Introduction

Recognizing the complexity of L2 development, applied linguistics researchers are turning to correspondingly sophisticated methods to analyze and statistically evaluate data. For instance, since L2 development is inherently about change over time, and applied linguists tend to operate with cluster-randomized rather than individual-randomized designs – e.g. students nested within classes and schools, children from different families, speakers in different neighborhoods and ethnolinguistic groups, etc. – statistical methods are required which adequately analyze change and account for variability and auto-correlation in linear as well as non-linear patterns. In this chapter, we advocate the generalized (mixed-effects) regression framework, including linear mixed effects regression models (mixed models), generalized additive (mixed) models (GAM[M]s) and multinomial logistic regression, which represent an important statistical development of recent years and provide invaluable tools for analyzing data within applied linguistics.

We focus on what these models are, how they work and why/when applied linguists should use them. \(^1\) We review three recent studies that employ sophisticated statistical methods to illustrate their power in investigating issues such as the age debate and factors determining source language transfer in L3 acquisition. However, most of the concepts and techniques discussed are applicable to other types of short-term and long-term temporal trajectories and other types of contexts. We close with a discussion of limitations of inferential statistics and desiderata for future research.

As a first step, however, we need to have a brief look at the core of most scientific investigations: causation, that is, cause–effect links as well as goals of inferential statistics. Because scientific studies are centrally concerned with causal relationships, a major question for applied linguists is how to understand causation in the social world. We begin, then, by exploring causation and how causes and consequences are envisaged in applied linguistics.
The theoretical side of the coin: inferential goals and the concepts of causation and causality as developed by applied linguists

Criteria for causal powers

Applied linguists, like other scientists, are interested in the “actual” cause of a wide variety of phenomena – from foreign language learning development to dyslexia, gender differences to the nature–nurture debate, language change to language attrition, and language aptitude. In the classical paradigm, theories are developed to describe, explain, and predict. This is also relevant to the recipients of applied linguistics research. Thus, policy-oriented discussion requires precision and causal directionality, and policy makers are interested in “factors” that favor or prevent a specific outcome, especially regarding factors that respond to intervention.

For many researchers “[e]xperimental research is the most powerful research method because it allows experimenters to manipulate, isolate, and control chosen variables, and thereby determine cause and effect” (Huffman, 2012, p. 21, our emphasis). The field of individual differences in second language acquisition (SLA), for instance, has frequently relied on the claim that traits cause language-related phenomena – that L2 learners possess something (motivation, anxiety, etc.) which causes/disposes them to behave in certain ways relevant to that trait (Furnham, 2016). The stronger the trait in learners, the more likely they are to act in such ways. Causal models of SLA thus provide evidence that attitude to the learning situation is the variable which initially promotes student classroom activity, and ultimately competence levels (Lorenzo, Moore, & Casal, 2011). In this context, linguistics is criticized as “reductionist”, attempting to locate causes in the smallest units possible (e.g. Larsen-Freeman, 2017).

Rerum cognoscere causas (“to know the cause of things”)

An extensive literature outlines the biases and “errors” people make in their understanding of causal powers (see e.g. Furnham, 2016); one might mention here the well-known “correlation is not causation” dictum. There is less discussion on how (applied) linguists themselves fall into “traps” when trying to assign cause and consequence to various phenomena.

There are three main problems with simple linear causal links that are often made in traditional approaches in applied linguistics: (1) the reasoner’s sensory input does not of itself explicitly illuminate causal relations (i.e. acquired causal relations must be somehow computed from the sensory input [Cheng, 1997]); (2) simple-causal links ignore the interconnections in a system, which render the isolation of variables difficult; and (3) simple-causal links limit the role ascribed to complexity (Larsen-Freeman, 2017). We examine now more closely each of these phenomena.

In contrast to the physical world, in the social world it is extremely difficult to isolate pertinent causal links owing to the existence of recognized counterexamples and uneliminable sources of unpredictability in human affairs; it has to be added, however, that the laws in the physical sciences are often imbued with an exaggerated precision attribute, and the social sciences tell us what usually, typically, or rarely happens rather than what always or never happens (Salmon et al., 1999). MacIntyre (1984) describes generalizations in the social sciences as statistical rather than universal. Statistical laws state some probability that an event of a certain type will co-occur with or be followed by an event of another type (Salmon et al., 1999). Finally, it is important not to confuse generalizability with causality, as causal relations can also be disclosed in case studies (see our section on “Introducing nonlinearity to the field of applied linguistics” below).

