Causation in Semantics and Grammatical Structure

Week 11: Structural equation models

Prerna Nadathur

December 19, 2019

Modeling causation: recap

Two main types of causal theory:

- **dependency theories**
  causal relationships are formal, abstract dependencies between objects (events), which may or may not be reducible to non-causal descriptions
  - logical dependency
  - counterfactual theories (Lewis 1973)
  - probability raising models
  - network models
- **production theories**
  "something more than a correlation or regularity is involved in causation" (Copley & Wolff 2014, p.23)
  - transmission theories (energy, momentum, conserved quantities)
  - force dynamics

Dependency theories: Lewis 1973

Main argument: a theory of causal dependence based on counterfactual necessity does better than *regularity* (sufficiency) theories on:

(A) distinguishing causes and effects,

(B) ruling out epiphenomena as potential causes

(C) cases of (early) pre-emption

> "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it." p.557

- roughly, to check a counterfactual dependency between actual events, cause C and effect E, we move to the *closest* world in which C does not occur, and check whether or not E occurs

Dependency theories: Lewis 1973

Lewis:

- we get around the problems of logical-necessity theories by considering causal chains
- causal dependence is equated with **stepwise counterfactual dependence** between events in a causal chain:

$$A \mathrel{\Box\!\!\rightarrow} B \mathrel{\Box\!\!\rightarrow} C \mathrel{\Box\!\!\rightarrow} D \mathrel{\Box\!\!\rightarrow} E$$

- since each step in the chain supports counterfactual dependence ($\Box\!\!\rightarrow$), E is **causally dependent** on A (for Lewis, A *causes* E)
- we noted that this is (perhaps inadvertently) a claim about the meaning of *cause*, not about causation writ large

Dependency theories: Lewis 1973

Stepwise counterfactual dependence allows us to deal with cases of **early pre-emption:**

(1) Billy and Suzy both have rocks to throw at a bottle. Billy will only throw if Suzy does not. Assume that they both have perfect aim. Suzy throws her rock and hits the bottle, so Billy does not throw his rock. Suzy's rock breaks the bottle.

  a. Billy's throw is pre-empted by Suzy's

  b. *intuition:* Suzy *caused* the bottle to break

  c. *problem:* If Suzy had not thrown her rock, the bottle would still have broken

$$\text{Suzy throws } \not\Box\!\!\!\rightarrow \text{ the bottle breaks}$$

- stepwise dependence to the rescue:

$$\text{Suzy throws } \Box\!\!\!\rightarrow \text{ Suzy's rock flies through the air } \Box\!\!\!\rightarrow \text{ Suzy's rock hits the bottle } \Box\!\!\!\rightarrow \text{ the bottle breaks}$$

Dependency theories: Lewis 1973

But it does not solve the problem of **late pre-emption:**

(2) Billy and Suzy both throw stones at a bottle on the wall. Their throws are on target. Suzy throws first, so her stone reaches the bottle before Billy's, and the bottle breaks.

a. Billy's rock hitting the bottle is *pre-empted* by Suzy's

b. *intuition:* Suzy caused the bottle to break, Billy did not

c. *problem:* Suzy throws $\not\Box\!\!\rightarrow$ the bottle breaks

- breaking the steps in the causal chain down does not restore counterfactual dependence

Suzy throws $\Box\!\!\rightarrow$ Suzy's rock flies through the air $\Box\!\!\rightarrow$ Suzy's rock hits the bottle $\not\Box\!\!\rightarrow$ The bottle breaks

  - if Suzy's rock had not hit the bottle, Billy's rock would have hit the bottle, and the bottle would still have broken

Dependency theories: network models

Causal network models: basic ideas (from last week)

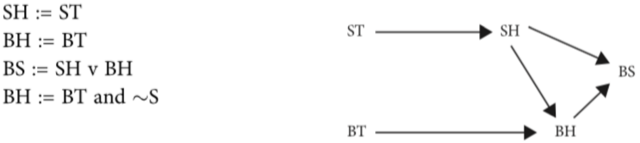
- causal relationships can be represented as a directed graph (a set of nodes with arrows linking them)
- events are the nodes in the graph (cf. Davidson 1969)
- arrows are indicators of causal influence (and directionality)
- causal network models can be accompanied by **structural equations** specifying the nature of dependencies

Positives:

- a network can represent more complex causal relationships than a (stepwise) chain of causation
- 'intervention' tests in network models capture some of Lewis's intuitions about similarity (and preserving facts vs. causal laws)

