Chapter 9

Beyond binary dependencies in language structure

Damián E. Blasi

University of Zürich,
Max Planck Institute for the Science of Human History

Seán G. Roberts

Max Planck Institute for Psycholinguistics

The study of the regularities in the structures present across languages has always been a quest in close contact with the analysis of data. Traditionally, causal dependencies between pairs of typological variables (like word order patterns or the composition of segment inventories) have been argued for on the basis of language counts, namely how many languages out of a sample exhibit certain patterns in contrast to others. Regularities of this kind have been used in virtually all theoretical camps, and researchers made them part of their discussion on functional pressures on language, cognitive schemes and the architecture of a putative common computational core underlying language, among other things. This popularity resides, without doubt, in the strength and simplicity of the idea: if a set of languages with no recent genealogical history nor traces of areal contact tend to share the same pair of properties again and again, then there seems to be something about the properties of probable languages in general.

While venerable and potentially useful, this procedure is complicated by many factors. First, the nature of a proposed dependency can affect how the pattern of observations translates into support for the dependency. In the first section, we show how different notions of causality and causal strength are appropriate for different types of dependencies involving two variables. Secondly, these dependencies can be distorted not only by historical relations between languages (as usually acknowledged in the literature) but also due to complex causal dependencies involving multiple variables. Addressing these concerns requires appropriate formalisms and statistical techniques. These exist and are widely used

Damián E. Blasi & Seán G. Roberts. 2017. Beyond binary dependencies in language structure. In N. J. Enfield (ed.), *Dependencies in language*, 117–128. Berlin: Language Science Press. DOI:10.5281/zenodo.573774

for addressing the problem of historical relations (which we cover in the second section), but methods for dealing with relationships between more than two variables are underdeveloped in linguistics. In the final section, we discuss some new approaches to detecting causal dependencies between more than two variables.

1 Probability and causation

There exist several possible formalizations of the concept of causality inspired in concepts from mathematics, logic, computation and philosophy (see Fitelson & Hitchcock 2011). For the kind of regularities and laws governing the language sciences causation appears more naturally described in terms of probabilities.

For the sake of simplicity, we will be dealing in these examples with a hypothesized cause (C) and an effect (E). These will be expressed in terms of total probabilities of the cause or the effect to occur ($P(C)$ and $P(E)$ respectively) and the related conditional probabilities (such as the probability of the effect occurring given that the cause is present $P(E|C)$, or the probability of the effect occurring given that the cause is absent $P(E|{\sim}C)$). In this context, we can think about causation as probability raising: the probability of the effect taking place is larger when the cause is present than when the cause is absent, $P(E|C) > P(E|{\sim}C)$.

It is critical to remark that these probabilities and the measures of strength are used as a way of thinking about causal relations instead of definitions suitable for statistical analysis. Identifying probabilities with type frequencies and determining causal dependencies by attesting patterns in language counts can be problematic, and as such the structure of the models we use to think about the data and the data themselves (and their statistical properties) should be always clearly distinguished.

Typically, probabilities are equated to frequencies of occurrence when the statistical assessment takes place. $P(E)$ is approximated to the proportion of times the cause is observed to occur compared to not occurring, and $P(E|{\sim}C)$ to the proportion of times the effect is observed when the cause is absent. For instance, given the contingency table in 1,

(1)

	C	∼ C
E	10	5
∼E	5	25

we could readily estimate $P(C)$=15/45=1/3 and $P(E|{\sim}C)$=5/30=1/5. This is the usual practice in the field, but it hides a number of assumptions about what is tested and the nature of the sampling process.

First of all, the strategy of counting languages has been used sometimes to say something about probable languages in general and not about the particular relations that hold in the necessarily contingent set of surveyed languages. This is as fundamental as it is uncontroversial and pervades scientific practice, and in particular the language sciences – we infer general properties of cognition from a limited sample of experimental participants and we determine the usage properties of words from samples of text that are diminishingly small in comparison to what is regularly produced by speakers.

In consequence, we assume that the frequency measured in a given set of typological data matches, in some way, the likelihood of picking at random any likely human language and finding that it has a certain property. This becomes explicit in the linguistic typology literature: in the absence of mechanisms or constraints shaping the structure of the grammar, we "would expect each type to have roughly an equal number of representatives" (Comrie 1989). The issue stems from the fact that what "roughly" means here is left unspecified and to a large extent at the discretion of the researcher. In fact, any reasonable sampling model will generate observable differences in the proportions even when no effect is present (Cysouw 2010). Specific distributions of typological variables have been motivated observationally (Nichols 1992), based on concrete models inspired by principles of language change (Cysouw 2010) or borrowed directly from the toolkit of machine learning, the Dirichlet process being a particularly popular choice that is plastic enough as to reflect our lack of certainty (Daumé III 2009; Piantadosi & Gibson 2014).