Generalizability, a crucial concept in quantitative research, is understood in a number of different ways in the literature. Most often, the term is used to denote statistical, sampling-based
generalizability, which refers (1) to the inferential link between an observation and an interpretation (internal validity) or (2) to an inference from the observation of a single case to a larger group (external validity). Regarding external validity, sample size is a design feature discussed critically in literature, as small samples debilitate statistical power. For instance, according to Duff (2006, p. 68):

> it is commonly accepted that quantitative research, with appropriate sampling (random selection, large numbers, etc.), research design (e.g. counter-balancing of treatments, ideally with a control group, pre-post measures, and careful testing and coding procedures), and inferential statistics where appropriate, has the potential to yield generalizable results.

However, an increase in sample size leads, one notes, to an increase in the generalizability of one sample to other samples that the same sampling procedure would produce; it does not lead to an increase in the generalizability of any sample estimate to its corresponding population characteristic (Pfenninger & Singleton, 2017). Moreover, it has been increasingly recognized that a theory may never be scientifically generalized to a setting where it has not been empirically verified (e.g. Larsen-Freeman, 2017). At the same time, case studies and qualitative methods are often dismissed as unscientific because of their apparent lack of generalizability and cause–effect linkages, although they have the potential to treat causation more broadly and include final causation, i.e. the purpose or meaning of things.

Another difficulty is ensuring that causal explanations are correctly attributed – and, relatively, there is the problem of multiple causation: often a complex set of distal and proximal factors operate at the same time. Hence, applied linguistics texts often include warnings about the complexity of identifying causes. Ellis (2006), for instance, cautions of the danger with the exclusive focus on a single independent variable, as the research is quick to attribute causality when significant relationships are uncovered. De Bot (2008, p. 173) illustrates the interconnectedness of variables in language learning:

> In a carefully designed quasi-experimental study, no differences among the conditions were found, which was of course disappointing for the researchers, but not really surprising. . . . In a way, it would have been surprising if such a single factor had explained differences in learning success.

This explains in part the difficulty of isolating variables or establishing direct cause–effect relationships. Even if we manage to isolate one particular cause to judge what its relation to a certain effect is, the other causes that enable or hinder the relation between the cause and effect are often unstable. How this impacts on research methods is still a matter of debate (see e.g. Harré & Moghaddam, 2016).

A final challenge is that applied linguists often deal with aggregate, social, non-individual entities (e.g. institutions, such as schools). Lowie (2017), for instance, states that “group observations are analyzed using statistical procedures that are widely accepted as proof for effects and relationships, and that results in generalizations of sample group scores to population” (p. 125). Studying the aggregate of a group to which the individual belongs raises the question whether an increase or decrease in aggregate-level error variance in the dependent variable outcomes documented by the results of, e.g., an ANOVA-type analysis warrants the inference that this effect is realized in each one of the individuals to whom the treatment condition or personal attribute producing that effect applies – whatever the exact nature of the causal mechanisms producing a given statistical effect might be (Furnham, 2016). The central question is, as Furnham (2016)
points out, whether aggregated and statistical techniques can do more than “hint at” possible causal patterns on an aggregated measure. In other words, there is no guarantee that the average “grand sweep” (Lowie, 2017, p. 137) is representative for any of the individuals of the population. Our current understanding of SLA views it as a complex, dynamic, ecologically situated, multivariate phenomenon – hence the impact of theories like chaos/complexity theory (Larsen-Freeman & Cameron, 2008) and dynamic systems theory (de Bot, Lowie, & Verspoor, 2007; Lowie, 2017). These characteristics of language development can make it difficult to adopt traditionally formulated, linearly framed research methods (Lorenzo et al., 2011).

