-based corrections to parameter standard errors exist within the path analysis framework, these will be ignored here for simplicity of illustration. For more information, the reader is referred to Stapleton (2013).

modification (but less so prepositional modification) has been found to differ across disciplines (Staples et al. 2016). Based on this, we hypothesize that disciplinary conventions affect adjectival and nominal modification, but not prepositional modification (our hypothesis is thus that there is no direct path between discipline and prepositional modification).

So far, we have a model with two independent variables (henceforth referred to as *register* and *discipline*) and three dependent variables (nominal, adjectival, and prepositional modification) that is built to test our *a priori* beliefs that both register and discipline are part of an explanatory system that can help us understand changes in the three measures of complexity investigated.[9] We will now turn to the question of how (or whether) the independent variables, as well as the dependent variables, relate among themselves.

To start, due to the completely balanced manner in which the sampling of register and discipline was conducted, there is no reason that our independent variables should covary; therefore, we are constraining their covariance to zero. Further, even if sampling of texts were not conducted in a balanced manner, we would have no theoretical reason to suspect that register and discipline covary, and would have likewise constrained their relation to zero. Having the option of choosing whether or not independent variables should be allowed to covary is an important strength of path models relative to multiple regression, the latter of which offers no such mechanism to impose one's theoretical beliefs.

With regard to the dependent variables, we also have complete control within the model, all the way down to the behavior of their residual error terms. To elaborate briefly, without any changes to the variable structure, our theoretical speculations above imply that the three measures of syntactic complexity should covary among themselves because, and only because, they share the common causal forces of register and/or discipline. However, to make a decision about this part of the model, we have to consider whether there are other reasons that these three variables should covary above and beyond their two mutual causal inputs in the model (i.e., whether there are other shared influential forces outside the model).

Here, we have two competing hypotheses. On the one hand, all three dependent variables are measures of NP complexity (the higher the value, the more complex the NP); as such, all three contribute to textual density and can be expected to be found, for example, in texts with set word limits. This suggests that some of the causal forces outside the model that are relegated to those dependent variables' error terms may, in fact, be shared among the variables, which means

9 Although path models can be used for investigations of moderating and/or mediating relations between variables (cf. Section 2.1), our hypotheses do not predict any such relations, and no relations of this kind are modeled in this article.

Table 1: Descriptive statistics for the frequency of the measures per NP and text.

	Adjectival modification	Prepositional modification	Nominal modification
Min; Max	0.13; 0.44	0.11; 0.34	0.03; 0.35
Mean (SD)	0.27 (0.06)	0.24 (0.05)	0.12 (0.07)
Median (IQR)	0.27 (0.08)	0.24 (0.06)	0.11 (0.09)

that those error terms would covary (and should be allowed to do so in our model). On the other hand, while the measures have been found to covary in different registers and disciplines, it is not a given that they will do so outside those contexts. That is, just because writers make frequent use of adjectival modification, it does not automatically mean that they will use nominal modification. This suggests that the common causal forces of register and/or discipline alone are the key reasons these NP complexity measures are likely to covary, and as such the error terms would not covary (and we should not allow them to do so in our model). Thus, these competing explanations lead us to two different hypothesized variable structures, illustrated in Figures 4 and 5, each of which will be tested in a separate model. Some descriptive statistics for the three measures are shown in Table 1.

The proposed models enable us to (i) evaluate the overall consistency of each with the data, (ii) compare the models' relative fit to each other to choose between

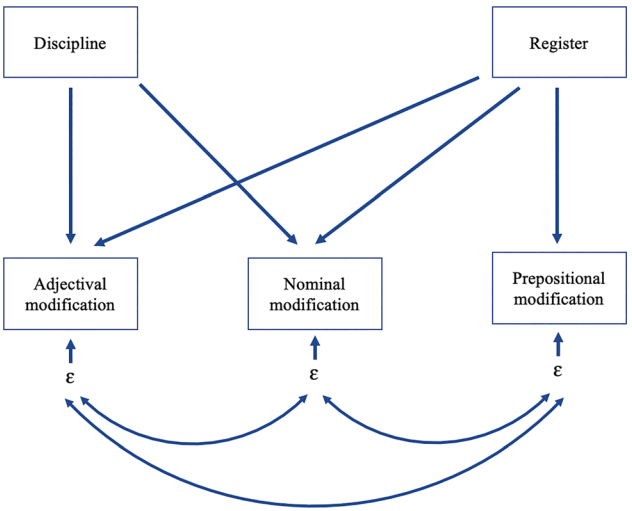

Figure 4: Path diagram for a Model 1.

