So many variables, but what causes what?

Cécile De Cat (University of Leeds & UiT Arctic University of Norway)
Sharon Unsworth (Radboud University, Nijmegen)

Paradis' keynote article provides a comprehensive overview of factors influencing bilingual children's dual language abilities. It includes the 'usual suspects', such as input quantity, and also highlights areas requiring further investigation, such as cognitive abilities. As such, it will no doubt serve as a valuable basis for the field as we move forward. Paradis quite rightly points out that whilst some of these factors may be unidirectionally related to language abilities, suggesting causality, for many others such relations are bi- or multidirectional and as such, caution is required in interpreting them. In order to pinpoint the nature and direction of these relations (currently absent from Figure 1 in the keynote), more complex analytic techniques are needed, as Paradis herself notes (p. 34). In this commentary, we provide an illustration of how the complex relationships between the variables discussed in Paradis's keynote article could be conceptualised within a causal inference approach. We offer a modest starting point by summarising key features of causal inference modelling and by illustrating how it might help us better understand what causes what.

Causal inference methods: a quick introduction

Causal inference aims to understand how one variable might influence another, given what is known (or suspected) about the relationships between these two and other relevant variables (Hernán, Hsu & Healy, 2019). It differs from predictive modelling in important ways (Arnold, Davies, de Kamps, Tennant, Mbotwa & Gilthorpe, 2020).

Predictive modelling, which is standardly used in language acquisition research, consists of building a model with the highest generalisability from the smallest subset of predictor variables in relation to an outcome variable. It seeks to maximise the variance explained by the model, while avoiding over-fitting the data. Crucially, the significant associations revealed by the model cannot be interpreted *causally:* the coefficients of the individual predictor variables do not reflect their unique relationship or relative importance to the outcome, since both their magnitude and sign depend on the overall set of predictors included in the model. The coefficients in a prediction model are, therefore, *individually* uninterpretable and should not be used to infer any associations, causal or otherwise (Westreich & Greenland, 2013). A common error when interpreting predictive models (which we too have committed in our own work), is to interpret all coefficients in the model in the same way *because* they are mutually adjusted. This is in fact inaccurate, as we will explain after having introduced causal inferences.

By contrast, quantitative causal inference seeks to explain the effect of an *exposure* on an *outcome*, by building a model to calculate the *estimate*, which is the effect of the exposure on the outcome in a dataset. An essential step prior to modelling is clarifying the data-generation process. This is done using external knowledge and hypotheses. We illustrate this below with concrete examples taken from Paradis' keynote, but in order to do so, we first need to introduce some key concepts in this approach. The role of the covariates included in the model is to adjust for any variable that could confound the relationship of interest (between the *exposure* and the *outcome*). Other variables have to be excluded from the model to avoid creating spurious associations, which in turn would bias the coefficient estimates. In other words, the set of variables included in the model solely aims at blocking the effect of confounders, which would otherwise distort the causal relationship under investigation. Including the appropriate set of covariates ensures *conditional exchangeability*, i.e., the ability of the model to compare the actual (observed) outcome with that of the *counterfactual*. For example, the counterfactual could be the outcome for a group of participants that would not have been affected by the *exposure*. If a variable affects both the *exposure* and the *outcome*, it is a *confounder*. If a variable is affected by the *exposure* and also affects the *outcome*, it is a *mediator*. This configuration of variables is illustrated in Figure 1. Concrete examples of confounders and mediators will be provided below in Figures 2 and 3. In Figure 2, the Diversity of HL interlocutor is a confounder in the relationship between the exposure (i.e., Amount of HL exposure in the home) and the outcome (i.e., Child's HL proficiency). In Figure 3, Media engagement in HL is a mediator in the relationship between the exposure (i.e., Family language policy) and the outcome (i.e., Child's HL proficiency).. Variables which have a direct path to the outcome or to the exposure are known as *ancestors*. Crucially, one and the same variable may fulfil any of these functions depending on the specifics of the model in question. *Confounders* have to be included in the model. *Mediators* must not, as they contribute to the causal path between *exposure* and *outcome*. If a *mediator* were included in the model, this itself would create bias

