A note on co-occurrence, transitional probability, and causal inference

Jeroen van Paridon* Phillip M. Alday

Max Planck Institute for Psycholinguistics, The Netherlands

Abstract

Much has been written about the role of prediction in cognition in general, and language processing in particular, with some authors even positing that prediction is the central *goal* of cognition. Attributing a specific *goal* to cognition is speculative, yet common theories of cognition posit that prediction plays a role in both perception and action. However, in studies on language processing, measures of predictability such as surprisal/forward transitional probability are no more, or even less effective in describing behavioral and neural phenomena than measures of post- or retrodictability such as backward transitional probability. We address this paradox by looking at the relationship between these different information theoretic measures and proposing a mechanistic account of how they are used in cognition. We posit that backward transitional probabilities support causal inferences about the occurrence of word sequences. Using Bayes' Theorem, we demonstrate that predictions (formalized as forward transitional probabilities) can be used in conjunction with the marginal probabilities of the current state/word and the upcoming state/word to compute these causal inferences. This conceptualization of causal inference in language processing both accounts for the role of prediction, and the surprising effectiveness of backwards transitional probability as a predictor of human behavior and its neural correlates.

*Jeroen.vanParidon@mpi.nl

On n-gram frequency and conditional probabilities

For at least half a century, it has been recognized high frequency[1] words are easier to produce (Oldfield & Wingfield, 1965; Jescheniak & Levelt, 1994) and to recognize, both in speech (Broadbent, 1967; Dahan et al., 2001; Cleland et al., 2006) and in print (Rayner, 1998; Cleland et al., 2006; Kuperman, 2013). However, when modeling language processing (be it speech perception, reading, etc.), we are often interested in processing beyond the single word level. Processing at the level of multi-word phrases (word n-grams) is more complex to model than single-word processing. This is partly due to the (linear) increase in lexical factors when modeling multi-word phrases, but more importantly, our understanding of phrase processing is not as well developed as our understanding of single-word processing. One easily accessible statistic relevant to phrase processing is word n-gram frequency, which has indeed been demonstrated to affect both language comprehension (Arnon & Snider, 2010) and language production (Janssen & Barber, 2012; Shao et al., 2019). These n-gram effects occur in addition to, and are distinct from, the effect of single-word frequency (Jacobs et al., 2016; Shao et al., 2019). It has been suggested, therefore, that these n-grams are stored as single units (*lexical bundles*, see e.g. Tremblay et al., 2011; Jacobs et al., 2016). However, given the combinatorial explosion of word n-grams that occurs for any value of n greater than 1, it is clear that storing n-grams (in some sort of expanded mental lexicon) is infeasible for all but the highest frequency n-grams (Baayen et al., 2013; Onnis & Huettig, in prep.), making whole n-gram storage inconsistent with the observation that n-gram frequency effects affect both high and low frequency n-grams (Arnon & Snider, 2010). We therefore reject the notion that apparent n-gram frequency effects are caused by the storage of whole n-grams and their frequencies, except for phrases with frequencies high enough to classify them as idioms (or compounds, cf. Jacobs & Dell, 2014), rather than phrases with purely compositional meaning. A more feasible mechanism than storing whole n-grams is to make use of conditional probabilities: The probability of a word occurring, given the

[1] We use the term frequency with regards to word occurrence in this article, which generally denotes a rate of occurrence (e.g. number of word occurrences per 1 million words, a scale from 0 to 1 million). Note however that for the purpose of comparing relative rates of occurrence, frequency is completely interchangeable with absolute counts (a scale from 0 to whatever the size of the corpus) and probabilities (a scale from 0 to 1).

occurrence of the preceding word. These conditional probabilities can be computed bidirectionally and are generally called transitional probabilities in the context of language (but note that these concepts are fundamentally equivalent). In studies of reading, the *forward transional probability* is generally referred to as *predictability*, which has been found to have a robust effects on various reading-related measures (e.g. first fixation duration, Balota et al., 1985; and inspection probability, Kliegl et al., 2004; for an alternative implementation of predictability see MacDonald & Shillcock, 2004). Transitional probabilities can also be reframed as *surprisal* ($-\log P_{\text{conditional}}$), an information theoretic measure that is often used in the field of Natural Language Processing. If we conceptualize the mental lexicon as a network of nodes and edges, transitional probabilities could feasibly be encoded in the edge weights, whereas storing whole n-grams requires an exponential increase in the number of nodes (cf. Baayen et al., 2013).