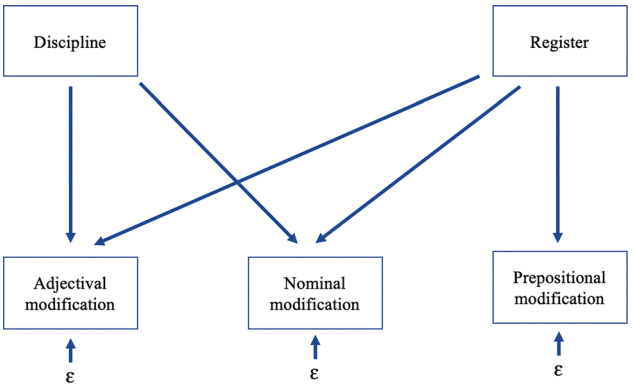

Figure 5: Path diagram for a Model 2.

them, and in the model selected thus (iii) test the relative impact of register and discipline on three measures of NP complexity. The following research questions were used to guide the analysis:

- Can the hypothesized relations among the measures of complexity be explained solely by register and discipline or are there external factors at play that cause these to covary? That is, what are the relative strengths of the two competing models?
- What is the relative importance of register and discipline on the measures of NP complexity?

4.2 Comparing the models and interpreting the results

We used *R* (R Core Team 2020) and the package *lavaan* (version 0.6-7; Rosseel 2012) to fit the models using maximum likelihood estimation. The code and full model output can be found in the Appendix. Table 2 summarizes the fit indices most often reported for SEM models, and what some of the recommended ranges are for acceptable fit.[10] The fit indices for the two models from our analysis are summarized in Table 3. It is worth noting, however, that the recommended values for good fit are not to be seen as cutoff points (the way *p*-values are commonly used in a

10 The values come primarily from work by Hu and Bentler (1999), which are understood to be rough guidelines and in no way universally applicable standards.

Table 2: Fit indices and their recommended values for acceptable fit.

Test	Reports:	Guidelines
χ^2 (Chi square)	A statistical test of the overall fit and discrepancy between the hypothesized model and the data. H_0: the model fits perfectly. Sensitive to sample size.	$p > 0.05$ ideally, but because models are expected to contain trivial misspecifications that become significant with increased sample size, this is typically ignored.
AIC (Akaike Information Criterion)	A measure incorporating fit and parsimony.	Lower values indicate a better model fit. Useful for comparing models, not assessing individual models.
CFI (Comparative Fit Index)	"absolute or parsimonious fit relative to a baseline model, usually the null/independence model that specifies no relations among observed variables" (Hancock and Schoonen 2015: 176).	>0.95 (>0.90 historically)
RMSEA (Root Mean Square Error of Approximation)	"the overall discrepancy between observed and implied covariance matrices while taking into account a model's complexity; fit improves as more parameters are added to the model as long as those parameters are making a useful contribution" (Hancock and Schoonen 2015: 176).	<0.06(<0.08 historically)
SRMR (Standardized Root Mean Square Residuals)	"the overall [standardized] discrepancy between an observed covariance matrix and the covariance matrix suggested by the parameter estimates from the hypothesized model specification; fit improves as more parameters are added to the model" (Hancock and Schoonen 2015: 176)	<0.08 (<0.05 historically)

NHST framework), but rather as recommendations to help researchers evaluate their models. This is especially important to keep in mind in cases where the model fit values provide somewhat contradictory recommendations (see also see McNeish and Wolf in press for a dynamic model of model fit).

Table 3: Fit indices for Models 1 and 2 in the present study.

Test	Model 1	Model 2
χ^2 (*df*)	1.446 (2 *df*); $p = 0.485$	30.605 (5 *df*); $p = 0.000$
AIC	−529.965	−506.806
CFI	1.00	0.70
RMSEA [90% CI]	0.000 [0.000; 0.201]	0.253 [0.171; 0.343]
SRMR	0.037	0.114

The fit indices for our models unanimously suggest that Model 1 is a better fit for our data: the chi-square, CFI, RMSEA, and SRMR all indicate a good fit for Model 1, but not for Model 2, and the AIC is lower for Model 1 than for Model 2.[11] A chi-square difference test confirms that the models are significantly different ($p < 0.001$, chi-square difference 29.159 with three *df*). We will therefore reject Model 2 and retain Model 1, and move to interpreting the results contained therein.

The model output (see the Appendix; summarized in Figure 6) provides the kind of information about effect sizes and *p*-values that we are accustomed to from the typical output of multiple regression.[12] For example, we can see that the difference between the registers for the nominal modifiers is not statistically significant ($p = 0.218$). Other than that, all other specified paths are significant at the 0.05 level, as marked in the figure.[13] While not pictured, the model output further provides separate R^2 values for the three dependent variables: 0.369 for adjectival modification, 0.093 for nominal modification, and 0.256 for prepositional modification.