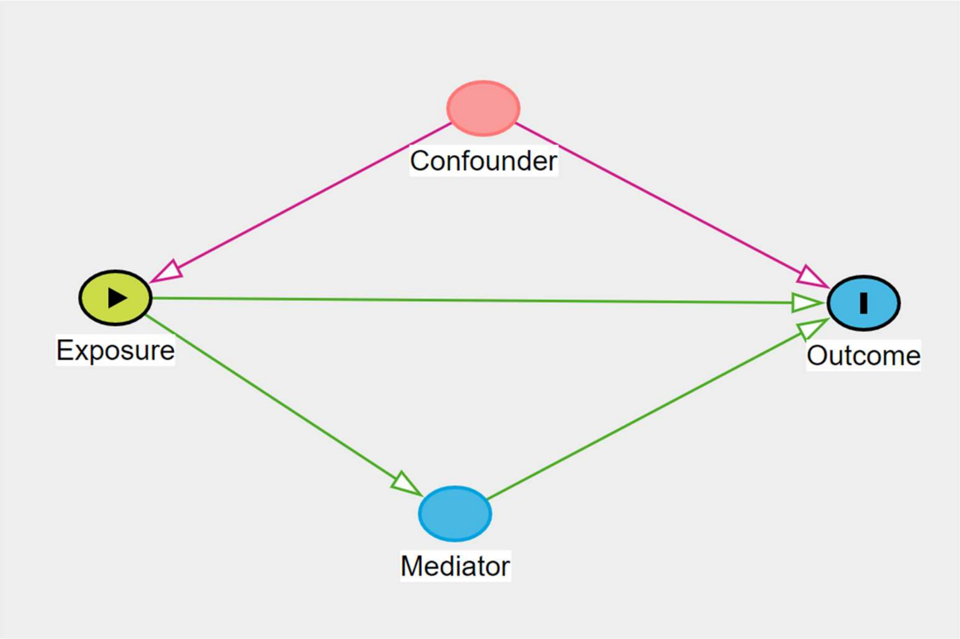

Figure 1. Directed Acyclic Graph illustrating two types of relationships affecting the causal link between an exposure and an outcome. The causal paths are shown in green. Confound-inducing paths are shown in red.

We are now in a position to understand why the coefficients for the control variables (i.e., the covariates included in the model) cannot be interpreted individually as causal. The coefficient of the exposure variable can be interpreted as the effect of that variable on the outcome, *at any given level of the control variables*. By contrast, the coefficient of a control variable can only be interpreted as the effect of that variable on the outcome *when the exposure variable is fixed,* i.e., blocking the effect of the control variable on the exposure variable. See Westreich & Greenland (2013) for a detailed explanation.

Ascertaining the status of variables in relation to the causal relationship of interest (and identifying the consequent set of variables to adjust for in the relevant model, i.e., the confounders) can be worked out visually with a user-friendly online tool: DAGitty http://www.dagitty.net/ (Textor et al, 2016). DAGitty is a browser-based environment[1] for creating, editing, and analyzing causal diagrams (also known as Directed Acyclic Graphs - DAGs), as illustrated theoretically in Figure 1 above and in practice in Figures 2 and 3 below.

Examples of DAGs based on Paradis' state-of-the-art review

After this all-too-brief introduction to causal inference modelling, we now attempt to apply this approach to some of the variables discussed in Paradis' paper. In doing so, we have used her state-of-the-art review to generate a DAG for the variables in Figures 2 and 3. The two figures include the same set of variables, but each focuses on a different causal relation. We kept the outcome variable constant (i.e., the child's proficiency in the Heritage Language (HL)), but chose a different exposure variable in each graph: the Amount of HL exposure in the home in Figure 2, and Family language policy (FLP) in Figure 3. In each case, the DAG situates the (hypothetically causal) relationship of interest within the context of variables known or suspected to be related (directly or indirectly) with either the exposure or the outcome variables. It therefore represents the analyst's understanding of the state-of-the-art. It also requires determining the temporal order of "crystallisation" of the variables: if a variable A is situated to the left of a variable B, it implies that A precedes B. For instance, Figure 2 assumes that Parental HL Proficiency is "set" before FLP (and influences it, as indicated by the arrow).

[1] Note that DAGitty does not perform statistical analyses. Is it not concerned with the actual data (nor with issues of sample size). DAGs help researchers identify and describe their assumptions in relation to the data collection process. When drawing a DAG, researchers should not start from the available data, but consider *all* the variables that *could potentially* cause the exposure and the outcome. Some of the variables in the DAG might in fact not be directly measurable.

