



HAL
open science

Mathematical definition of the BRAID probabilistic model: Technical report companion to (Phénix et al., 2025)

Thierry Phénix, Emilie Ginestet, Sylviane Valdois, Julien Diard

► To cite this version:

Thierry Phénix, Emilie Ginestet, Sylviane Valdois, Julien Diard. Mathematical definition of the BRAID probabilistic model: Technical report companion to (Phénix et al., 2025). Université Grenoble - Alpes; CNRS. 2025. hal-04920250

HAL Id: hal-04920250

<https://hal.science/hal-04920250v1>

Submitted on 30 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mathematical definition of the BRAID probabilistic model: Technical report companion to (Phénix et al., 2025)

Thierry Phénix, Émilie Ginestet, Sylviane Valdois, and Julien Diard
Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France

1. Introduction

This is a technical note, to provide the complete mathematical definition of the BRAID model. Unfortunately, it could not be included in the main paper (Phénix et al., 2025). Therefore, it technically has not been peer-reviewed, in this form, during the publication process, in 2025 in the journal *Psychonomic Bulletin & Review*. However, it can be seen as an English translation and summary of the mathematical definition of the BRAID model as found in Thierry Phénix’s PhD thesis manuscript, which, of course, has been peer-reviewed as part of Thierry’s PhD defense.

The first section serves as a preamble: it recalls the two schematic representations of the BRAID model, reprised from Figure 1 and Figure 2 of the published paper (Phénix et al., 2025).

The second section contains the full mathematical definition of the model, with the definition of all variables, the decomposition of the joint probability distribution (this is a slight adaptation of Appendix A of the published paper (Phénix et al., 2025)), and the definition of all terms appearing in this decomposition. The section then provides details on how Bayesian inference yields probabilistic computation for all tasks of interest: letter recognition, word recognition, letter recognition with lexical influence, and lexical decision.


The third and final section provides the mathematical definition and properties of controlled coherence variables.

2. Schematic representations of the BRAID model

Figure 1 shows, side by side, two schematic representations of the BRAID model.

3. BRAID model definition

We apply the Bayesian Programming methodology (Bessière et al., 2013; Lebeltel et al., 2004), in which a “Bayesian program” is a structure that consists of two parts. In the first part, called a “description”, we provide a complete, operational mathematical definition

Correspondence concerning this paper should be addressed to Julien Diard  <https://orcid.org/https://orcid.org/0000-0003-0673-477X>, Laboratoire de Psychologie et NeuroCognition, CNRS UMR 5105; Université Grenoble Alpes, BMD; 1251 Av. des Universités, CS40700, 38058 Grenoble Cedex 9, France. Email: Julien.Diard@univ-grenoble-alpes.fr

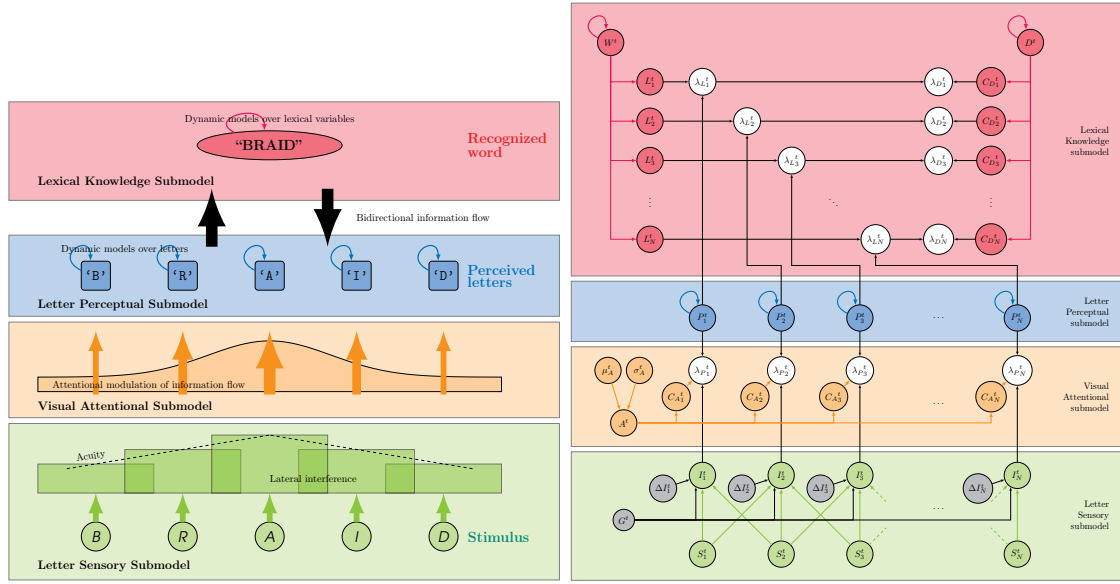


Figure 1

Graphical representations of the BRAID model. The left schema is a conceptual representation of the model, the right schema is the graphical representation of the structure of the model, according to the usual convention of probabilistic graphical models (except for the self-looping arrows, which represent temporal dependencies).

of the joint probability distribution of interest. In the second part, the joint probability distribution is used, thanks to Bayesian inference, to compute answers to probabilistic questions of interest.