Similarly to (unstandardized) coefficients in a multiple regression, with all other variables kept constant, the path coefficients in the present models show the predicted change when discipline changes from *natural sciences* (coded as 0) to

11 It can also be noted that the modification indices for Model 2 (see the Appendix) suggest that allowing for error covariance (in particular between ADJ and PP) would improve the model fit considerably.

12 Although not illustrated here, we can test whether the paths are different from one another either by running a competing model in which the paths are constrained to be equal and comparing the fit to our proposed model, or by creating an additional parameter in our current model whose value is set to the difference between the paths of interest and then seeing whether this parameter's estimate comes out as statistically significantly different from zero.

13 Note, however, that while *p*-values offer useful information for specific paths, any decisions pertaining to whether to reject or retain a given model are based on the other indices we report (cf. our discussion about NHST in Sections 2.1 and 2.2).

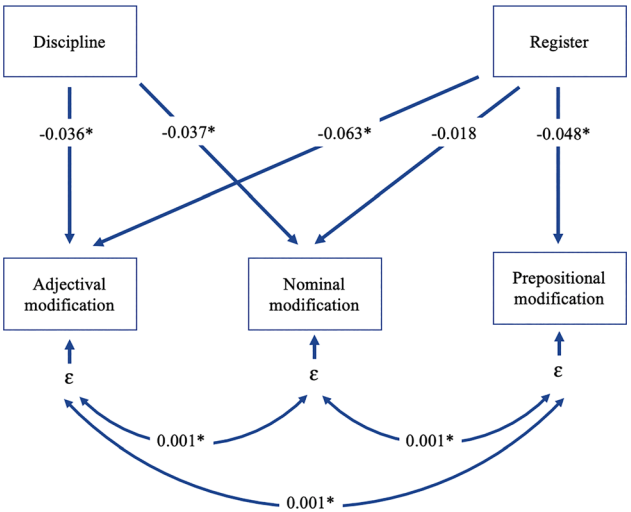

Figure 6: Path diagram for Model 1.

social sciences (coded as 1) or when register changes from *academic prose* (coded as 0) to *popular science* (coded as 1). As the path coefficients are negative, it predicts, for example, that the social science texts contain 0.036 adjectival modifiers *fewer* per NP than the natural science texts, and that the popular science texts include 0.063 adjectival modifiers *fewer* per NP than the academic prose texts, which (given the range found in the data; see Table 1) are not immaterial. As can further be seen, the path coefficients for the effects of discipline on adjectival and nominal modification are −0.036 and −0.037, respectively, indicating almost identical effect size.

Returning to our research questions, we can conclude that Model 1 fit the data better than our competing model, Model 2. For the latter model, the fit indices and the modification indices suggested that there were relations in the data that the model did not capture, whereas the model fit indices for Model 1 indicated good fit. The retained model, Model 1, showed that there was evidence to support the hypotheses based on previous studies stating that both discipline and register have an impact on NP complexity. Discipline has a statistically significant effect on both adjectival and nominal modification. Specifically, the texts from the social sciences have a lower average per NP for adjectival and nominal modification than the texts from the natural sciences. Regarding register, the less formal

popular science texts had a lower average per NP for all three measures compared to the more formal academic texts. At a more global level, we can note that register has a comparatively strong effect on adjectival and prepositional modification, and only a very minor, nonsignificant effect on nominal modification. This suggests that the nominal modification patterns in the data are better explained by discipline than by register. A closer look at the data suggests that a possible explanation for this is that nominal modification is often used for technical terms and thus a common feature of technical descriptions, which are more typical of the natural sciences than the social sciences; an example of such a use is provided in Example (3).

(3) This is particularly true if an **emission spectrum** can be observed: **absorption spectra** arise mainly from the vibrational ground-state (Acad_natural_sci.H9R.sampled.txt).

Lastly, the error covariances estimated were statistically significant. The values in Figure 6 are in an unstandardized metric, and as they carry the units of their associated variables are hard to interpret. Standardized values for these relations, that is, error correlations, are available in a standardized solution. They are 0.456 for adjectival and prepositional modification, 0.334 for adjectival and nominal modification, and 0.263 for prepositional and nominal modification. This suggests that there are causal forces outside the model that exert similar influence on all three measures, which means that the measures do not vary independently from one another. Adjectival and prepositional modification exhibited particularly similar behavior in the data.