Hard to tell the difference: Equivalences

In the following sections, we transform all relevant quantities to a logarithmic scale (which is common practice) for reasons of computational convenience and cognitive plausibility[2].

Forward transitional probability (FTP) is a function of bigram and word$_1$ frequency:

$$\log P(w|w_{\text{prev}}) = \log \frac{P(w_{\text{prev}}, w)}{P(w_{\text{prev}})} = \log P(w_{\text{prev}}, w) - \log P(w_{\text{prev}})$$

(Eq. 1)

[2]Evidence suggests that word frequencies are experienced (both consciously and subconsciously) on a logarithmic scale. Contrast *angry* and *enraged*, for instance: *angry* is a fairly frequent word and *enraged* is fairly infrequent (in fact, in our dataset, *angry* is 56 times more frequent than *enraged*), however the effect of this difference in frequency on the difference in e.g. reading times or lexical decision times will not be proportional to the frequency, but to the logarithm of the difference in frequency. Similarly, when asked for explicit ratings in the difference in word frequency between different words, people are likely to give answers proportional to the logarithm of the frequency. This is in line with other power laws in cognition and perception and the reason why common measurement scales such as decibels for sound intensity are logarithmic in the physical unit, but linear in perception. Note also that the base of the logarithm is irrelevant, in general, because every logarithms is a multiple of every other logarithm, so when rescaling predictors to their standard deviation (common practice for linear regression in cognitive science), because the standard deviation of a log-transformed predictor is proportional to the base of the logarithm, the rescaled predictor will be invariant with respect to the base of the logarithm.

Backward transitional probability (BTP) is a function of bigram and word$_2$ frequency:

$$\log P(w_{\text{prev}}|w) = \log \frac{P(w_{\text{prev}}, w)}{P(w)} = \log P(w_{\text{prev}}, w) - \log P(w)$$

(Eq. 2)

So, using Bayes' Theorem[3] we can compute FTP from BTP (and vice versa):

$$\log P(w|w_{\text{prev}}) = \log \frac{P(w_{\text{prev}}|w) \cdot P(w)}{P(w_{\text{prev}})} = \log P(w_{\text{prev}}|w) + \log P(w) - \log P(w_{\text{prev}})$$

(Eq. 3)

Ergo, information-theoretic surprisal, which is equivalent to negative log FTP can be computed from word frequencies and BTP[4]:

$$-\log P(w|w_{\text{prev}}) = \log P(w_{\text{prev}}) - \log P(w_{\text{prev}}|w) - \log P(w)$$

(Eq. 4)

Similar results can be derived for other information theoretic measures.

The surprising effect of surprisal: Multicollinearity in linear models of behavior

The practical consequence of the equivalences outlined above is that when multiple measures of frequency and co-occurrence are used simultaneously as predictors in a linear model,

[3]Bayes' Theorem as used here is simply the law of conditional probabilities. Its use here is not specific to Bayesian statistics.

[4]Similarly, pointwise mutual information (PMI) can be computed from frequencies:

$$\log \frac{P(w_{\text{prev}}, w)}{P(w_{\text{prev}}) \cdot P(w)} = \log P(w_{\text{prev}}, w) - \log P(w) - \log P(w_{\text{prev}})$$

(Eq. 5) And considering Equations 1 and 2, that means we can compute PMI from transitional probability (symmetrically):

$$\log \frac{P(w_{\text{prev}}, w)}{P(w_{\text{prev}}) \cdot P(w)} = \log P(w|w_{\text{prev}}) - \log P(w) = \log P(w_{\text{prev}}|w) - \log P(w_{\text{prev}})$$

(Eq. 6) Many information-theoretical measures can be trivially computed from frequencies and transitional probabilities in this fashion.

this tends to result in collinearity between linear combinations of predictors. If this multicollinearity is perfect, it is impossible to perform the linear algebra necessary to fit the regression models. Most statistics packages will issue a warning regarding this multicollinearity and which predictors it concerns. However, even in cases where there is not perfect multicollinearity, the use of two or more co-occurrence measures can lead to unexpected consequences.