3.1 BRAID model: description

To define the joint probability distribution that constitutes the probabilistic knowledge base of the BRAID model, first we define the variables it involves, second we decompose the joint probability distribution as a product of terms, introducing conditional independence hypotheses to simplify some terms, and third and finally, we mathematically define each term of the decomposition.

Variables

To define each variable of the model, we give it a name and variation domain, i.e., a set of possible values. Subscript indexes $X_{1:N}$ refer to spatial position in a left-to-right letter sequence, superscript indexes $X^{1:T}$ refer to time evolution of variable X from time index 1 to T .

- $S_{1:N}^{1:T}$, $I_{1:N}^{1:T}$, $P_{1:N}^{0:T}$ and $L_{1:N}^{1:T}$ are variables over letter identity, that is, they each have the domain $\mathcal{D}_L = \{ 'a', 'b', \dots, 'z', '$' \}$, with '\$' a symbol representing the unknown or missing letter. We note $|\mathcal{D}_L|$ the cardinal of the space \mathcal{D}_L , i.e. $|\mathcal{D}_L| = 27$.

- $W^{0:T}$ are variables over word identity, that is, they each have the domain $\mathcal{D}_W = \{w_1, w_2, \dots, w_K\}$.
- $D^{0:T}$ are Boolean variables for lexical decision, that is, their values are noted $\{T, F\}$ (for *True*, *False*, respectively).
- $\Delta I_{1:N}^{1:T}$ are variables over relative position shifts, that is, they each have the domain $\mathcal{D}_{\Delta I} = \{-1, 0, 1\}$.
- $\lambda_{L_{1:N}}^{1:T}$, $\lambda_{P_{1:N}}^{1:T}$, $\lambda_{D_{1:N}}^{1:T}$ are coherence variables, that is, binary variables that can be used a “Bayesian switches”. Their values are noted $\{0, 1\}$.
- $C_{D_{1:N}}^{1:T}$ are binary variables, their values are noted $\{0, 1\}$.
- $C_{A_{1:N}}^{1:T}$ are “control variables”, that is, binary variables that can pilot the state of coherence variables. Their values are noted $\{0, 1\}$.
- $A^{1:T}$ are variables denoting the spatial positions of letters. Their domains are the discrete set of possible letter positions, i.e. for N -letter words, $\mathcal{D}_A = \{1, 2, \dots, N\}$.
- $\mu_A^{1:T}$ and $G^{1:T}$ are variables over spatial positions. Their domains are the continuous sets of possible attention and gaze position, i.e. $\mathcal{D}_{Pos} = [1, N]$.
- $\sigma_A^{1:T}$ are variables over spatial dispersion. Their domains are the interval $(0, 100]$.

Decomposition

The core definition of the BRAID model is the joint probability distribution

$$JD_{BRAID} = P \left(\begin{array}{cccccccc} W^{0:T} & L_{1:N}^{1:T} & \lambda_{L_{1:N}}^{1:T} & P_{1:N}^{0:T} & A^{1:T} & \mu_A^{1:T} & \sigma_A^{1:T} & C_{A_{1:N}}^{1:T} & \lambda_{P_{1:N}}^{1:T} \\ G^{1:T} & S_{1:N}^{1:T} & \Delta I_{1:N}^{1:T} & I_{1:N}^{1:T} & \lambda_{D_{1:N}}^{1:T} & D^{0:T} & C_{D_{1:N}}^{1:T} & & \end{array} \right),$$

which is defined by the following decomposition:

$$\begin{aligned}
 JD_{BRAID} = & \tag{1} \\
 & P(W^0)P(D^0) \prod_{n=1}^N P(P_n^0) \\
 & \prod_{t=1}^T \left[\begin{array}{l}
 P(W^t | W^{t-1}) \prod_{n=1}^N P(L_n^t | W^t) \\
 P(D^t | D^{t-1}) P(C_{D_{1:N}}^t | D^t) \\
 \prod_{n=1}^N \left[P(\lambda_{D_n}^t | \lambda_{L_n}^t C_{D_n}^t) P(\lambda_{L_n}^t | L_n^t P_n^t) \right] \\
 \prod_{n=1}^N P(P_n^t | P_n^{t-1}) \\
 P(A^t | \mu_A^t \sigma_A^t) P(\mu_A^t) P(\sigma_A^t) \prod_{n=1}^N P(C_{A_n}^t | A^t) \\
 \prod_{n=1}^N P(\lambda_{P_n}^t | P_n^t I_n^t C_{A_n}^t) \\
 P(G^t) \prod_{n=1}^N \left[P(S_n^t) P(\Delta I_n^t) P(I_n^t | S_{1:N}^t \Delta I_n^t G^t) \right]
 \end{array} \right]
 \end{aligned}$$

The top terms (first line of Equation (1), outside of the main product) concern the initial state of the model (at time $t = 0$), whereas the innermost product contains the temporally local portion of the model, i.e., the model that is iterated at each time step $t \neq 0$. The innermost product is laid out over seven lines, roughly following a top-down traversal of the dependency structure shown Figure 1 (right), from lexical knowledge to stimulus.

Parametric forms and parameter identification

We now define each term of the model, that is to say, each term appearing in Equation (1), by defining their parametric forms.

- $P(W^0)$ is the prior probability distribution over word identity. It is a discrete probability distribution $P([W^0 = w_i]) = p_{w_i}$, whose parameters p_{w_i} are identified from a frequency database provided by a chosen lexicon.
- $P(D^0)$ is the prior probability distribution over the lexical decision binary variable, that is to say, over whether a word is present or not. It is a Bernoulli distribution $P([D^0 = T]) = p_{d^0}$, whose parameters $p_{d^0} = 1/2$ in our experiments, assuming ignorance of the frequency of encountered words and non-words, or, equivalently, assuming half of stimuli are words.
- $\forall n, P(P_n^0)$ are the prior probability distributions over letter identity. They are discrete uniform probability distributions.
- $\forall t, P(W^t | W^{t-1})$ are the probability distributions for the (stationary) dynamical model of word identity evolution in lexical memory. These discrete conditional probability distributions are generalized version of the Laplace succession law. In other

words, they are “almost Dirac” discrete conditional probability distributions: most of the probability mass is set on the same word as in the previous time step, and the rest is distributed as is distributed the prior probability distribution over words, $P(W^0)$. When iterated with itself, $P(W^t | W^{t-1})$ therefore converges towards $P(W^0)$, which can be seen as its resting state. Recalling that $P([W^0 = w_i]) = p_{w_i}$, and introducing parameter $Leak_W$ to control convergence speed, we define:

$$\begin{aligned} & P([W^t = w^t] | [W^{t-1} = w^{t-1}]) \\ &= \begin{cases} \frac{1+p_{w^t} Leak_W}{1+Leak_W} & \text{if } w^t = w^{t-1} \\ \frac{p_{w_i} Leak_W}{1+Leak_W} & \text{otherwise.} \end{cases} \end{aligned}$$

- $\forall t, n, P(L_n^t | W^t)$ are the probability distributions for the (stationary) model of how letters compose words, i.e., a lexical database of known words. They are “almost Dirac” discrete conditional probability distributions: given a lexicon, a word and a position, almost all the probability mass ($1 - \epsilon$) is set on the correct letter, and the rest of the probability mass is uniformly distributed over all other letters. Recall that the letter space is of size $|\mathcal{D}_L|$, so that, mathematically:

$$\begin{aligned} & P([L_n^t = l] | [W^t = w^t]) \\ &= \begin{cases} 1 - \epsilon & \text{if } l \text{ is the correct letter} \\ & \text{for word } w^t \text{ at position } n \\ \epsilon / (|\mathcal{D}_L| - 1) & \text{otherwise.} \end{cases} \end{aligned}$$

- $\forall t, n, P(\lambda_{L_n^t} | L_n^t P_n^t)$ are the probability distributions for the coherence models between the lexical knowledge model and the perceptual letter representation model. As any coherence model, they are Dirac probability distributions over value $\lambda_{L_n^t} = 1$ when $L_n^t = P_n^t$. Mathematically:

$$P([\lambda_{L_n^t} = 1] | [L_n^t = l_L] [P_n^t = l_P]) = \begin{cases} 1 & \text{if } l_L = l_P \\ 0 & \text{otherwise.} \end{cases}$$