Although this study is included for illustrative purposes, and thus not necessarily expected to yield results of particular linguistic significance, we can still take note that the weak causal effect of register on nominal modification might suggest that future studies should focus exclusively on adjectival and prepositional modification in relation to register to further investigate the ways in which these covary. Moreover, as we found evidence to support the hypothesis stating that the measures of complexity do not vary independently of one another, a natural next step would be to look more closely into what factors—linguistic or extralinguistic—other than discipline and register may have an impact on NP complexity and extend the model to include these as well. We might also want to compare other registers and/or disciplines get a more complete picture.

So far, as an introduction to SEM, we have opted for an in-depth discussion of (measured) path models, rather than a more superficial treatment of several

different techniques from the broader SEM family. Nonetheless, to help interested readers know what they might want to turn their attention to next, we will briefly introduce two techniques related to path models: *confirmatory factor analysis* and *latent path models*.[14]

4.3 Branching out: Other covariance structure models

Measured path models, confirmatory factor analysis (CFA), and latent path models are different, but interrelated, types of *covariance structure models*, designed to help us learn why and how variables relate. One important difference between path models of the kind that we have discussed above and the two other techniques is that the latter incorporate so-called *latent variables*.

Latent variables are variables hypothesized to exist but which have no observed realizations and instead need to be measured via observable variables (see, e.g., Bollen 2002). In corpus linguistics, we often talk about how we "operationalized a given variable as/through measures X, Y, and Z;" for example, "we operationalized L2 proficiency through scores from speaking, listening, reading, and writing tests." However, in many cases, such variables could instead be conceptualized as latent variables and modeled as such. Doing so is advantageous in that it reduces the dimensionality of the data and in that the resulting model captures "relations among the constructs of interest rather than among their error-prone measured indicators" (Hancock and Schoonen 2015: 165). Although not always labeled as such, latent variables are included and discussed in corpus linguistics studies employing Multidimensional Analysis (e.g., Biber 1988) which employs exploratory factor analysis (EFA) to identify potential latent variables.

EFA differs from CFA in that in EFA, the latent variables are unknown, and the factors that emerge from the analysis are believed to be the latent variables (i.e., the underlying sources of covariation among the measured variables). As the statistical analysis is not constrained by theory in EFA, it can therefore "build on chance correlations among features as well as theoretically significant correlations" (Biber 2001: 220). By contrast, the type of factor analysis that we focus on here, CFA, does not involve data exploration in search of such underlying sources.

14 While not covered in this introductory account of path models, readers familiar with mixed-effects models might be interested in also looking into multilevel path/structural models (see, e.g., Stapleton 2013 for a didactic treatment).

Instead, measured variables are selected *a priori* with the purpose of serving as indicators of a hypothesized latent variable (Bollen 2002: 624). The model is an analytical framework that is used to "gather support for, or to refute, hypothesized constructs' existence and the relations to their observed indicators" (Hancock and Schoonen 2015: 164). An example of a study that has used CFA in corpus linguistics is Biber (2001) where the patterning of linguistic markers of complexity across register was modeled.

Latent path models combine CFA's ability to model latent variables with the structural goals of measured path models. This typically yields models with hypothesized structural relations among the latent variables rather than among the measured operationalizations of those latent variables. Latent path models thus have the added benefits of

> (1) directly representing the construct relations that are typically of interest to the researcher; (2) estimating those relations without attenuation and with increased statistical power; and (3) being rejectable, that is, allowing for the assessment of consistency or inconsistency with the pattern of covariation in the data to determine whether or not the model as a whole is viable. (Hancock and Schoonen 2015: 165)

Elaborating upon point (2) in the above quote, such models allow for disattenuation; that is, they account for measurement error. It is well-known that there is error in everything we measure. Measurement error can, for example, result from imprecise measures, tagging errors, and manual coding (see Larsson et al. in press for more examples and a discussion of measurement error in Learner Corpus Research). If we do not account for the fact that measured operationalizations of our latent constructs contain error, we are forced to ignore the attenuating and often misleading effects of measurement error; this is highly problematic given that doing so ultimately leads to less precise (and sometimes even erroneous) results (Grewal et al. 2004).

Application of latent path models would be beneficial to a broad range of corpus linguistic studies. For example, it would allow researchers working on learner data to more accurately model concepts such as L2 proficiency and formality. In addition, historical linguists can use this framework to better understand social causes of language change. Moreover, researchers interested in sociolinguistic inquiries would greatly benefit from modeling causes of linguistic variation. For examples of applications of latent path models in Second Language Acquisition research, readers are referred to Hancock and Schoonen (2015).