A hypothetical example: To predict reading times of a word of interest, w, we use w frequency and FTP from w_{prev} to w as predictors (the former as a measure for the ease of retrieving the current word, the latter as a measure for the predictability of the upcoming word). Counterintuitively, we find that low FTP is associated with *faster* reading. Does this mean that surprising words are somehow also more predictable? That seems contradictory. However, let's consider a simple linear model of reaction time with word frequency and FTP as predictors. For simplicity, we leave out the error term:

$$\log RT = \beta_0 + \beta_1 \cdot \log P(w) + \beta_2 \cdot \log P(w|w_{\text{prev}})$$

Now, using Equation 3, we note that:

$$\log P(w|w_{\text{prev}}) = \log P(w_{\text{prev}}|w) + \log P(w) - \log P(w_{\text{prev}})$$

And we substitute this back into the model:

$$\log RT = \beta_0 + \beta_1 \cdot \log P(w) + \beta_2 \cdot (\log P(w_{\text{prev}}|w) + \log P(w) - \log P(w_{\text{prev}}))$$

From this, it becomes clear that the apparent negative effect for FTP on actually reflects an effect of BTP minus w_{prev} frequency.

$$\log RT = \beta_0 + (\beta_1 + \beta_2) \cdot \log P(w) + \beta_2 \cdot (\log P(w_{\text{prev}}|w) - \log P(w_{\text{prev}}))$$

Because the effect of word frequencies is so large and stable, w_{prev} frequency will tend to dominate BTP, and therefore the coefficient β_2 will be negative. At the same time, because β_2 is negative, β_1 will be inflated making the effect of w frequency seem larger than it is.