- $\forall t, n, P(P_n^t | P_n^{t-1})$ are the probability distributions for the (stationary) dynamical model of letter identity evolution in the perceptual letter representation model. These discrete conditional probability distributions are “almost Dirac” discrete conditional probability distributions: almost all the probability mass is set on the same letter as in the previous time step, and the rest of the probability mass is uniformly distributed over all other letters (so that, when iterated with itself, P_n^t would decay to a uniform probability distribution, with decay speed controlled by parameter $Leak_P$). Mathematically, $P(P_n^t | P_n^{t-1})$ are Laplace succession laws:

$$\begin{aligned} & P([P_n^t = l_P^t] | [P_n^{t-1} = l_P^{t-1}]) \\ &= \begin{cases} \frac{1+Leak_P}{1+|\mathcal{D}_L| Leak_P} & \text{if } l_P^t = l_P^{t-1} \\ \frac{Leak_P}{1+|\mathcal{D}_L| Leak_P} & \text{otherwise.} \end{cases} \end{aligned}$$

- $\forall t, P(A^t \mid \mu_A^t, \sigma_A^t)$ are the probability distributions for the model of attention distribution. They are discrete and bounded approximations of Gaussian probability distributions over the space of possible letter positions. Their parameters are their means μ_A^t and standard-deviations σ_A^t .
- $\forall t, P(\mu_A^t)$ and $P(\sigma_A^t)$ are prior probability distributions on the parameters of attention distribution. They both are uniform probability distributions over their respective domains. This choice is of no practical consequence, as all inferences considered in this manuscript are conditioned on given values for variables $\mu_A^t = \mu, \sigma_A^t = \sigma$, so that $P([\mu_A^t = \mu])$ and $P([\sigma_A^t = \sigma])$ can always be aggregated into normalization constants. The values μ and σ for μ_A^t and σ_A^t depend on the experiment, and are provided in the main text.
- $\forall t, n, P(\lambda_{P_n^t} \mid P_n^t, I_n^t, C_{A_n^t})$ are the probability distributions for the coherence models between the perceptual letter representation model and the visual letter recognition model, controlled by attention. When the control variable $C_{A_n^t} = 1$, the probability distribution over coherence variable $\lambda_{P_n^t}$ is a Dirac distribution centered on value 1 when $P_n^t = I_n^t$, and centered on value 0 otherwise. When the control variable $C_{A_n^t} = 0$, the probability value for the case $[\lambda_{P_n^t} = 1]$ is $1/|\mathcal{D}_L|$. Mathematically:

$$\begin{aligned}
 & P([\lambda_{P_n^t} = 1] \mid [P_n^t = l_P] [I_n^t = l_I] [C_{A_n^t} = c]) \\
 &= \begin{cases} 1 & \text{if } c = 1 \text{ and } l_P = l_I \\ 0 & \text{if } c = 1 \text{ and } l_P \neq l_I \\ 1/|\mathcal{D}_L| & \text{if } c = 0. \end{cases}
 \end{aligned}$$

Details about control variables and demonstrations of their properties, in the general case, as well as the rationale and consequence of value $1/|\mathcal{D}_L|$ for the case $c = 0$, are provided in the last section of this document.

- $\forall t, n, P(C_{A_n^t} \mid A^t)$ are the probability distributions for the models of attention allocation to spatial positions. They are Bernoulli probability distributions, with the attention probability allocated to position n by $P(A^t)$ being the probability for value $C_{A_n^t} = 1$:

$$\begin{cases} P([C_{A_n^t} = 1] \mid A^t) = P([A^t = n]) \\ P([C_{A_n^t} = 0] \mid A^t) = 1 - P([A^t = n]) \end{cases} .$$
- $\forall t, P(G^t)$ are the probability distributions for the model of gaze positioning. They are discrete uniform probability distributions over the space of possible letter positions. This choice is of no practical consequence, as all inferences considered in this manuscript are conditioned on given values of gaze position g^t .
- $\forall t, n, P(S_n^t)$ are the prior probability distributions over stimuli. They are discrete uniform probability distributions. Here again, this choice is of no practical consequence, as all inferences considered in this manuscript are conditioned on given values s_n^t for variables S_n^t , so that $P([S_n^t = s_n^t])$ can always be aggregated into normalization constants. The values s_n^t describe the stimuli used in experiment simulations, i.e. the input words, pseudo-words, etc.