Causal concepts and modern research methods

The previously discussed issues with causation lead to particular demands on good research design for finding causes in applied linguistics. Modern explorations of cause and consequence suggest that in answering causal questions, the role of context is central. According to Furnham (2016) a satisfactory account of cause–effect relationships requires the scientist to go beyond the mere establishment of statistical covariations between variables and into a theoretical elaboration of the dynamics that could possibly be producing those statistical covariations. This is particularly relevant for applied linguistics, since several qualities of language-related phenomena make the non-reductionist, ecological, systems view that dynamics and complexity theories – henceforth Complex Dynamic System Theory (CDST) – afford appropriate (Larsen-Freeman, 2017). CDST is a theory of change describing relationships among embedded subsystems that cannot be assumed to be linear, given that intra-learner variability is a necessary condition for development (de Bot et al., 2007). This approach questions the feasibility of investigating cause–effect relations: since it is highly unlikely that a single cause (or a handful of independent variables) will give rise to such a complex event as L2 learning, the suggestion is to focus on ‘tendencies, patterns and contingencies’ rather than simple cause–effect explanations (de Bot & Larsen-Freeman, 2011, p. 23). In its strictest form, CDST also requires us to revisit the idea of predicting behavior, since no two situations can be similar enough to produce identical behavior (Larsen-Freeman & Cameron, 2008). To put bluntly, in a complex world, “we lose predictability; the nature of explanation changes; cause and effect work differently” (Larsen-Freeman & Cameron, 2008, p. 72).

With advancing statistical techniques, applied linguists can study important relationships holding between linguistic processes, features, and forms that conform to the main premises of complexity theories. For instance, outlining alternative methods in the context of CDST methodological principles, Lowie (2017) advocates the use of statistical techniques that can handle non-deterministic (i.e. probabilistic) cause–effect relationships. In the following, we explore more closely some of the possibilities that applied linguists have at their disposal.

The methodological side of the coin: three studies

Beyond the general linear model: some thoughts on the generalized (mixed-effects) regression framework

Our first case revolves around the age factor, one of the most frequently used variables in applied linguistics, be it as a central variable (i.e. the main variable of interest) or as a socio-demographic variable to assist the study of something else. Applied linguists have a way of anticipating age matter-of-factly, through parlance about “age effects”, and reporting it as a cause of deficient behavior, which makes this an interesting case for our discussion.
Age-related research has followed the trend towards using sophisticated statistical tests as well as in multiplying the range of tests used. From the 1990s, a trend appeared for researchers to move beyond a focus on age as a stand-alone variable, and to explore its interaction with other variables. This period saw a marked increase in the use of statistical methods: inferential statistics such as *t*-tests (e.g. Jia & Fuse, 2007) or (multivariate) analyses of (co)variance (e.g. Flege, Yeni-Komshian, & Liu, 1999), multiple regression analyses (e.g. Muñoz, 2014), or a factor analytic approach (e.g. Csizér & Kormos, 2008), as well as correlations (e.g. Kinsella & Singleton, 2014) and structural equation modeling (SEM) (e.g. Jackel, Schurig, Florian, & Ritter, 2017).

The typical linear regression model is a generalized linear model with a Gaussian distribution and “identity” link function. It requires each observation to be independently distributed and be determined solely by its underlying parameters plus some random noise. Violation of this assumption can lead to problematic estimates (see e.g. Sóskuthy, 2017). For instance, if we run an ANOVA to compare differences between early and late starters in a school context, the model will automatically assume that measurements for a given age group have “uncorrelated errors” (e.g. that all early starters are in the same class, school, family, ethno-linguistic group, neighbourhood, etc.). However, that is not the way most data are structured in applied linguistics, where groupings and clusters abound. When we fit a regression model to our data set without taking the grouping/temporal/spatial structure into account, we ignore the fact that individual data points are dependent on the presence (or absence) of grouping structure as well as temporal/spatial structure in the data. Smoothing over variability in this way obliterates the very information that we may need (Ellis & Larsen-Freeman, 2006). Also, we may obtain age effects which are illusory (see below). This necessitates finding methods that take account of environmental influences.

Recently, we have been able to observe the emergence of the citing of effect sizes and confidence intervals in applied linguistics, which can be attributed to the requirements of some journals (see e.g. Ellis, 2000). What is more, the class of statistical models called linear mixed-effects regression modelling, which has been extensively applied in other fields such as ecological research, appears to be increasing in applied linguistics.