Assuming for a moment now that we do have access to the true probabilities of causes and effects and their relation (perhaps via a careful consideration of the observed frequencies), let us consider now the two simplest cases of causal relations between C and E (illustrated in Figure 1). Greenberg's seminal work on implicational typological universals already presented a binary classification of dependencies into which we will tap due to its popularity (Greenberg 1966, see Culbertson, this volume Cristofaro, this volume).

Some of Greenberg's universals are bidirectional implications, such as the order of adposition and noun implying the order of genitive and noun, and vice versa. Bidirectional implications contrast with unidirectional implications, which allow the possibility of the effect being present without the cause, but the cause makes the effect more probable. For instance, Greenberg suggested that lan-

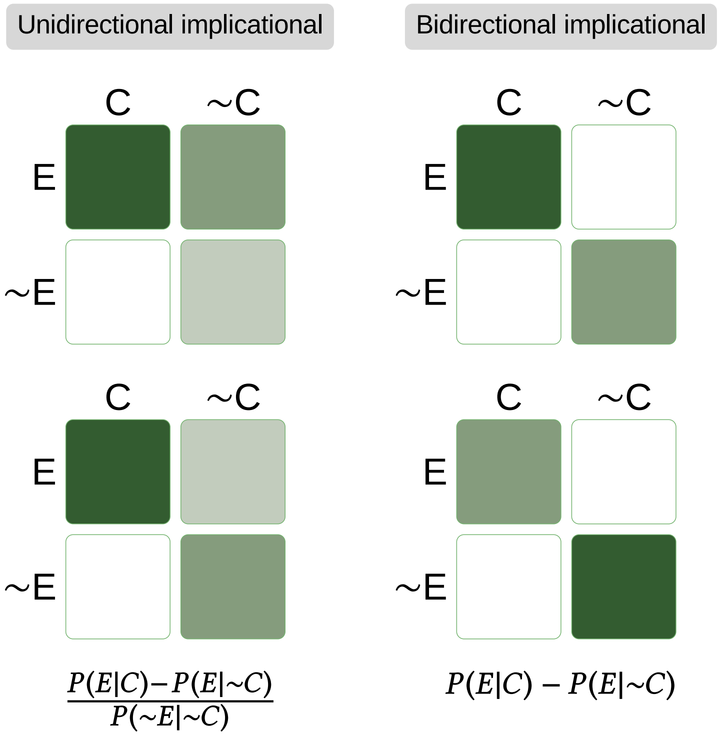
Unidirectional implicational
Bidirectional implicational
C
~C
E
~E
$\frac{P(E|C)-P(E|{\sim}C)}{P({\sim}E|{\sim}C)}$
$P(E|C) - P(E|{\sim}C)$

Figure 1: Contingency tables that maximise different measures of causal strength when language type frequencies are equated to type probabilities. On the left are two tables which maximize unidirectional implications and on the right are two tables which maximize bidirectional implication. More intense colour stands for more cases attested with those properties; cells in white represent no counts. The formulas for different notions of causal strength appear at the bottom.

guages with VSO canonical order tend to be prepositional, though this does not claim that all prepositional languages will be VSO: prepositions occur with virtually all other word order combinations, prominently SVO.

While these ideas are intuitive, the formalization of causal strength by means of probabilities sheds light on the kinds of evidence that are needed in order to put forward any claim about causal influence. For the sake of convenience, causal measures are often defined in such a way that 1 stands for the strongest causal relation and 0 for the absence of any evidence of a causal effect, with intermediate values reflecting strengths between these extremes. To start with, Eells (1991)'s view of causal strength captures adequately the causal strength underlying a bidirectional implication, which is defined as:

(2) $CS_e = P(E|C) - P(E|{\sim}C)$

That is, the change in the probability of the effect when the cause is present and when the cause is absent. The largest difference (CS_e=1) will be achieved when the cause deterministically triggers the effect ($P(E|C)$=1) and where the absence of the cause also implies the absence of the effect ($P(E|{\sim}C)$=0) – as represented in Figure 1. On the other hand, when the cause does not change the probability of the effect occurring ($P(E|C)$=$P(E|{\sim}C)$), Eells' measure of causal strength is minimised (CS_e=0). Notice that the strength of the assertion of a bidirectional implicational universal does not rely on the relative frequencies of each type, i.e. $P(C)$ and $P(E)$ and their complements.