- $\forall t, n, P(\Delta I_n^t)$ are the probability distributions over relative input shifts, for the model of lateral interference between letters. They are discrete probability distributions over position indexes $\mathcal{D}_{\Delta I} = \{-1, 0, 1\}$, that we choose to constrain to be symmetric for internal letters, with parameter θ_I regulating the strength of interference from neighboring letters. Mathematically:

$$P([\Delta I_n^t = i]) = \begin{cases} 1 - \theta_I & \text{if } i = 0 \\ \theta_I/2 & \text{otherwise.} \end{cases}$$

For the first and last letter of a word, however, all interference is provided by the only neighboring letter, and the distribution is then normalized:

$$P([\Delta I_n^t = i]) = \begin{cases} \frac{2\theta_I}{1+\theta_I} & \text{if } i = 0 \\ \frac{1-\theta_I}{1+\theta_I} & \text{otherwise.} \end{cases}$$

- $\forall t, n, P(I_n^t | S_{1:N}^t \Delta I_n^t G^t)$ are the probability distributions for the model of stimulus decoding. They are discrete probability distributions over letter identity, as a function of stimulus input, relative position shift, and gaze position. Describing this term is easier done in pedagogical steps.

First, consider a probability distribution of the form $P(I_n^t | S_n^t)$: this would be a conditional table of discrete probability distributions, whose parameters can be identified from any number of experimental confusion matrices in the literature (i.e., from one specific confusion matrix to capture the specifics of experimental conditions it was measured in, or from an average of several to simulate overall statistics of confusions in letter identification). Let $\{p_{i,s}\}_{i,s \in \mathcal{D}_L^2}$ be such experimental parameters; we scale them down by a constant factor $Scale_I$, to affect the quantity of perceptual information that enters the model at each time step (in effect, this slows down simulation time by scaling the arbitrary time unit).

Second, eccentricity from gaze position is used to augment the scaling factor $Scale_I$, as a linear function of distance between gaze g^t and letter position n . The slope of this linear function is related to a parameter noted θ_G . Gaze position G^t thus also conditions the term $P(I_n^t | S_{1:N}^t G^t)$.

Third and finally, we implement interference from neighboring stimuli by refining $P(I_n^t | S_n^t G^t)$ into $P(I_n^t | S_{1:N}^t \Delta I_n^t G^t)$: ΔI_n^t represents a possible lateral shift between stimulus and letter recognition. When $\Delta I_n^t = 0$, the stimulus at position n correctly feeds the letter representation at the same position n . However, when $\Delta I_n^t = -1$, the stimulus at position $n - 1$ incorrectly feeds the letter representation at position n , etc. Given the value of ΔI_n^t , the adequate set of $\{p_{i,s}\}$ parameters is used.

- $\forall t, n, P(\lambda_{D_n}^t | \lambda_{L_n}^t C_{D_n}^t)$ are the probability distributions for the coherence models between coherence variables of the word recognition model and the decision control variables. As any coherence model, they are Dirac probability distributions over value $\lambda_{D_n}^t = 1$ when $\lambda_{L_n}^t = C_{D_n}^t$. Mathematically:

$$P([\lambda_{D_n}^t = 1] | [\lambda_{L_n}^t = l] [C_{D_n}^t = c]) = \begin{cases} 1 & \text{if } l = c \\ 0 & \text{otherwise.} \end{cases}$$

- $\forall t, P(D^t | D^{t-1})$ are the probability distributions for the (stationary) dynamical model of lexical decision evolution. They are conditional Bernoulli distributions featuring decay parameter $Leak_D$, that controls convergence speed for each hypothesis $D^t = T$ and $D^t = F$:

$$\begin{cases} P([D^t = T] | [D^{t-1} = T]) &= 1 - Leak_D \\ P([D^t = F] | [D^{t-1} = F]) &= 1 - Leak_D . \end{cases}$$

- $\forall t, P(C_{D_{1:N}^t} | D^t)$ are joint probability distributions of control variables that define the lexical decision knowledge. There are two cases to consider. Firstly, when reading a word ($D^t = T$), the perceived letters should perfectly match those of a word in orthographic knowledge, so that all $C_{D_n^t}$ should be 1. Secondly, when reading something that is not a word ($D^t = F$), some or all of control variables $C_{D_n^t}$ can be 0. However, to limit computation time, we limit the model so that it only consider non-words which have at most 1 letter that differs from a real word (in practice, this has a negligible impact, as having two differing letters will also be recognized as having probably at least 1 letter differing). The position of the error is unconstrained. Mathematically:

$$\begin{aligned} &P([C_{D_{1:N}^t} = \{c_1, \dots, c_N\}] | [D^t = d]) \\ &= \begin{cases} 1 & \text{if } d = T \text{ and } \forall n, c_n = 1 \\ 1/N & \text{if } d = F \text{ and } \exists! n, c_n = 0 \\ 0 & \text{otherwise .} \end{cases} \end{aligned}$$

(We recall that $\exists! n, c_n = 0$ reads “there exists a single n such that $c_n = 0$, i.e., $P(C_{D_{1:N}^t} | [D^t = F])$ is uniformly distributed over the N cases that contain a single error $c_n = 0$ at position n .)