To illustrate the advantages of mixed models and the problematic nature of traditional analyses, we compare a traditional multivariate analysis of variance with a mixed model, using the same data set from the Beyond Age Effects (BAE) study (Pfenninger & Singleton, 2017, 2019). The overarching research question of this study is: how does age of foreign language learning onset affect the foreign language proficiency level of different learner populations in the short and the long run? In total, 636 secondary school students were recruited at the beginning and end of mandatory school-time (ages 13 and 18 respectively); all had learned Standard German and French at primary school, but only half had had English from third grade (age 8) onwards, the remainder having started five years later at secondary school. There were 200 monolinguals, 144 simultaneous bilinguals with biliteracy skills, 107 simultaneous bilinguals without biliteracy skills, and 185 sequential bilinguals. The sampled students were nested in a hierarchical fashion within classes within schools. For instance, the monolingual students were integrated in 12 different classes in five different schools. Among other tasks, each participant filled in 60 gaps in a receptive vocabulary task (i.e. the academic sections in Schmitt, Schmitt, and Clapham’s [2001] Versions A and B of Nation’s Vocabulary Levels Test), which were later rated for correctness.

We first use a general linear model with the group of the 200 monolingual students (i.e. ANOVA at Time 1 and ANCOVA at Time 2) as well as effect sizes eta-squared (η²). In this analysis the data are first aggregated, averaging first over participants (i.e. the two groups of
early and late starters), and second over the 60 items. ANOVA at Time 1 reveals significant differences in terms of receptive vocabulary knowledge between the 100 monolingual early starters and the 100 monolingual late starters ($F = -6.12, p < .001**$, $\eta^2 = .320$). At Time 2, ANOVA indicates that receptive vocabulary still reaches statistical significance in favour of the early starters ($F = -6.12, p < .001**$), with a small effect size ($\eta^2 = .112$).

However, assuming that measurements for a given age group category have uncorrelated errors is problematic, as described earlier. We need to know whether between-group and within-group differences are due to age effects or to something else such as effects of class/school context and climate. While correlated data are explicitly forbidden by the assumptions of standard (between-subjects) (M)AN(C)OV A and regression models, this is where mixed-effects models really shine (Cunnings & Finlayson, 2015). Also, these models run on the unaveraged data and thus elegantly work around the problem of discretizing continuous variables such as age and foreign language proficiency. Problems associated with such discretization include: (1) it draws categorical boundaries where none exist; (2) it spuriously suggests the presence of cut-offs or threshold effects; (3) it makes investigating the dynamics of the learning trajectory impossible; and (4) it leads to a decrease in the probability of finding a pattern if there is one (Vanhove, 2013).

Within the mixed model framework, a distinction is made between “fixed” and “random” effects. Fixed effects can be continuous (such as “time”) or categorical (such as a condition difference) (Winter & Wieling, 2016). The BAE study examined how L2 development was influenced by age of onset (AO), so the analysis included a fixed effect for AO. L2 proficiency was also expected to change over time, so a fixed effect of “time” was needed. A particularly important aspect of model fitting is the presence of interactions. In the BAE case, we wanted to find out if age of onset influenced the change in proficiency over time, and thus included the interaction of time and the respective condition variables (“Will age of onset effects differ with respect to how the effects unfold over time?”).

Age effects may be restricted to certain items or tasks (or certain subjects, items, classes, or schools). Furthermore, whereas age of onset may vary between classes and schools, in a longitudinal research design, the continuous predictor “time” varies within them, as each student, class, and school is tested at multiple points in time. Thus, students, classes, and schools may differ not only in overall average proficiency, but also in their sensitivity to the change in proficiency over time. Random slopes are required to model this type of variance (see Cunnings & Finlayson, 2015).

Including such random effects makes a significant difference regarding results yielded by our data set. If we subject it to a multilevel analysis – which yields adequate estimates of variances and therefore correct standard errors, correct inferences, and (likelihood-based) $p$ values – we also find a significant advantage for an earlier start in terms of receptive vocabulary at Time 1 ($\beta = -2.20, SD = .80, t = -2.75, p = .130$). However, there is no longer any sign of an age effect for receptive vocabulary at Time 2 ($\beta = -2.20, SD = .80, t = -2.75, p = .130$). The conclusions drawn, then, are fundamentally different: in the ANOVA analysis, we conclude that an earlier age of onset is advantageous for receptive vocabulary in the short and the long run, but in the mixed model we posit receptive vocabulary benefits from an earlier age of onset only in the short run, such effects disappearing during secondary schooling.