5 Conclusion

The present article has sought to introduce one member of the SEM family thought to be particularly useful to corpus linguists, namely path models. It has outlined general advantages of path models vis-à-vis another, more commonly employed techniques (multiple regression analysis), as well as discussed benefits of moving toward model-based reasoning. A number of advantages of path models were reviewed. For example, path models enable us to draw conclusions about causality in model structures with multiple dependent and independent variables, which, whether expressed overtly in our articles or not, underlie many corpus linguistic inquiries. That is, we oftentimes have a confirmatory mindset in study designs involving multiple independent and dependent variables ("do *q* and *w* affect *x*, *y*, and *z*?"), but the commonly employed analytical tools do not allow us to directly address such questions.

Furthermore, using SEM techniques such as path models would enable a change in perspective, where we can move away from almost complete reliance on the more narrow NHST paradigm toward model-based reasoning, cumulative knowledge building and big-picture thinking, all of which seem to be, to varying degrees, missing in the field. However, that is not to say that we should *replace* careful linguistic analysis with model-based thinking; quite the opposite: these techniques should always be preceded by (and their accuracy crucially rely on) rigorous linguistic analysis and in-depth knowledge of the literature, followed by linguistic interpretation. We are in no way suggesting that SEM techniques are a solution in themselves; they merely aid the corpus analyst in understanding the linguistic phenomena of interest. Nor are we saying that we should abandon all other statistical techniques used in the field; this is not a matter of finding "one technique to rule them all."

We would posit, however, that adding SEM techniques to our toolbox would enable greater flexibility in terms of study design, thus moving the borders for what can and cannot be concluded using statistical methods on corpus data. We therefore hope to have inspired readers to start exploring the great potential path models and other members of the SEM family has for studies in corpus linguistics.

Acknowledgments: We are grateful to the editorial team and the anonymous reviewers for their helpful comments and feedback.