This concludes the description of probabilistic terms featured in the BRAID model.

4. Using Bayesian inference to model cognitive tasks

The BRAID model is used to solve three main cognitive tasks, that we used in several simulated experiments: isolated letter recognition, word recognition and lexical decision. In the case of isolated letter recognition, we develop two variants, depending on whether lexical knowledge is allowed to influence letter recognition or not. Each cognitive task is modeled by a probabilistic question to BRAID, solved automatically by Bayesian inference.

Indeed, we call a *probabilistic question* any term of the form $P(\textit{Searched} | \textit{Known})$, where *Searched*, *Known* are subsets of variables appearing in the joint probability distribution of the model (such that $\textit{Searched} \neq \emptyset$, and such that *Searched*, *Known*, along with *Free*, form a partition of all variables in the joint probability distribution). Whatever the question, Bayesian inference allows to automatically compute its answer; this has been demonstrated in the general case elsewhere (Bessi ere et al., 2013; Lebeltel et al., 2004).

In this annex, we provide the mathematical definition of the probabilistic questions that model our cognitive tasks, and show what solution results from Bayesian inference from the joint probability distribution JD_{BRAID} of the BRAID model.

4.1 Isolated letter identification without lexical influence: $Q_{P_n^T}$

Letter recognition, in its simplest form, is modeled by the probabilistic question $Q_{P_n^T} = P(P_n^T \mid S_{1:N}^{1:T} [\lambda_{P_n}^{1:T} = 1] \mu_A^{1:T} \sigma_A^{1:T} G^{1:T})$: given the stimulus $S_{1:N}^{1:T}$, given attention distribution parameters $\mu_A^{1:T}, \sigma_A^{1:T}$ and gaze position $G^{1:T}$, and given that information is allowed to propagate in the model from the stimulus to the perceived letter at position n (i.e., $[\lambda_{P_n}^{1:T} = 1]$), what is the probability distribution over the perceived letter at position n , P_n^T ? Using the shorthand $k^t = s_{1:N}^t g^t \mu_A^t \sigma_A^t$, it is rewritten as $Q_{P_n^T} = P(P_n^T \mid k^{1:T} [\lambda_{P_n}^{1:T} = 1])$.

We now derive the answer to question $Q_{P_n^T}$ by applying Bayesian inference. We first notice that many coherence variables are left unspecified in the question (λ_L and λ_D at all positions, $\lambda_{P_m^t}$ for positions $m \neq n$). A key feature of coherence variables is that, in this case, they can be interpreted as open ‘‘Bayesian switches’’, so that we can simplify portions of the model beyond these coherence variables (Gilet et al., 2011). Therefore, even if we start from the entire joint probability distribution JD_{BRAID} , only variables between stimulus $S_{1:N}^{1:T}$ and letter percept P_n^T are involved. In other words, we can solve $Q_{P_n^T}$ in a ‘‘lighter’’ version of the model, which would be the joint probability distribution $JD_{Percept}$:

$$JD_{Percept_n^T} = P \left(\begin{array}{cccccc} P_n^{0:T} & A^{1:T} & \mu_A^{1:T} & \sigma_A^{1:T} & \lambda_{P_n}^{1:T} & \\ C_{A_n}^{1:T} & G^{1:T} & S_{1:N}^{1:T} & \Delta I_n^{1:T} & I_n^{1:T} & \end{array} \right),$$

defined by the following decomposition:

$$JD_{Percept_n^T} = P(P_n^0) \prod_{t=1}^T \left[\begin{array}{l} P(P_n^t \mid P_n^{t-1}) \\ P(A^t \mid \mu_A^t \sigma_A^t) P(\mu_A^t) P(\sigma_A^t) \\ P(\lambda_{P_n^t} \mid P_n^t I_n^t C_{A_n^t}) P(C_{A_n^t} \mid A^t) \\ P(G^t) \prod_{n=1}^N [P(S_n^t)] P(\Delta I_n^t) P(I_n^t \mid S_{1:N}^t \Delta I_n^t G^t) \end{array} \right].$$

To answer the question $Q_{P_n^T}$, we first involve this partial joint distribution $JD_{Percept}$ by marginalizing over missing variables. Then we reorder terms to make appear a temporal

recursion term.