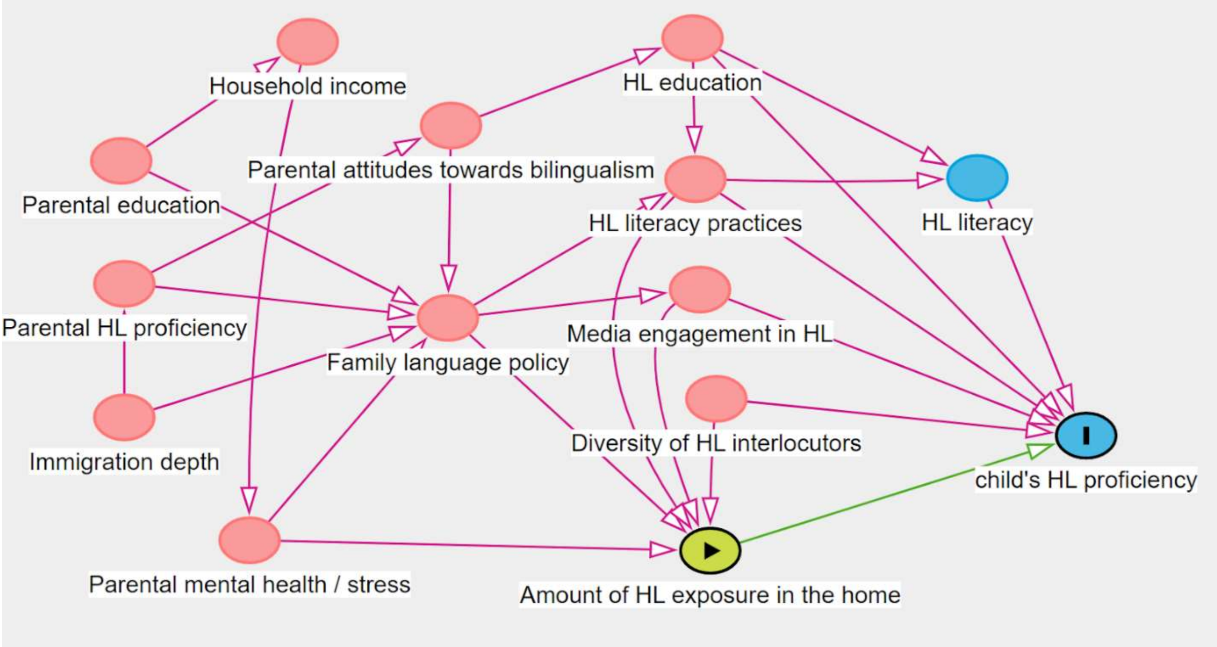

Figure 2. DAG representing the causal path (cf. green line) between the Amount of Heritage Language Exposure in the Home and Child's Heritage Language Proficiency and the biasing paths (cf. red lines). Ancestors of the outcome appear as blue dots; ancestors of the exposure and *outcomes appear as red dots.*

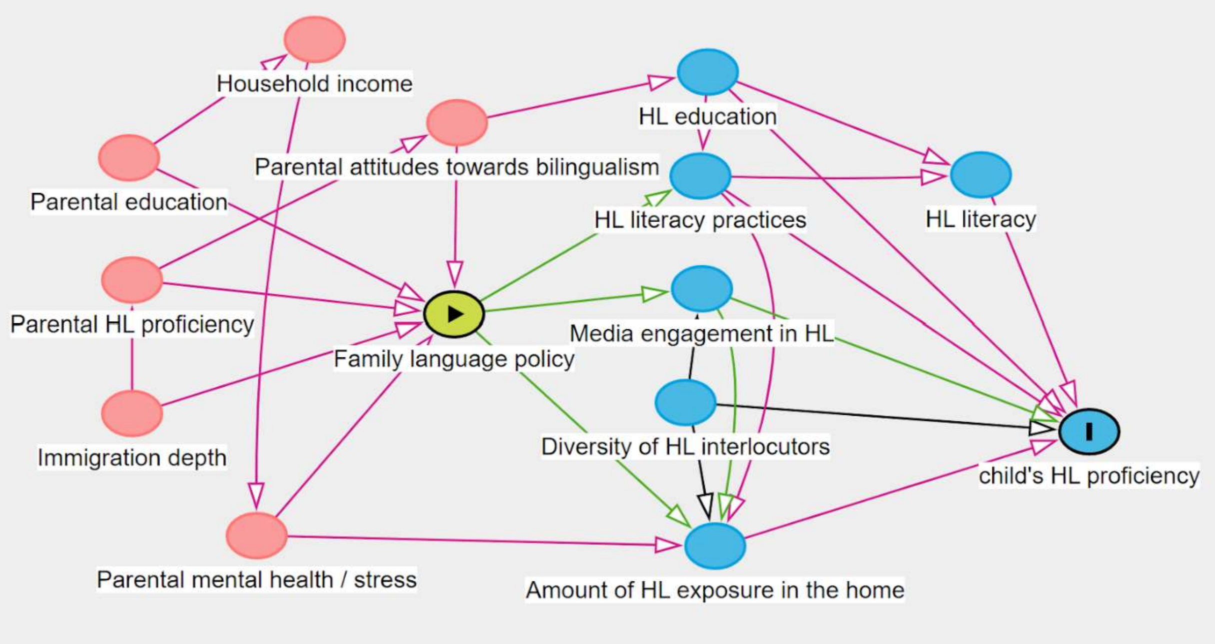

Figure 3. DAG representing the causal path (cf. green line) between the Family Language Policy and Child's Heritage Language Proficiency and the biasing paths (cf red lines). Ancestors of the outcome appear in blue; ancestors of the exposure *and* outcomes appear in red.