Appendix

R code used for the analysis, along with the model output.

```
> ## Loading lavaan package
> library(lavaan)
> ## Fitting the models
> # Model 1
>
> model.1 <-
+ 'ADJ ~ DISCIPLINE
+ NN ~ DISCIPLINE
+ ADJ ~ REGISTER
+ PP ~ REGISTER
+ NN ~ REGISTER
+ DISCIPLINE ~~ 0*REGISTER #discipline and register are not allowed to
covary
+ PP ~~ ADJ
+ NN ~~ ADJ
+ NN ~~ PP'
>
> ## Fitting the model to an object
> model.1 <- sem(model.1, data = data_SEM)
> ## Writing out the model summary and the fit measures
> summary(model.1, fit.measures = TRUE, standardized = TRUE,
rsquare = TRUE, modindices = TRUE)
> ## Model summary and fit measures:

lavaan 0.6-7 ended normally after 81 iterations

Estimator                                         ML
Optimization method                           NLMINB
Number of free parameters                         13

Number of observations                            80

Model Test User Model:

Test statistic                                 1.446
Degrees of freedom                                 2
P-value (Chi-square)                           0.485
```

```
Model Test Baseline Model:

Test statistic                                95.433
Degrees of freedom                            10
P-value                                       0.000
User Model versus Baseline Model:
Comparative Fit Index (CFI)                   1.000
Tucker-Lewis Index (TLI)                      1.032
Loglikelihood and Information Criteria:
Loglikelihood user model (H0)                 277.982
Loglikelihood unrestricted model (H1)         278.705
Akaike (AIC)                                  -529.965
Bayesian (BIC)                                -498.999
Sample-size adjusted Bayesian (BIC)           -539.992
Root Mean Square Error of Approximation:
RMSEA                                         0.000
90 Percent confidence interval - lower        0.000
90 Percent confidence interval - upper        0.201
P-value RMSEA <= 0.05  0.551

Standardized Root Mean Square Residual:

SRMR                                          0.037

Parameter Estimates:
Standard errors                             Standard
Information                                 Expected
Information saturated (h1) model          Structured

Regressions:
             Estimate   Std.Err   z-value   P(>|z|)   Std.lv   Std.all
ADJ  ~
 DISCIPLINE    -0.036     0.009    -3.824     0.000   -0.036   -0.302
NN  ~
 DISCIPLINE    -0.037     0.014    -2.674     0.007   -0.037   -0.275
ADJ  ~
 REGISTER      -0.063     0.011    -5.931     0.000   -0.063   -0.527
PP  ~
 REGISTER      -0.048     0.009    -5.252     0.000   -0.048   -0.506
NN  ~
 REGISTER      -0.018     0.014    -1.232     0.218   -0.018   -0.131
```

```
Covariances:
                Estimate   Std.Err   z-value   P(>|z|)   Std.lv   Std.all
DISCIPLINE ~~
  REGISTER         0.000                                0.000     0.000
.ADJ ~~
  .PP              0.001     0.000     3.714     0.000    0.001     0.456
  .NN              0.001     0.000     2.833     0.005    0.001     0.334
.NN ~~
  .PP              0.001     0.000     2.276     0.023    0.001     0.263

Variances:
                Estimate   Std.Err   z-value   P(>|z|)   Std.lv   Std.all
  .ADJ             0.002     0.000     6.325     0.000    0.002     0.631
  .NN              0.004     0.001     6.325     0.000    0.004     0.907
  .PP              0.002     0.000     6.325     0.000    0.002     0.744
   DISCIPLINE      0.250     0.040     6.325     0.000    0.250     1.000
   REGISTER        0.250     0.040     6.325     0.000    0.250     1.000

R-Square:
                Estimate
   ADJ              0.369
   NN               0.093
   PP               0.256

Modification Indices:

          lhs op         rhs     mi    epc  sepc.lv  sepc.all sepc.nox
6  DISCIPLINE ~~    REGISTER 0.000  0.000    0.000     0.000    0.000
19         PP ~~  DISCIPLINE 1.433  0.003    0.003     0.134    0.134
25         PP ~          ADJ 1.433 -0.301   -0.301    -0.382   -0.382
26         PP ~           NN 1.433 -0.296   -0.296    -0.420   -0.420
27         PP ~   DISCIPLINE 1.433  0.011    0.011     0.115    0.115
28 DISCIPLINE ~          ADJ 0.461  0.990    0.990     0.119    0.119
29 DISCIPLINE ~    ~      NN 1.125  3.127    3.127     0.419    0.419
30 DISCIPLINE ~           PP 1.065  1.223    1.223     0.115    0.115
31 DISCIPLINE~~    REGISTER 0.000  0.000    0.000     0.000    0.000
32   REGISTER ~          ADJ 0.000  0.000    0.000     0.000    0.000
33   REGISTER ~           NN 0.000  0.000    0.000     0.000    0.000
35   REGISTER ~   DISCIPLINE 0.000  0.000    0.000     0.000    0.000
> ## Model 2
>
```

```
> model.2 <-
+ 'ADJ  ~  DISCIPLINE
+ NN  ~  DISCIPLINE
+ ADJ  ~  REGISTER
+ PP  ~  REGISTER
+ NN  ~  REGISTER
+ DISCIPLINE ~~ 0*REGISTER
+ PP ~~ 0*ADJ
+ NN ~~ 0*ADJ
+ NN ~~ 0*PP'
>
> ## Fitting the model to an object
> model.2 <- sem(model.2, data  =  data_SEM)
> ## Writing out the model summary and the fit measures
>    summary(model.2,    fit.measures  =  TRUE,    standardized  =  TRUE,
rsquare  =  TRUE, modindices  =  TRUE)
> ## Model summary and fit measures:

lavaan 0.6-7 ended normally after 50 iterations

Estimator                                         ML
Optimization method                           NLMINB
Number of free parameters                         10
Number of observations                            80
Model Test User Model:
Test statistic                                30.605