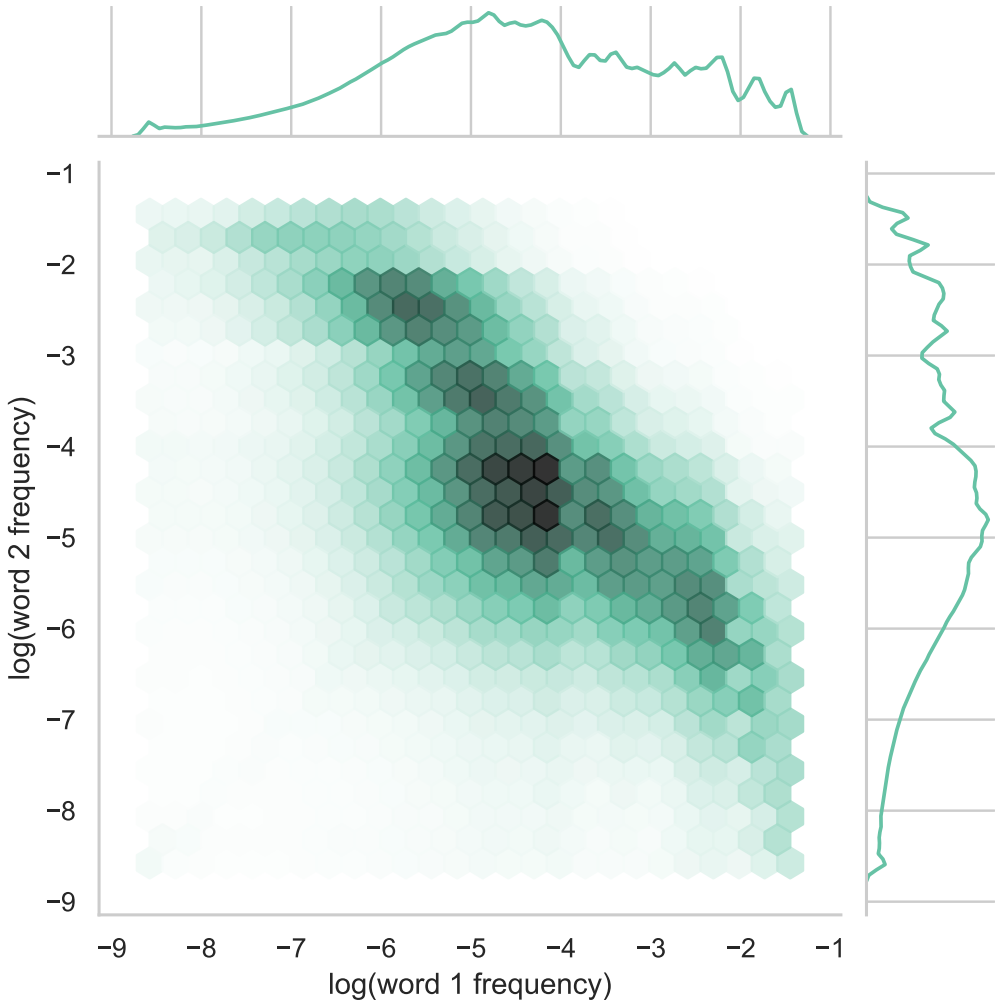

Figure 1: Joint distribution of first and last word frequencies for bigrams.

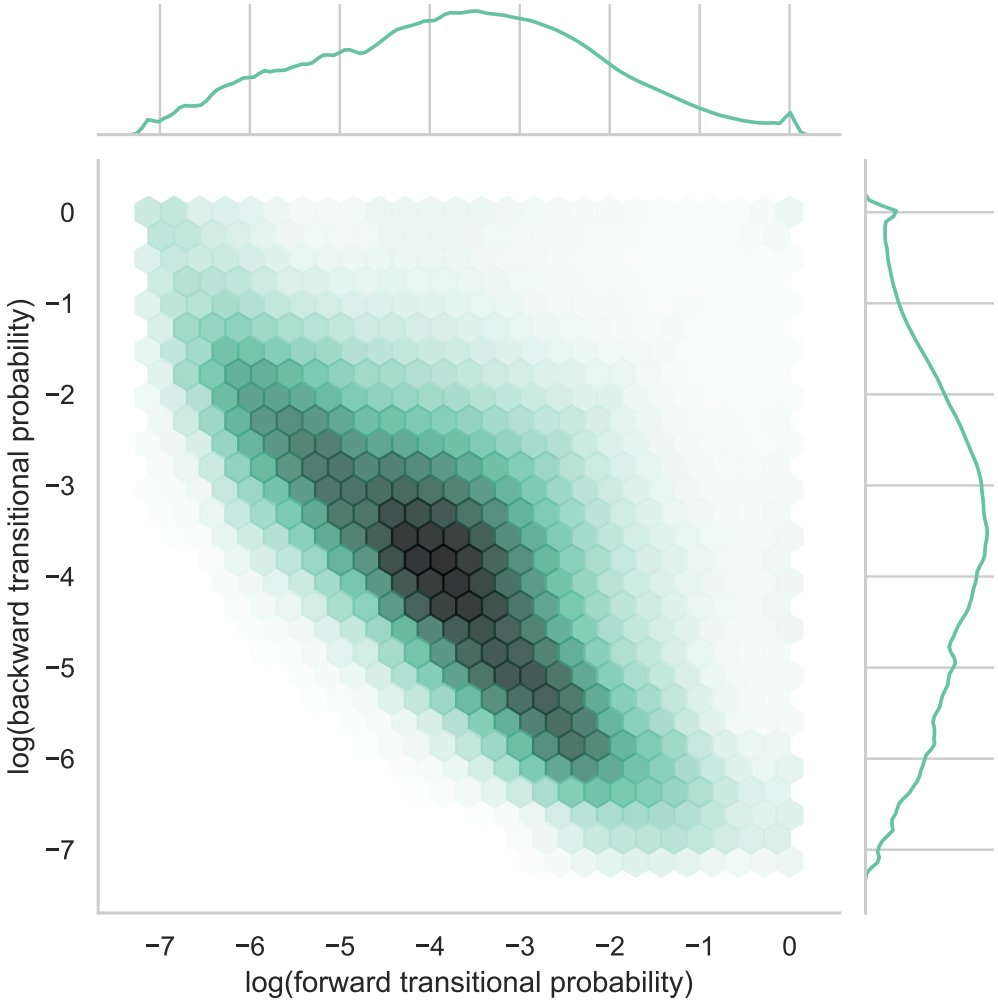

Figure 2: Joint distribution of forward and backward transitional probabilities for bigrams.