On the other hand, unidirectional implications do not make any predictions with respect to the case in which the cause is absent. $P(E|{\sim}C)$ could be close to either 1 or 0 without affecting our confidence on the efficacy of the cause – e.g. that smoking leads convincingly to cancer is independent of the fact that cancer might arise due to other factors as well. However, rather than using the plain conditional probability as a measure of the causal strength of a unidirectional implication ($P(E|C)$) the probability $P(E|{\sim}C)$ plays the role of a baseline to compare against. Thus, a good normalized measure of causal strength for unidirectional implications would be one that (1) becomes 0 when the cause does not make the effect more or less probable than its absence and (2) is 1 only when the cause yields the effect determinstically ($P(E|C)$=1). This leads to none other than Cheng (1997)'s notion of causal strength:

(3) $CS_c = \left[P(E|C) - P(E|{\sim}C)\right] / P({\sim}E|{\sim}C)$

That is, the causal power increases as we observe the effect with the cause and decreases as we observe the effect with the cause, but only to the extent that we also observe no effect without the cause.

In contrast to the idea that causality constitutes a monolithic phenomenon, there are many other approaches to the notion of causal strength (see Fitelson & Hitchcock 2011), each one being suitable for the study of different dependencies. The notion of causal measure will also impact the strategy of inference of the involved probabilities. For example, a unidirectional implication could be assessed by collecting data only on languages which are known to exhibit the cause, while a bidirectional implication requires knowing about languages both with and without the cause.

2 Moving towards statistical support

The formalisms above rely on knowing the real probabilities of each cell in the contingency table. The question of practical interest, then, is how to make a statistically valid case for a dependency based on language counts. These counts might differ considerably from the true probabilities since simple co-occurrence in a sample of data does not guarantee dependency. The most well-known sources of inflated co-occurrences without substantial causal links are shared history or contact. For instance, in the Mesoamerican linguistic area, languages frequently display a vigesimal numeral system and they lack switch-reference, traits that distinguish them from neighbouring languages (Campbell, Kaufman & Smith-Stark 1986). A contingency table displaying the number of languages in that region would give the impression that both variables are associated, which will be simply reflecting the fact that those traits have been transmitted together all the way down the genealogical tree or horizontally from other language(s). This confound – known as Galton's problem – applies to any study trying to detect causal connections between traits in languages. Roberts & Winters (2013) demonstrate how pervasive this problem can be by finding co-occurrences between traits with no causal dependencies between them.

These problems can be overcome if the history of contact between languages is taken into account. For example, bidirectional implications can be easily captured by the many regression methods available. Jaeger et al. (2011) recommend a mixed effects model framework so as to be able to account for areal and genealogical dependencies as random effects for that purpose. Another alternative is to use explicit phylogenetic information and map branch lengths to covariance (so languages that diverged more recently in time are expected to have more similar feature values) (Verkerk 2014). The Family Bias method (Bickel 2013) continues the tradition of comparing larger linguistic groupings in a regular regression setting (without any special specification of the covariance between languages) but instead infers the biases of the groupings by assessing the evidence in favour or against one particular typological variant (or set of variants). The literature on the statistical assessment of unidirectional implications is much less restricted, however. Researchers have devised ways of resolving this issue within the frequentist (Everett, Blasi & Roberts 2015) and Bayesian traditions (Daumé III & Campbell 2009).

Another way that co-occurrence probabilities can be distorted, and one that is rarely addressed, involves more complicated causal dependencies. The statistical methods mentioned above become limited when more than two variables are taken into account at a time and indeed, perhaps as an implicit acknowledgement of this difficulty, most typological generalizations are limited to pairs of variables rather than more complex constellations.

Let us see more precisely how complex dependencies might yield spurious dependencies by considering the simplest possible case beyond the two-variable case, which is naturally when there are three variables causally linked in some way. If we regard causal relations graphically as arrows going from the causes to the effects, then this setting will correspond to any of four different possible arrangements depicted in Figure 2.