Causal network models

The late pre-emption example in a (deterministic) network model:



SH := ST
BH := BT
BS := SH v BH
BH := BT and ∼S

Fig. 2.2 An example of late pre-emption in terms of a direct graph (Hitchcock 2010)

- the network model allows us to include the influence that Suzy's hit (SH) has on Billy's hit (BH)
- Suzy's throw (ST) only influences Suzy's hit (SH), but the value of SH determines the value of BH (assuming we don't change the facts about whether or not Billy throws)
- this explains why we think Suzy's throw is the cause and Billy's throw is not: there's an asymmetry in the arrows in the model

Today: A closer look at network models

Causal network models

Sloman (2005) presents an overview of network/structural equation models as a tool in cognitive science (psychology, computer science, linguistics, philosophy):

- in their practical uses, the idea is that a representation of causal relationships can allow us to replicate or model human behaviour in key areas of cognition (decision-making, planning, etc)
- **Sloman's basic assumption:** the key to understanding human behaviour is that humans are agents, who "actively pursue goals", and "talk, think, and act as if [we have free will]":

> " . . . to understand the mind requires a way of representing agency." p.5

Causation and agency

What does it mean to be an agent?

- for Sloman, *agency* is the ability to intervene on the world to change it (this is familiar!)
  - specifically, our cognitive representations involve entities with this ability
- our knowledge about intervening in the world is knowledge of causal relationships: which inputs produce which outputs

Upshot: "agency is the ability to represent causal intervention"

- the consequence of humans cognitively representing themselves as agents is that we have cognitive representations of causation

Causation, events, and agency

Some of the broader cognitive-science ideas are familiar to ideas we've seen crop up in linguistic investigations of causation:

> "...causality is a form of construal. We impose a causal frame on the world to understand it, a frame that tells us about the mechanisms that not only produced the world as it is but that (counterfactually) govern the world however it had turned out." p.6

Judgements about alternative situations and counterfactuality are relevant to causation

- causation allows us to make reasonable guesses about how the world will turn out based on certain hypothetical (counterfactual facts), because we expect causal laws to be a constant

Causation, events, and agency

Causal frames allow us to make sense of the actions of others and the consequences of our own actions:

- the type of implementation Sloman describes is a **Bayesian network** some of the ways that network models work draw on Bayes's work on probability theory

> "What's unique about the causal model framework is that it gives us a way to think about the effect of action in the world. It makes the claim that the cognitive apparatus that people use to understand the world has a specialized operation that encodes certain changes as the effect of agency, of intervention." p.7

Causal models: central assumptions

The causal model framework is based on the idea that, in looking at a complex system, we learn how it works by paying attention to what is **invariant**:

- *Example:* when learning to drive, we learn about the relationships between motion and pedals, and orientation and direction of motion
- we don't make any links between how to drive and the colour of the vehicle or the material of the seats
- getting into a car of a new colour doesn't affect our ability to drive because this knowledge isn't represented as invariant with respect to cause and effect relationships
- one piece of evidence that humans DO represent the world causally comes from **selective attention** to invariants
- expertise works like this: an expert mechanic knows which parts of a system to examine, because s/he knows about the invariant causal relationships in the system

Events and causation

> "In the domain of events, causal relations are the fundamental invariants." p.17

- although we're now trying to model world understanding, not just language, this is basically the same as Davidson's point
- causal relations can be identity conditions for events because of this notion of invariance
- scientific study is also based on the idea that systems have invariants
    - when we manipulate an *independent variable* and measure the change in an associated *dependent variable*, the 'dependency' we are looking at is causal
    - the correlation in variation between an independent and dependent variable is the invariant in this system (not the values themselves)

Events and causation

Another familiar assumption: causes and effects are events

- causal relationships involve change over time
- changing a cause asymmetrically produces (makes more likely) a change in the effect
- causes and effects pattern together over time (i.e., they are statistically learnable)
- "causal relations relate entities that exist in and are therefore bounded in time – events" (compare Davidson)

This information allows us to do experiments:

- based on the asymmetry, we expect that causes can be tweaked to produce changes in their effects
- we 'wiggle' (intervene) on something and then measure changes in events that we think are **causally downstream**
- we don't expect anything upstream of our intervention to change: if it does, this prompts new causal hypotheses

Causation and counterfactuals

> "To say that *A* caused *B* seems to mean something like the following: *A* and *B* both occurred, but if event *A* had not occurred (and *B* had no other sufficient cause), *B* would not have occurred either." p.24