On making theoretically motivated choices

Mathematically exchangeable entities are convenient because they allow us to chose the formulation that is most convenient for us to work with. They are however also problematic because there is no reason to assume that the mind and brain use the same formulation that we do. Naive methods stand no chance of distinguishing between the different accounts, due to issues such as correlation and multicollinearity (see e.g. Levy, 2008, for a footnote on how FTP/BTP correlation complicates predictor selection). The psychological or neurobiological implementation matters, but we are not able to determine that from these measures alone. Instead, we have to look for fundamental structural reasons why one representation would be more compatible with other structures and mechanisms, such as neural connectivity, much the same way that arguments about frequency versus time domain representation in M/EEG are resolved by proposing fundamental mechanisms and not by computing the Fourier transform.

Rethinking predictive coding: Retrodiction as inference

Rather than putting prediction central in cognition, we posit that cognition functions by making probabilistic causal inferences. We start from the assumption that at its core, cognition subserves a perception-action loop. In this light, a particularly useful mechanism is to compute inferences about the state of the external world and the things that led to the current state (i.e. causality), as this is can guide both (imperfect) perception and subsequent action planning. Inferences about the current state of the world and the chain of states leading to the current state are encoded as backward transitional probabilities. The backward transitional probability directly answers the question "how probable is it, that the currently observed state was preceded by a given state?". This probabilistic notion of causality is the same type used in *Granger causility*: it does not imply causality in the philosophical or physical sense, but it does imply stochastic sequential dependence (Granger, 1969). This inference is computed via Bayes' Theorem, as above (Equation 3). In particular, we use information about marginal probabilities (of the current state (marginal likelihood)

and the next state (prior)) combined with the conditional probability of the next state based on the current state (likelihood, here FTP) to compute probabilistic causality. Note that prediction occurs here as an intermediate step in determining causality: the likelihood, i.e. FTP, is a critical piece in computing causality. Note that this account also explains the relative success of measures such as cloze probability. In this framework, cloze probability corresponds to the maximum likelihood. In a typical experiment, where word frequencies have been carefully controlled, we thus have a manipulation of the likelihood under nearly constant priors. Because the maximum likelihood under constant priors is proportional to the maximum a posteriori value (MAP), i.e. the peak of the posterior, the standard cloze manipulation corresponds to a manipulation of the BTP. At the same time, we do not have perfect control over the prior (word frequency) in experiments, and so the maximum likelihood does not directly correspond to the MAP.

Bayesian brains with *some* probabilities?

Although we have presented our account as a direct manipulation of the relevant probabilities, this is not a necessity. Indeed, our account is compatible with both sampling perspectives with or without direct knowledge of probabilities (cf. Sanborn & Chater, 2016) and with variational accounts (Friston, 2005; Friston, Thornton, & Clark, 2012). It is consistent with the "reversal" of the flow of prediction and error in prominent accounts such as Friston's (2005) theory of cortical responses. In this theory, prediction flows upward through the cortical hierarchy, while error propagates downward. In our account, prediction is used to compute the probability of a given cause, which corresponds to the goodness of fit, or equivalently error, associated with that cause.

Conclusion

The rise of information theory in the brain and behavioral sciences has presented researchers with a plethora of potential quantitative measures. We have shown that the arbitrary choice of measure cannot be driven by purely statistical concerns, because commonly used measures

are linear combinations of each other and thus statistically indistinguishable. This complex interrelation gives rise to apparent paradoxes, such as an illusory facilitation in processing surprising words when controlling for absolute frequency. However, these paradoxes should not be overinterpreted, as they are spurious, introduced by a particular decomposition. Instead, we should focus on mechanistic accounts, such as the one proposed here. By assuming that inferences about causality are instrumental in both perception and action, two of the core operations of cognition, we arrive at an account of prediction as a side effect, rather than a goal of cognition. This account allows us to make theoretically motivated choices between information theoretic measures as predictors for language processing and human behavior more generally.