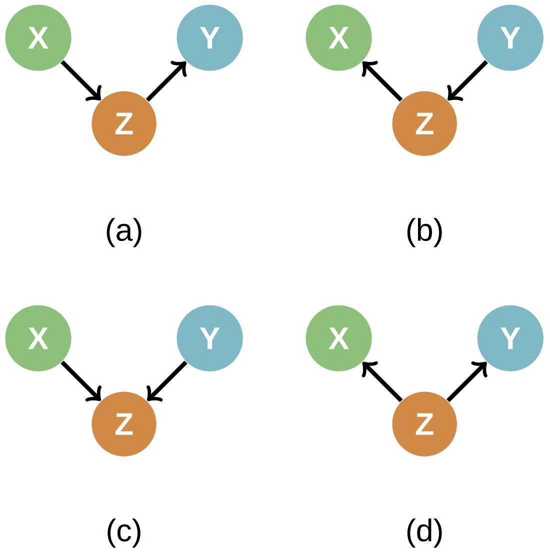

Figure 2: Four possible (non-trivial) ways in which variables X, Y and Z could be causally linked. Arrows represent the flow of causality, so that an arrow pointing from X to Z indicates that changes to X cause changes in Z.

The first two cases (a and b in Figure 2) correspond to Z simply serving as a communicator of the effect of X on Y or vice versa. For instance, it has been suggested that population size and morphological complexity are causally connected via the number of L2 speakers (Lupyan & Dale 2010): the larger the population (X), the more likely it is that the language comes into contact with other languages, increasing the number of L2 speakers (Z) which act upon the language by simplifying the system of its morphological structures (Y).

The third possibility is that Z is causing both X and Y (d in Figure 2), so the observed causal link between the two is an artifact of Z being a common cause.

As an example, many languages of the world have noun classes (X) also have applicative voice (Y) (Aronoff & Fudeman 2011). The common cause behind the joint occurrence of these features is that many of these languages come from the Atlantic-Congo family (Z), one of the largest linguistic families.

Finally, it could be that both X and Y contribute jointly to cause Z (c in Figure 2). Languages with isolating morphology (X) will naturally have shorter words in average (Z), and the same is true for languages with tones (Y).

The qualitative Greenbergian implications presented before had a transparent formal counterpart and they can be evaluated statistically with well established methods. However, the discussion and evaluation of dependencies involving three or more variables become increasingly unsuitable without a proper formalization. The probabilistic framework discussed at the beginning finds a justification at this point. In addition to it, we need to briefly review some definitions and concepts from graph theory (see Pearl 2009).

A graph consists of a set of nodes and a set of edges. Directed edges bind two nodes in an asymmetrical fashion – so if A and B are nodes, either A⟶B or A⟵B. A sequence of nodes from A to B where each adjacent pair is bound by a directed edge going from the first to the second member is referred to as a path between A and B. A path that starts and finishes in the same node is referred to as a cycle. A directed graph is one in which all edges are directed, and a directed graph with no cycles is called a directed acyclic graph (DAG).

The set of nodes that can be reached through a path from A are A's descendants, and the nodes that are directly connected to A such that their common edge points to A (like B⟶A) are the parents of A. In DAGs there are no paths which go from a descendant back to one of its parents.

This graphical framework allows a straightforward visualization of causal connections between variables. Variables are represented as nodes and causal relations (of any kind discussed in the binary case) are represented as directed edges, so A⟶B will be read as “A causes B”. The assumption linking this graph representation to the ideas of probabilistic causation discussed before is that of the Markov Causal Condition. If two variables are dependent and one is not a descendant of the other then their dependency can be explained away by appealing to a common ancestor of the pair. Put another way, a variable is only affected by its immediately connected (ancestor) causes.

Embracing this representation of the relations in the data opens up new statistical possibilities. One that partially relies on regression is to use structural equation models (Duncan 2014). Structural equation modelling is a cover term for a number of techniques that allows the testing of more or less well-specified functional dependencies between variables as embedded in DAGs. To take a very basic example (based on a specific case of structural equation modelling called path analysis), suppose that we want to decide between situations (a) and (b) of Figure 2. Assuming that we are in possession of good guesses about what could be the functional dependencies, we then could contrast the model fit (how well the model predicts the observed data) between (a) and (b). The possibilities provided by structural equation modelling include the inclusion of hidden variables and non-parametric functional dependencies.