- this is an update w.r.t. Lewis: it takes into account the possible existence of other causes
- a network (as opposed to a causal chain) representation can make sense of this type of reasoning
- Sloman: the counterfactual relationship (however local) is what distinguishes causation from correlation (echoes Lewis)
- counterfactual conditionals are about what *would have happened* if some intervention had been made or not made
- statistical learning is an important assumption here: we can never gather *all* possible data, so we can never be certain about causal relationships
  - Hume: seeing that the sun comes up every morning of your life does not guarantee that it will come up the next day

Problems with counterfactuals, revisited

The pre-emption examples:

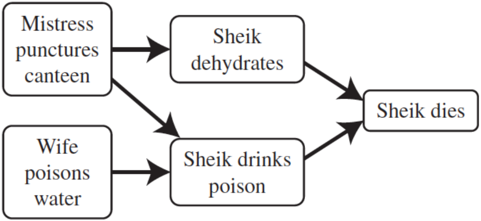
- recall that these examples showed that counterfactual dependence is not necessary for a causal judgement
- the no-other-sufficienct cause condition helps with pre-emption, but causes other problems

(2) **The poisoned water scenario.** A sheikh's wife wants to kill him, and so does his mistress. The sheikh is going on a long desert journey, so the wife puts poison in his water canteen. The mistress makes a small hole in the canteen so that the water drips out. The sheikh dies of thirst in the desert.

- if we consider each potential cause in the absence of alternative sufficient causes, then BOTH are causes of the king's death
- but we only judge the mistress to be guilty of murder

Counterfactuals and networks

A graph representation of the poisoned-water scenario:

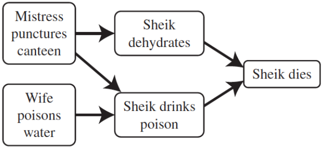


- there's an asymmetry in the relationship between the wife's action and the sheik's death as compared to the relationship between the mistress's action and the sheikh's death
- crucially, the puncture causally affects the potential drinking-poison event: it prevents it
- note that we don't know that the link is preventative from the graph itself, but from the story
    - (so, we still need a way of representing the link type)

Counterfactuals and networks

Network models revise the counterfactual relationship:

- event C actually causes event E if
  1. both C and E occur
  2. if C had not occurred, then E would not have occurred, even when all events not on the path from C to E are assumed to have occurred



- if everything not on the path from the poisoning to the death occurred, but the wife had not poisoned the water, the sheikh would still have died
- thus, the wife's actions did not **actually cause** the death

*There is a good deal of literature devoted to developing an accurate definition of **actual cause**: Pearl (2000), Halpern & Pearl (2001, 2005)

Causal network models

Network models **do not define causation**:

- the arrows in the graph represent causal links, but the arrows are not analyzable

The make-up of a model:

1. The system being represented [world]
2. A set of **structural equations** (probabilistic or deterministic) [algebraic model]
3. A **directed acyclic graph** (DAG) [graphical model]
    - a **graph** is a network of nodes and links (a mathematical object)
    - it's **directed** because the arrows have directionality
    - we assume that causality is **acyclic**

Modeling a real-world system

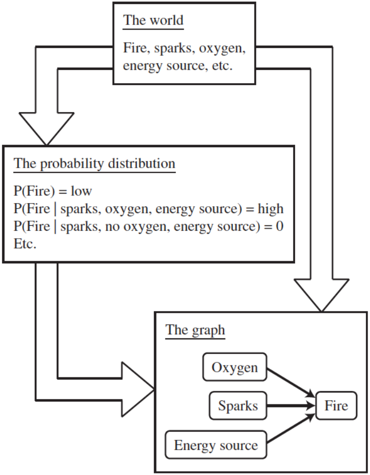
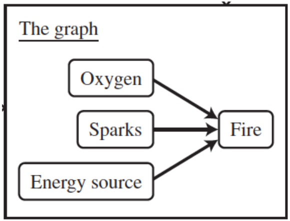

The world
Fire, sparks, oxygen, energy source, etc.
The probability distribution
P(Fire) = low
P(Fire | sparks, oxygen, energy source) = high
P(Fire | sparks, no oxygen, energy source) = 0
Etc.
The graph
Oxygen
Sparks
Energy source
Fire