Degrees of freedom                                 5
P-value (Chi-square)                           0.000

Model Test Baseline Model:

Test statistic                                95.433
Degrees of freedom                                10
P-value                                        0.000

User Model versus Baseline Model:

Comparative Fit Index (CFI)                    0.700
Tucker-Lewis Index (TLI)                       0.401
Loglikelihood and Information Criteria:
Loglikelihood user model (H0)               263.403
Loglikelihood unrestricted model (H1)       278.705
Akaike (AIC)                               -506.806
```

```
Bayesian (BIC)                                -482.985
Sample-size adjusted Bayesian (BIC)           -514.519
Root Mean Square Error of Approximation:
RMSEA                                          0.253
90 Percent confidence interval - lower         0.171
90 Percent confidence interval - upper         0.343
P-value RMSEA ≤ 0.05                           0.000
Standardized Root Mean Square Residual:
SRMR                                           0.114
Parameter Estimates:
Standard errors                             Standard
Information                                 Expected
Information saturated (h1) model          Structured
Regressions:
             Estimate   Std.Err   z-value   P(>|z|)   Std.lv   Std.all
ADJ ~
 DISCIPLINE    -0.030     0.011    -2.861     0.004   -0.030    -0.257
NN ~
 DISCIPLINE    -0.032     0.014    -2.266     0.023   -0.032    -0.243
ADJ ~
 REGISTER      -0.063     0.011    -5.942     0.000   -0.063    -0.535
PP ~
 REGISTER      -0.048     0.009    -5.252     0.000   -0.048    -0.506
NN ~
 REGISTER      -0.018     0.014    -1.232     0.218   -0.018    -0.132
Covariances:
             Estimate   Std.Err   z-value   P(>|z|)   Std.lv   Std.all
DISCIPLINE ~~
 REGISTER       0.000                                0.000     0.000
.ADJ ~~
 .PP            0.000                                0.000     0.000
 .NN            0.000                                0.000     0.000
.NN ~~
 .PP            0.000                                0.000     0.000
Variances:
             Estimate   Std.Err   z-value   P(>|z|)   Std.lv   Std.all
 .ADJ           0.002     0.000     6.325     0.000    0.002    0.648
 .NN            0.004     0.001     6.325     0.000    0.004    0.923
 .PP            0.002     0.000     6.325     0.000    0.002    0.744
  DISCIPLINE    0.250     0.040     6.325     0.000    0.250    1.000
```

```
  REGISTER       0.250     0.040     6.325     0.000     0.250    1.000
R-Square:
             Estimate
  ADJ           0.352
  NN            0.077
  PP            0.256
Modification Indices:
          lhs op       rhs       mi    epc   sepc.lv   sepc.all   sepc.nox
6  DISCIPLINE ~~   REGISTER  0.000 0.000    0.000      0.000      0.000
7         ADJ ~~         PP 16.136 0.001    0.001      0.449      0.449
8         ADJ ~~         NN  8.852 0.001    0.001      0.333      0.333
9          NN ~~         PP  5.350 0.001    0.001      0.259      0.259
19         PP ~~ DISCIPLINE  1.433 0.003    0.003      0.134      0.134
21        ADJ ~          NN  8.852 0.248    0.248      0.279      0.279
22        ADJ ~          PP 16.136 0.525    0.525      0.419      0.419
23         NN ~         ADJ  8.852 0.447    0.447      0.397      0.397
24         NN ~          PP  5.350 0.406    0.406      0.288      0.288
25         PP ~         ADJ 11.981 0.315    0.315      0.395      0.395
26         PP ~          NN  3.796 0.135    0.135      0.189      0.189
27         PP ~  DISCIPLINE  1.433  0.011    0.011      0.115      0.115
28 DISCIPLINE ~         ADJ  0.000  0.000    0.000      0.000      0.000
29 DISCIPLINE ~          NN  0.000  0.000    0.000      0.000      0.000
30 DISCIPLINE ~          PP  1.065  1.223    1.223      0.115      0.115
31 DISCIPLINE ~    REGISTER  0.000  0.000    0.000      0.000      0.000
32   REGISTER ~         ADJ  0.000  0.000    0.000      0.000      0.000
33   REGISTER ~          NN  0.000  0.000    0.000      0.000      0.000
35   REGISTER ~  DISCIPLINE  0.000  0.000    0.000      0.000      0.000
```

```
> ## Model comparison
> anova(model.1, model.2)
Chi-Squared Difference Test
   Df  AIC  BIC Chisq Chisq diff Df diff Pr(>Chisq)
model.1 2 -529.96-499.00 1.4457
model.2 5 -506.81-482.99 30.6049  29.159  3 2.073e-06 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2001. On the complexity of discourse complexity: A multi-dimensional analysis. In Biber Douglas & Susan Conrad (eds.), *Variation in English: Multi-dimensional studies*, 215–240. Harlow: Longman.

Biber, Douglas. 2020. Inspecting the foundation of corpus linguistic research to build for the next generation: Forward to the past. In *Plenary talk presented at the ICAME conference [online], 21 May, 2020*.

Biber, Douglas & Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9. 2–20.

Biber, Douglas & Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Bethany Gray, Shelley Staples & Jesse Egbert. 2020. Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *International Journal of Academic Purposes* 46. https://doi.org/10.1016/j.jeap.2020.100869.

Bollen, Kenneth. 2002. Latent variables in psychology and the social sciences. *Annual Review of Psychology* 53. 605–634.

Breiman, Leo & Jerome H. Friedman. 1997. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society* 59(1). 3–54.

Burnard, Lou. 2007. Reference guide for the British national corpus (XML edition). Available at: www.natcorp.ox.ac.uk/docs/URG/.