What do we see in our two DAGs? In the DAG in Figure 2, the causal path between Amount of HL Exposure in the Home and Child's HL proficiency is not mediated by any other factor. There are however a number of confounder variables (e.g., Diversity of HL interlocutors, Media engagement in HL and HL literacy practices) indicated in red in the DAG. These variables would need to be included as covariates in the statistical model. In the DAG in Figure 3, the outcome remains the same but the exposure differs. As a result, the confounding and mediating variables also differ. Here, the hypothetical effect of FLP on HL proficiency is mediated by three variables (as shown by the green lines between FLP and HL literacy practices, Media engagement in HL and Amount of HL exposure in the home). As mediators, these variables cannot be included in the model. There are also confounders (e.g. HL education, which determines both HL literacy practices and HL proficiency, and Parental attitudes towards bilingualism, which determines both FLP and HL education). Only confounders have to be included in the model as covariates, as explained above. The exact set of variables to include in the model is computed automatically by the DAGitty tool.

Importantly, not having an arrow between two nodes means that the two variables in question are assumed to *not* influence each other. For instance, in our DAGs, we assume that FLP does not influence the diversity of HL interlocutors (as this can vary substantially outside the home). This could be seen as contentious, and calls for debate. We are not committed to either view, and just raise this point as an illustration of how the approach works.

Reflections and outlook

What have we learned from our own first dabble into the world of DAGs? One very clear consequence of having to specify the nature and direction of the relations between the various internal, proximal and distal factors affecting bilingual children's language abilities is that it forces us to explicate unarticulated assumptions (e.g., FLP is determined by Parental HL proficiency) and acknowledge the biases inherent in our data. DAGs force us to consider what variables really mean (rather than just what we can measure). Whilst the two DAGs shown here are clearly incomplete and subject to debate, they are a first step toward implementing advanced data modelling techniques in order to better understand the types of relations which exist between variables of interest. They take us one step further than the standard predictive models used in most of the research in our field thus far, and beyond the somewhat simplistic model in Figure 1 in Paradis' paper. A key characteristic of DAGs is the temporal ordering of variables: those appearing on the left are assumed to be determined earlier in time than those appearing to their right (e.g. Parental education is likely to be set before FLP). To some extent, the nodes appearing on the left in our DAG tend to align with Paradis' distal variables, with the exception of literacy and education in the HL. The DAG thus invites us to rethink what it means for a variable to be distal or proximal.

The DAGs we have generated for the purposes of this commentary illustrate the challenge involved when determining the (potential) relations between the myriad of variables involved in bilingual language development. The further steps required to navigate this challenge and to arrive at any kind of satisfactory solution will need to involve a collaborative effort as DAGs by necessity reflect the state-of-the-art. Using causal inference modelling thus requires us to think more carefully, as a field, about the types of relationships between variables of interest, and about the biases inherent to our data. This approach also gives us the means to better understand so-called "risk factors" and "protective factors", and to conceptualise them properly within a causal approach (Huitfeldt, 2016).

We look forward to the exchanges arising from this invitation to embrace a causality approach to the study of individual differences in bilingual/multilingual language outcomes.

References

Arnold, K., Davies, V., de Kamps, M., Tennant, P., Mbotwa, J. & Gilthorpe, M . (2020). Reflection on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning, *International Journal of Epidemiology*, *49(6)*, 2074–2082. doi:10.1093/ije/dyaa049.

Hernán, M.A., Hsu, J. & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks, *CHANCE*, *32(1)*, 42-49. doi:10.1080/09332480.2019.1579578

Huitfeldt, A. (2016). Is caviar a risk factor for being a millionaire? *British Medical Journal*, *355*. doi:10.1136/bmj.i6536

Textor, J., van der Zander, B., Gilthorpe, M., Liskiewicz, M., & Ellison, G. (2016). Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology 45(6)*, 1887-1894.

Westreich, D., & Greenland, S. (2013). The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *American Journal of Epidemiology*, 177(4), 292-298. doi:10.1093/aje/kws412