Figure 4.2

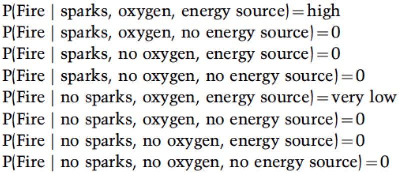
- probability distribution:
  - marginal (raw): e.g., Pr(fire)
  - conditional: e.g., Pr(fire given other conditions)
- the graph constrains the probability distribution:
  - the probability of fire is conditional on the probabilities of oxygen, sparks, and energy
  - this is reflected in the presence of directed arrows; if there's no arrow between two nodes, they are **independent**

Structural equations



- looking at the graph, we know that whether or not fire occurs depends on oxygen, sparks, and an energy source

- the "Fire" node needs to come along with a set of equations to specify the nature of the causal relationships, for instance:



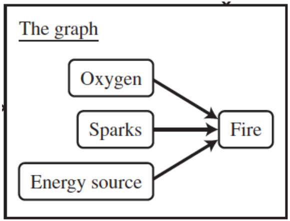
P(Fire | sparks, oxygen, energy source) = high
P(Fire | sparks, oxygen, no energy source) = 0
P(Fire | sparks, no oxygen, energy source) = 0
P(Fire | sparks, no oxygen, no energy source) = 0
P(Fire | no sparks, oxygen, energy source) = very low
P(Fire | no sparks, oxygen, no energy source) = 0
P(Fire | no sparks, no oxygen, energy source) = 0
P(Fire | no sparks, no oxygen, no energy source) = 0

- this distribution specifies the likelihood of fire for every possible combination of causes

Structural equations

We can also have **deterministic** structural equations, as in the pre-emption example:

- suppose fire is guaranteed in the presence of its three **causal ancestors,** and impossible otherwise



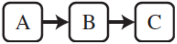
$$\textsc{Val}(\text{fire}) := \textsc{Val}(\text{oxygen}) \wedge \textsc{Val}(\text{sparks}) \wedge \textsc{Val}(\text{energy}),$$
$$\text{where } \textsc{Val}\colon \text{Events} \to \{T, F\}$$

- a probabilistic, as opposed to a deterministic, model leaves room for uncertainty
- (but there are ways to include uncertainty in deterministic models as well)

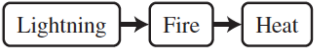
Dependence and independence

We can have different types of dependencies in a graph:

- chain:



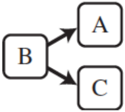
A and *C* are dependent because *A* indirectly causes *C*



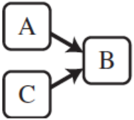
…you're more likely to have heat if you have lightning than if you don't (causal)

  - *A*, *C* are **conditionally independent**, given *B* (once we fix *B*, they aren't connected; *B* screens off *C* from *A*)

- fork:



- inverted fork (collider):

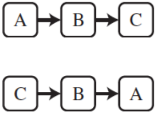


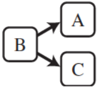
- *B* is a common cause or a common effect, but *A* and *C* are independent

Dependence and independence

Dependency is reflected in correlation, but correlation cannot fully define a causal model:

- *A* and *B* can be correlated in any one of the following situations (and others)





- we need to do more than make observations
- we make two assumptions when we infer causal structure based on the world:
  1. **causal Markov condition**: the direct ancestors of a node screen it off from all nodes except its direct descendants
  2. **stability:** we assume that correlations are real, not matters of chance (sometimes we're wrong)

Bayesian networks

These assumptions are what make a causal graph **Bayesian**:

- to work out the truth value/probability of a given event, we just need to consider the truth values/probabilities of its immediate ancestors
  - this is a consequence of **Bayes's theorem** (see Ch. 5) which says that we can calculate the *posterior* probability (probability given some other event A) of an event B by considering the prior probabilities of A and B and the probability of A, given B

$$\Pr(B|A) = \frac{\Pr(A|B) * \Pr(B)}{\Pr(A)}$$

- thus, having the graphical structure greatly simplifies the information we need to specify a probability distribution over a set of event variables
- by representing directionality, it also gives us a way of making tests of the system and our hypotheses about it

Intervention in causal networks

If we observe a correlation between two events, we can test for causal structure by running experiments:

- suppose you notice that every time I touch a certain place on the wall, the light goes off
- it's possible that it's a coincidence, but you can test this by touching the wall yourself and seeing what happens

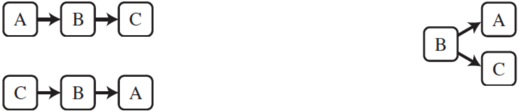
Experiments and 'surgery':

- Pearl's *do* operator represents intervention on a causal system: $do(X = x)$ resets the value of X to a value x of choice
- after an intervention, X is no longer informative about its causal ancestors, but the change is expected to change its causal descendants
- so, if we 'change' X and Y changes, then we suspect that Y is causally 'downstream' of X