Egbert, Jesse, Tove Larsson & Douglas Biber. 2020. *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge: Cambridge University Press.

Fong, Cathy Y.-C. & Connie S.-H. Ho. 2017. What are the contributing cognitive-linguistic skills for early Chinese listening comprehension? *Learning and Individual Differences* 59. 78–85.

Grewal, Rajdeep, Joseph A. Cote & Hans Baumgartner. 2004. Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Marketing Science* 23(4). 519–529.

Gries, Stefan Th. 2003. Grammatical variation in English: A question of 'structure vs. function'? In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 155–173. Berlin/New York: Mouton de Gruyter.

Gries, Stefan Th. 2005. Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2). 277–294.

Gries, Stefan Th. 2008. Phraseology and linguistic theory: A brief survey. In Sylviane Granger & Fanny Meunier (eds.), *Phraseology: An interdisciplinary perspective*, 3–25. Amsterdam: John Benjamins.

Gries, Stefan Th. 2015a. Quantitative designs and statistical techniques. In Biber Douglas & Randi Reppen (eds.), *The Cambridge handbook of English corpus linguistics*, 50–71. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2015b. The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1). 95–125.

Hancock, Gregory R. & Rob Schoonen. 2015. Structural equation modeling: Possibilities for language learning researchers. *Language Learning* 65(Supp. 1). 160–184.

Hu, Li-Tze & Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling* 6(1). 1–55.

Hu, Xianyao, Richard Xiao & Andrew Hardie. 2019. How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis. *Corpus Linguistics and Linguistic Theory* 15(2). 347–382.

Kaatari, Henrik. 2017. *Adjectives complemented by that or to-clauses: Exploring semantico-syntactic relationships and genre variation*. Uppsala, Sweden: Uppsala University Unpublished Doctoral Dissertation.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–275.

Kline, Rex B. 2005. Principles and practice of structural equation modeling, 2nd ed. New York: Guilford.

Koplenig, Alexander. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory* 15(2). 321–346.

Kyle, Kristoffer. 2016. *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* Doctoral Dissertation. Available at: http://scholarworks.gsu.edu/alesl_diss/35.

Kyle, Kristoffer & Scott Crossley. 2018. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal* 102(2). 333–349.

Larsson, Tove & Henrik Kaatari. 2020. Syntactic complexity across registers: Investigating (in) formality in student writing. *Journal of English for Academic Purposes* 45. https://doi.org/10.1016/j.jeap.2020.100850.

Larsson, Tove, Egbert Jesse & Douglas Biber. *On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective*. under review.

Larsson, Tove, Magali Paquot & Luke Plonsky. Inter-rater reliability in learner corpus research: Insights from a collaborative study on adverb placement. *International Journal of Learner Corpus Research* 6(2). 237–251, in press.

Lee, Peter M. 2012. Bayesian statistics: An introduction, 4 ed. Chichester: Wiley.

Levy, Roy & Jaehwa Choi. 2013. Bayesian structural equation modeling. In Gregory R. Hancock & Ralph O. Mueller (eds.), *Structural equation modeling: A second course*, 2 ed, 563–623. Charlotte: IAP Information Age Publishing.

Levy, Roy & Robert J. Mislevy. 2016. *Bayesian psychometric modeling*. Boca Raton: Taylor & Francis Group.

McNeish, Daniel & Melissa G. Wolf. Dynamic fit index cutoffs for Confirmatory Factor Analysis models. Preprint Available at: https://psyarxiv.com/v8yru, in press.

Paquot, Magali & Luke Plonsky. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3. 61–94.

Pearl, Judea. 2012. The causal foundations of structural equation modeling. In Rick H. Hoyle (ed.), *Handbook of structural equation modeling*, 68–91. New York: The Guilford Press.

Plonsky, Luke. 2015. Statistical power, *p* values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In Luke Plonsky (ed.), *Advancing quantitative methods in second language research*, 23–45. New York: Routledge.

R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rodgers, Joe L. 2010. The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist* 65. 1–12.

Rosseel, Yves. 2012. lavaan: An R package for structural equation modeling. *Journal of Statistical Software* 48(2). 1–36.

Schoonen, Rob, Amos Van Gelderen, Reinoud Stoel, Hulstijn Jan & Kees De Glopper. 2011. Modeling the development of L1 and EFL writing proficiency of secondary-school students. *Language Learning* 61. 31–79.

Schumacker, Randall E. & Richard G. Lomax. 2016. *A beginner's guide to structural equation modeling*, 4th edition. New York: Routledge.

Staples, Shelley, Jesse Egbert, Douglas Biber & Bethany Gray. 2016. Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication* 33. 149–183.

Stapleton, Laura. M. 2013. Multilevel structural equation modeling with complex sample data. In Gregory R. Hancock & Ralph O. Mueller (eds.), *Quantitative methods in education and the behavioral sciences: Issues, research, and teaching. Structural equation modeling: A second course*, 521–562. Charlotte: IAP Information Age Publishing.

Variyath, Asokan M. & Anita Brobbey. 2020. Variable selection in multivariate multiple regression. *PloS One* 15(7). e0236067.

Wallis, Sean. 2020. *Statistics in corpus linguistics: A new approach*. New York: Routledge.