In cases where uncertainty about the correct model is high, model comparison might not be the best ally. In those cases, it is possible to appeal to the predictions that come "for free" by assuming the Markov Causal Condition along with the DAG. The idea is that the Markov Causal Condition entails a series of conditional dependency statements involving the variables, and that given appropriate conditions it is possible to estimate the most likely underlying causal graph from observational data. There are multiple methods for doing this (Shalizi 2013), a popular efficient and computationally inexpensive method being the PC algorithm (Spirtes, Glymour & Scheines 2000; Kalisch et al. 2012). These techniques are only starting to be explored by researchers in the language sciences (Blasi et al. 2018; Baayen, Milin & Ramscar 2016).

3 Conclusion

The inference of causal dependencies based on surveys of languages has a long history in the field. This methodology faces several complications, like the difficulty of estimating probabilities from counts of languages or the lack of consideration of higher-order dependencies between multiple variables. Methods and formalisms based on probability can address these problems, and help linguists to better test and think about the nature of dependencies in language.

Acknowledgements

We would like to thank the participants of the Dependencies workshop, and in particular Jennifer Culbertson for helpful feedback. SR is supported by an ERC Advanced Grant No. 269484 INTERACT to Stephen Levinson. We thank the Max Planck Society for additional support.

References

Aronoff, Mark & Kirsten Fudeman. 2011. *What is morphology*. Hoboken, New Jersey: John Wiley & Sons.

Baayen, R. Harald, Petar Milin & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiology* 30 (11). 1174–1220.

Bickel, Balthasar. 2013. Distributional biases in language families. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency*, 415–444. Amsterdam: John Benjamins Publishing Company.

Blasi, Damián E., Seán G. Roberts, Marloes Maathuis & Emmanuel Keeulers. 2018. *Inferring causality in lexical properties from observational data.*

Campbell, Lyle, Terrence Kaufman & Thomas C. Smith-Stark. 1986. Meso-America as a linguistic area. *Language* 62(3). 530–570.

Cheng, Patricia W. 1997. From covariation to causation: A causal power theory. *Psychological Review* 104(2). 367.

Comrie, Bernard. 1989. *Language universals and linguistic typology: Syntax and morphology*. Chicago: University of Chicago Press.

Cysouw, Michael. 2010. On the probability distribution of typological frequencies. In Makoto Kanazawa, András Kornai, Marcus Kracht & Hiroyuki Seki (eds.), *The mathematics of language*, 29–35. New York: Springer.

Daumé III, Hal. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the Association for Computational linguistics*, 593–601. Association for Computational Linguistics.

Daumé III, Hal & Lyle Campbell. 2009. A Bayesian model for discovering typological implications. *arXiv*. preprint arXiv:0907.0785.

Duncan, Otis Dudley. 2014. *Introduction to structural equation models*. Amsterdam: Elsevier.

Eells, Ellery. 1991. *Probabilistic causality*. Vol. 1. New York: Cambridge University Press.

Everett, Caleb, Damián E. Blasi & Seán G. Roberts. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences* 112(5). 1322–1327.

Fitelson, Branden & Christopher Hitchcock. 2011. Probabilistic measures of causal strength. In Phyllis McKay Illari, Federica Russo & Jon Williamson (eds.), *Causality in the sciences*, 600–627. Oxford: Oxford University Press.

Greenberg, Joseph H. (ed.). 1966. *Universals of language*. Cambridge: MIT Press.

Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15(2). 281–320.

Kalisch, Markus, Martin Mächler, Diego Colombo, Marloes H. Maathuis & Peter Bühlmann. 2012. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* 47(11). 1–26.

Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PloS One* 5(1). e8559.

Nichols, Johanna. 1992. *Linguistic diversity in space and time.* Chigago: University of Chicago Press.

Pearl, Judea. 2009. *Causality.* New York: Cambridge University Press.

Piantadosi, Steven T & Edward Gibson. 2014. Quantitative standards for absolute linguistic universals. *Cognitive Science* 38(4). 736–756.

Roberts, Seán G. & James Winters. 2013. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PloS One* 8(8). e70902.

Shalizi, Cosma Rohilla. 2013. Advanced data analysis from an elementary point of view. http://www.stat.cmu.edu/cshalizi/ADAfaEPoV/13.

Spirtes, Peter, Clark N Glymour & Richard Scheines. 2000. *Causation, prediction, and search.* Cambridge, MA: MIT press.

Verkerk, Annemarie. 2014. The correlation between motion event encoding and path verb lexicon size in the Indo-European language family. *Folia Linguistica Historica* 35. 307–358.