$$\begin{aligned}
 Q_{P_n^T} &= P(P_n^T \mid k^{1:T} [\lambda_{P_n^{1:T}} = 1]) \\
 &\propto \sum_{\substack{I_n^{1:T} \\ \Delta I_n^{1:T}}} \sum_{P_n^{1:T-1}} \sum_{\substack{A^{1:T} \\ C_{A_n^{1:T}}}} JD_{Percept_n^T} \\
 &\propto \sum_{P_n^{T-1}} \left[\begin{array}{l} P(P_n^0) \\ \sum_{\substack{I_n^{1:T-1} \\ \Delta I_n^{1:T-1}}} \sum_{P_n^{1:T-2}} \sum_{\substack{A^{1:T-1} \\ C_{A_n^{1:T-1}}} } \prod_{t=1}^{T-1} \left[\begin{array}{l} P(P_n^t \mid P_n^{t-1}) \\ P(A^t \mid \mu_A^t \sigma_A^t) \\ P(\lambda_{P_n^t} \mid P_n^t I_n^t C_{A_n^t}) P(C_{A_n^t} \mid A^t) \\ P(G^t) \prod_{n=1}^N [P(S_n^t)] P(\Delta I_n^t) P(I_n^t \mid S_{1:N}^t \Delta I_n^t G^t) \end{array} \right] \\ P(P_n^T \mid P_n^{T-1}) \\ \sum_{\substack{I_n^T \\ \Delta I_n^T}} \sum_{\substack{A^T \\ C_{A_n^T}}} \left[\begin{array}{l} P(A^T \mid \mu_A^T \sigma_A^T) \\ P(\lambda_{P_n^T} \mid P_n^T I_n^T C_{A_n^T}) P(C_{A_n^T} \mid A^T) \\ P(G^T) \prod_{n=1}^N [P(S_n^T)] P(\Delta I_n^T) P(I_n^T \mid S_{1:N}^T \Delta I_n^T G^T) \end{array} \right] \end{array} \right]
 \end{aligned}$$

In the second line of the outermost summation, we first recognize the model at the previous time step, and thus a recursion term:

$$\begin{aligned}
 Q_{P_n^T} & \\
 &\propto \sum_{P_n^{T-1}} \left[\begin{array}{l} \sum_{\substack{I_n^{1:T-1} \\ \Delta I_n^{1:T-1}}} \sum_{P_n^{1:T-2}} \sum_{\substack{A^{1:T-1} \\ C_{A_n^{1:T-1}}} } JD_{Percept_n^{T-1}} \\ P(P_n^T \mid P_n^{T-1}) \\ \sum_{\substack{I_n^T \\ \Delta I_n^T}} \sum_{\substack{A^T \\ C_{A_n^T}}} \left[\begin{array}{l} P(A^T \mid \mu_A^T \sigma_A^T) \\ P(\lambda_{P_n^T} \mid P_n^T I_n^T C_{A_n^T}) P(C_{A_n^T} \mid A^T) \\ P(G^T) \prod_{n=1}^N [P(S_n^T)] P(\Delta I_n^T) P(I_n^T \mid S_{1:N}^T \Delta I_n^T G^T) \end{array} \right] \end{array} \right] \\
 &\propto \sum_{P_n^{T-1}} \left[\begin{array}{l} Q_{P_n^{T-1}} \\ P(P_n^T \mid P_n^{T-1}) \\ \sum_{\substack{I_n^T \\ \Delta I_n^T}} \sum_{\substack{A^T \\ C_{A_n^T}}} \left[\begin{array}{l} P(A^T \mid \mu_A^T \sigma_A^T) \\ P(\lambda_{P_n^T} \mid P_n^T I_n^T C_{A_n^T}) P(C_{A_n^T} \mid A^T) \\ P(G^T) \prod_{n=1}^N [P(S_n^T)] P(\Delta I_n^T) P(I_n^T \mid S_{1:N}^T \Delta I_n^T G^T) \end{array} \right] \end{array} \right]
 \end{aligned}$$