This illustrates how drastically clustering effects of streamed classes may minimize age effects. While any statistical test will always have a 5% chance of detecting a “significant effect” even if nothing is going on – known as the Type-I error rate – the chances of getting such false positives may here increase substantially. In the BAE study, when clustered data was analyzed without taking the clustering into account (e.g. in ANOVA-type analyses), the Type-I error rate rose by 50%, implying age effects where there were none (Pfenninger & Singleton, 2017).
**Introducing nonlinearity to the field of applied linguistics**

Applied linguistics research has not been to the fore in employing sophisticated procedures to analyze truly longitudinal data. This is a serious limitation, particularly for SLA, where many questions concern time and timing. Given the centrality of time in this domain, more attention to longitudinal research practices and findings is desirable (see also Ortega & Iberri-Shea, 2005). However, longitudinal data in applied linguistics research are often analyzed utilizing the same inferential statistics that are employed in cross-sectional research. While ANOVA methods can provide a basis for a longitudinal analysis where the study design is very simple, they have many limiting shortcomings. For instance, they cannot capture the non-linear nature of L2 learning trajectories. This is disappointing for models seeking to make causal inferences about such trajectories. Thus, if visualization of the temporal trajectories reveals nonlinearities (e.g. a sinusoidal pattern reflecting the fluctuations in L2 development), more complex models require consideration. As described earlier, the most appropriate method of analyses of development and change in language as processes in time will involve non-linear analyses of longitudinal case studies, focusing on variability, trends, and interactions over time – particularly methods with dense measurements that focus on studying change as it occurs in the data covering the entire period during which development is studied (Lowie, 2017).

There are many process-oriented approaches for the analysis of time series data; according to Lowie (2017), three main types of analysis have been used to unveil aspects of L2 development: “the analysis of variability over time at different time scales, the analysis of dynamic interactions between subsystems, and the analysis of dynamic interactions between persons using the language” (p. 129). We here outline one method that is general and flexible enough to deal with any kind of data structure (continuous, categorical or count) and can model any type of nonlinearity (Winter & Wieling, 2016), thus being applicable to many kinds of data sets: Generalized Additive (Mixed) Models (GAM[M]), originally developed by Hastie and Tibshirani (1986).

The principle behind GAMs resembles that of regression, except that instead of summing effects of individual predictors, GAMs sum smooth functions, resulting in a trend line that best fits the data so as to derive a trend across time, which might be used to predict future values (Sóskuthy, 2017; Winter & Wieling, 2016). Functions permit the modeling of more complex patterns, which can be averaged to obtain smoothed curves that are more generalizable. GAMM is an extension from GAM for grouped or clustered data. For these models, a within-group variance-covariance structure can be used to account for the corresponding within-group autocorrelation. Time series data, treated as single cluster data, can also be modeled by GAMM. GAMMs are ideal for analyzing non-linear change where there are nested dependencies, such as time points within dyads (in repeated interaction experiments) or time points within learning trajectories (in longitudinal research with dense data) (Winter & Wieling, 2016). Increasing the number of measurements per participant results in more information about the individual trajectories. Furthermore, adding measurements from additional subjects contributes more information towards the underlying parameters. Besides incorporating random intercepts and slopes, GAMMs also offer a third option: random smooths. Random smooths are similar to random slopes, but they are more flexible than the latter: while random slopes can capture only by-group variation in linear effects, random smooths can also deal with by-group variation in non-linear effects (Sós kuthy, 2017).

To exemplify, GAMMs are currently being used to analyze data from a longitudinal study with dense measurements called Age and Immersion Milestones (AIM), conducted by the first author (Pfenninger, in press.). It explores the impact of age of first bilingual language exposure.
on the development of writing complexity in 91 children in Swiss bilingual programs with 50% of the content being taught via the community language German and 50% via English. They varied in their home language and age of first bilingual instruction (5, 7, or 9) and were tested up to 32 times (four times a year over a maximum period of eight years in [pre]primary school). The analysis of linguistic data was conducted with the goal of testing whether age of onset had a significant effect on the trajectories under investigation. Another key question was when L2 development was statistically significantly increasing (or decreasing). We included age of onset as a continuous variable, and looked at the (non-linear) literacy pattern over time in order to examine potential latency effects, while simultaneously accounting for subject and item variation. “Time series” was included as a “random effect”. In other words, a smoothed function was fitted by combining several low-level functions (such as a linear function, a quadratic function, a logarithmic function, etc.) across the whole time span (Winter & Wieling, 2016). The analysis revealed the learners’ individuality in the learning process and enabled the first author to identify meaningful, consistent patterns of individual differences as well as phases with significant development (both increase and decline). Furthermore, the fine-grained GAM analysis revealed evidence of a clear discontinuity in the age curve that could not be captured in a more traditional analyses. In this context it is also worth mentioning the Meulman, Wieling, Sprenger, Stowe, & Schmid (2015) study of the effect of age of arrival on grammatical processing in L2 learners as measured by event-related brain potentials. GAM revealed no main effect of age of arrival, but only effects in interaction with time. Furthermore, ANOVA suggested the presence of a critical period before the age of 17 – learners with an age of arrival between 7 and 16 but not those in the group with an age of arrival of 17 and later showing a P600 for gender violations. The absence of such a discontinuity in the GAM analysis argued against the existence of a critical period, justifying Birdsong’s (2004) call for more granularity in studies of age effects.