Intervention in causal networks

At the cognitive level, interventions are about counterfactual reasoning: *if you'd wiggled A, then B would have changed*

- (we can counterfactual conditionals with respect to a causal network graph, rather than trying to define causal relations in terms of counterfactuals – see Schulz 2011, Ciardelli et al 2018, others)
- we can use interventions to determine the nature of a correlation:



- to determine which model is relevant, we need to independently 'wiggle' *A*, *B*, *C*, and measure any changes
- this is often just done at the level of common-sense reasoning when we're observing the world (we never think that a light coming on causes the switch to change)

Causal network models

Causal network models make a lot of simplifying assumptions:

- we don't always understand the precise method of connection between a cause and an effect
- we often leave out factors that seem relatively stable (e.g. whether or not a circuit is shorted)
- what gets put into the model or left out depends on the level of granularity we're interested in

But these representations capture some patterns of reasoning:

- directed arrows capture the idea that causal relationships are asymmetric: effects do not diagnose their causes
- interventional 'experiments' build in the notion that effects do not precede their causes (and some models actually incorporate time; Schulz 2007)
- this approach to counterfactuals explains Lewis's observation that preserving facts is more important for similarity than preserving the 'backwards integrity' of causal relations

Causal networks and counterfactuals

Last week: transitivity fails in counterfactual reasoning

(3) a. If Hoover had been born in Russia, he would have been a communist

b. If Hoover had been a communist, he would have been a traitor in the eyes of the US government

c. ⊭ If Hoover had been born in Russia, he would have been a traitor . . .

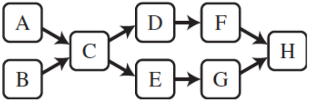
- (3a): we 'intervene' on Hoover's birthplace; consequently he becomes a communist
- (3b): we intervene only on Hoover's political stance, leaving him American (*even if* it's more likely that he would have been born outside of the US, given his political sympathies)
- but in (3c), we go back to an earlier point, so we lose the facts that are relevant to Hoover's being a traitor

Intervention explains the failure of transitivity

Causal networks and linguistic causation

Importing causal network models into linguistics is not strictly about experimentation, but about comparing the kind of reasoning we do in a network model to the linguistic judgements we make:

- counterfactual reasoning is one area that suggests that language cares about network representations
- a network model can be quite complex graphically:



- based on the structural equations associated with a model, the dependencies can be of different *types*:
  - $C := A \wedge B$, but $H := F \vee G$
  - the relationship between A and C is importantly different from the relationships between F and H

Causal networks and lexical semantics

The main ideas:

- the different relationships and configurations that can be encoded in the structure of a causal model have 'cognitive reality'
- at the linguistic level, we expect these differences to be reflected in linguistic judgements
- . . . and the causal language we use or accept as an appropriate description of a system or situation
- last week, we saw that some of the difficulty philosophers have had in defining causation is because they (unintentionally) were only looking at uses of the word *cause*
- next week, we'll look at an approach to representing different causatives in terms of different network configurations
  - (and consider how the other kinds of causal parameter (agency, intention, volition) fit into a network picture)

References

1. Copley, B. & P. Wolff. 2014. Theories of causation should inform linguistic theory and vice versa. In B. Copley & F. Martin, eds., *Causation in Grammatical Structures,* 11–57. Oxford: Oxford University Press.
2. Davidson, D. 1969. The individuation of events. In N. Rescher, ed., *Essays in Honor of Carl G. Hempel.* Dordrecht: Springer.
3. Dowty, D. 1979. *Word Meaning and Montague Grammar*. Dordrecht: Reidel.
4. Lewis, D. 1973. Causation. *The Journal of Philosophy,* 70: 556–567.
5. Halpern, J. & J. Pearl. 2001. Causes and explanations: a structural-model approach. Part I: Causes. *Proceedings of 17th Conference on Uncertainty in Artificial Intelligence*, 194–202.
6. Pearl, J. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge: Cambridge University Press.
7. Schulz, K. 2007. Minimal models in semantics and pragmatics: free choice, exhaustivity, and conditionals. Ph.D., Universiteit van Amsterdam.
8. Schulz, K. 2011. If you'd wiggled A, then B would've changed. *Synthese* 179: 239–251.
9. Sloman, S. 2005. *Causal Models: How People Think About the World and Its Alternatives*. Oxford: Clarendon Press.