While such a microgenetic method may differ from the causal-mechanical tradition, it does not exclude causality: according to Larsen-Freeman and Cameron (2008, p. 233), the process of “co-adaptation” describes a kind of mutual causality, in which change in system leads to change in another system connected to it, and this mutual influencing continues over time. CDST types of explanations do not exclude generalization either. As Lowie (2017, p. 137) points out, even though observations in case studies cannot – and need not – be generalized to groups of individuals or populations of language learners, causal relations can certainly be disclosed in individual trajectories of single case studies and can have a direct bearing on underlying theory (see also Pfenninger & Singleton, 2019). If we want to make general claims about L2 development with this research design, the time series data, treated as single cluster data, becomes the target of inferential statistics. That is, we may consider each set of clusters in the research design as a sample from a population of clusters that we wish to generalize upon.

**Analyzing categorical data using logistic regression**

The third case we would like to discuss in this chapter relates to lexical transfer in multilingual learners. The most important difference between a second language (L2) learner and a multilingual learner is that the multilingual learner can rely on and is influenced by more than just one language when acquiring a new language. For a L2 learner, crosslinguistic influence can only ever arise from their L1, while the multilingual learner may experience transfer from their L1 and their L2(s). Understanding what factors determine the predominance of one background language as source of transfer over another in multilingual learners helps us develop
psycholinguistic models of learners’ mental lexica and thus represents an important area of research within multilingual language acquisition.

The vast majority of studies on lexical transfer in third language acquisition to date have relied on qualitative and exploratory methods to investigate the source language of transfer in multilingual learners (Williams & Hammarberg, 1998; Bardel & Lindqvist, 2007). These methods represent an excellent tool to uncover new phenomena, delineate patterns of behaviour, generate hypotheses, and establish potential explanatory factors. The purpose of the study discussed here (Neuser, 2017) was to test the factors established through previous qualitative research using inferential statistics. Due to the categorical nature of the dependent variable, with more than two possible outcomes (L1 + multiple L2s), multinomial logistic regression presented itself as the most appropriate statistical method.

The analytical advantages of the logit transformation in dealing with discrete binary outcomes were put forth in a series of papers in the 1960s by Sir David Cox, at the same time that multinomial logistic regression was put forth by several other scholars.

Multinomial Logistic Regression (MLR) is an extension of logistic regression, which analyzes a dichotomous/binary (i.e. with two possible outcomes) dependent variable. All logistic regressions are generalized linear model procedures which use the same basic formula, but instead of regressing for the probability of a continuous dependent variable, they regress for the probability of a categorical outcome. Ordinary linear regression and logistic regression both have the same goal of predicting a particular outcome. Both types of regression can be used to fit a predictive model to an observed data set of y and x values. If one then encounters an additional value of x without its accompanying value of y, the fitted model can be used to predict the value of y (Seal, 1967).

Multinomial logistic regression is used to predict categorical placement in, or the probability of, category membership of a dependent variable (DV) based on multiple independent variables (IVs) (Menard, 2010). As discussed earlier in the chapter, the focus of linguistic analysis should be on tendencies and patterns, rather than simple cause–effect explanations. MLR offers us the discovery of non-binary probabilistic patterns in linguistic behaviours. One example of when multinomial logistic regression can be used would be to determine the probability of a particular pronunciation of *thought* (DV) to occur (possible outcomes: /θɔt/, /tɔt/, and /ft/), given particular characteristics (IVs) (e.g. location, L1, education, gender). For this study, it was used to explore which background language (BL) learners will transfer from when learning a new language, given a number of variables: proficiency in the BLs, exposure to the BLs, psychotypology, and the L1/L2 status of the BLs.

Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership. MLR compares multiple groups (i.e. outcomes) through a combination of binary logistic regressions, using one of the groups as reference group.

One important advantage of multinomial logistic regression over other statistical models, such as linear regression models or ANOVAs, is that it does not make any assumptions of normality and homogeneity of variance for residuals, or linearity of variance for the independent variables. This makes it an extremely useful method in research that uses complex human data, which often do not fulfil the required assumptions of other statistical analysis tools (Menard, 2010).

In addition, MLR offers a solution to the common problem of confounding variables that was identified in a number of previous studies on lexical transfer (e.g. Bardel & Lindqvist, 2007), through their ability to control other relevant factors when one is being tested for its predictive power. The results of this study, for example, have shown how well proficiency,
exposure, psychotypology, and the L1/L2 status, *each in their own right*, predict which background language will dominate as source of lexical transfer in any given individual, while controlling for the other variables. As previously discussed, given the complexity of language learning, we have to assume that for most phenomena, there is multiple causation, and that focusing on one independent variable alone can be dangerous. This model allows us to test the predictive power of a variety of factors, while controlling for each one of them.

Finally, the analysis using MLR is based on individual tokens of transfer, rather than percentages and/or group averages. This means that the source language of every transferred item is taken as a data point for the dependent variable, and not, for example, an average of how many items are transferred from a particular source language within the whole group.

**Conclusion**

Our chapter on inferential statistics aimed to illustrate a more general feature of the analysis of human practices – the way it is easy to slip into taking a certain picture for granted. As a matter of fact, research in many areas of applied linguistics (e.g. research into individual differences variables) has noted the need for more care about inferences from statistical results. For instance, considering that linguistic phenomena and processes are complexly multidetermined and not really amenable to simple explanations, the measure of success clearly cannot be to make simple causal statements, attributing a consequence to a single cause. Thus, when working with clustered designs, “the hierarchical sources of variability cannot be ignored without seriously contributing to errors of inference, compromising the validity of results and research conclusions” (O’Connell & McCoach, 2008, p. 5). Linear cause and effect links do not produce satisfying explanations that are respectful of the interconnectedness of the many nested levels and timescales of language development. Given this, Ortega and Iberri-Shea (2005) caution that if “more large-size longitudinal quantitative studies are conducted in SLA, it will be important to train ourselves in the use of statistical analytical options that are available specifically for use with longitudinal designs and data” (p. 41).

In our discussion of recently developed statistical options, we have attempted to include both product-oriented research (i.e. observations of group behavior at one moment in time) and process-oriented research (i.e. observations of individuals over time; Lowie, 2017). Specifically, we have outlined the mechanisms of three statistical analyses that, in some way, work against the “decontextualizing, segregating, and atemporalizing” (Larsen-Freeman & Cameron, 2008, p. 257) of language. They illustrated how recent attempts have been made in applied linguistics to avoid interpreting significance in a binary fashion. For instance, Larson-Hall and Plonsky (2015) emphasize the relevance of descriptive statistics and effect sizes (see also Baffoe-Djan & Smith, and Loewen & Godfroid in this volume). Methods such as mixed effects models, like Bayesian statistics, offer solutions to the increasing criticism that null hypothesis significance testing and generalizations based on group observations using Gaussian statistics have recently faced. They also offer a compromise between generalizability and variability.

Finally, it needs to be mentioned that even though statistical tests like mixed models can take account of the fact that a student’s performance is dependent on which class he or she is in, what exactly leads to these class differences (group dynamics, teacher personality, quality of instruction, school environment, etc.) can often not be clarified with quantitative methods due to limitations of the research design. Here, qualitative and quantitative research can complement each other. For example, when statistical results about the effects of causes are reported,
Inferential statistics

the qualitative analysis is helpful for understanding e.g. the direction of causation in a specific case. This complementarity is one reason why mixed-method research is possible and highly recommended (see Singleton & Pfenninger, 2015; see also Hashemi in this volume).

Note
1 Note that we are not going to make recommendations for choices about model fitting.

References


Pfenninger, S. E. (in press). About the INTER and the INTRA in age-related research: Evidence from a longitudinal CLIL study with dense time serial measurements. Linguistics Vanguard.