As in a standard Markov model, the temporal transition based on the recursion term can thus be isolated from the rest of the equation. This latter part is reorganized to make

appear the summation over the attentional switch variables $C_{A_n}^T$:

$$\begin{aligned} & \left[\sum_{P_n^{T-1}} \left[P(P_n^T | P_n^{T-1}) Q_{P_n^{T-1}} \right] \right. \\ & \propto \left[\sum_{\substack{I_n^T \\ \Delta I_n^T \\ C_{A_n}^T}} \sum_{A^T} \left[\begin{array}{l} P(A^T | \mu_A^T \sigma_A^T) \\ P(\lambda_{P_n^T} | P_n^T I_n^T C_{A_n}^T) P(C_{A_n}^T | A^T) \\ P(G^T) \prod_{n=1}^N [P(S_n^T)] P(\Delta I_n^T) P(I_n^T | S_{1:N}^T \Delta I_n^T G^T) \end{array} \right] \right] \\ & \propto \left[\sum_{P_n^{T-1}} \left[P(P_n^T | P_n^{T-1}) Q_{P_n^{T-1}} \right] \right. \\ & \left. \sum_{\substack{A^T \\ C_{A_n}^T}} \left[\sum_{I_n^T} \left[\sum_{\Delta I_n^T} \left[\begin{array}{l} P(\Delta I_n^T) \\ P(I_n^T | S_{1:N}^T \Delta I_n^T G^T) \end{array} \right] \right] \right] \right] \left. \right] \end{aligned}$$

It remains only to write explicitly the two terms of the sum over the binary variable $C_{A_n}^T$ and to simplify writing. We first remark that, since only position n is considered, we can collapse the summation over A^t . We also note $\alpha_n = P([C_{A_n}^T = 1] | [A^T = n])$, and p the current value considered for variable P_n^T , which propagates to variable I_n^T , in the collapse of the summation over I_n^T due to the Bayesian switch $\lambda_{P_n^T}$ being closed. We obtain:

$$\propto \left[\sum_{P_n^{T-1}} \left[P([P_n^T = p] | P_n^{T-1}) Q_{P_n^{T-1}} \right] \right. \\ \left. \left[\begin{array}{l} \alpha_n \sum_{\Delta I_n^T} \left[\begin{array}{l} P(\Delta I_n^T) \\ P([I_n^T = p] | S_{1:N}^T \Delta I_n^T G^T) \end{array} \right] + \\ \frac{(1 - \alpha_n)}{|\mathcal{D}_L|} \sum_{\Delta I_n^T} \left[\sum_{I_n^T} P(I_n^T | S_{1:N}^T \Delta I_n^T G^T) \end{array} \right] \right] \right] \left. \right]$$

Finally, applying the normalization rule to simplify sums, we obtain:

$$Q_{P_n^T} \propto \left[\sum_{P_n^{T-1}} \left[P([P_n^T = p] | P_n^{T-1}) Q_{P_n^{T-1}} \right] \right. \\ \left. \left[\begin{array}{l} \alpha_n \sum_{\Delta I_n^T} \left[\begin{array}{l} P(\Delta I_n^T) \\ P([I_n^T = p] | S_{1:N}^T \Delta I_n^T G^T) \end{array} \right] + \\ \frac{(1 - \alpha_n)}{|\mathcal{D}_L|} \end{array} \right] \right] \quad (2)$$

This result is easily interpreted, as it contains the classic components of inference in a Hidden Markov Model (HMM; Rabiner & Juang, 1993). Indeed, the temporal recursive question $Q_{P_n^{T-1}}$ is first multiplied by state transition probabilities $P(P_n^T | P_n^{T-1})$. This multiplication is then summed over P_n^{T-1} , which ‘‘predicts’’ letter percepts for next time step (in our case, this step involves memory decay, so that information about letter percepts

decays and uncertainty increases). Finally, information from the stimulus is acquired via the observation model. This observation model, in the BRAID model, is itself hierarchical and structured, and already involves mechanisms such as lateral interference (with summations over ΔI_n^T and I_n^T) and attention distribution (with summation over $C_{A_n}^T$).

4.2 Word recognition: Q_W^T

The second cognitive task we consider is word recognition, mathematically modeled by the probabilistic question $Q_W^T = P(W^T \mid S_{1:N}^{1:T} [\lambda_{L_{1:N}}^{1:T} = 1] [\lambda_{P_{1:N}}^{1:T} = 1] \mu_A^{1:T} \sigma_A^{1:T} G^{1:T})$: given the stimulus $S_{1:N}^{1:T}$, given attention distribution parameters $\mu_A^{1:T}, \sigma_A^{1:T}$ and gaze position $G^{1:T}$, and given that information propagates in the model from stimulus to words (i.e., $[\lambda_{L_{1:N}}^{1:T} = 1], [\lambda_{P_{1:N}}^{1:T} = 1]$), what is the probability distribution over words W^T ?

As previously, we use the property of open Bayesian switches, and observe that λ_D variables are unspecified: we can simplify out the part of the BRAID model dedicated to lexical decision. The “reduced” version of the joint distribution we use to solve the probabilistic question of word recognition is:

$$JD_{Word}^T = P \left(\begin{array}{cccccccc} W^{0:T} & L_{1:N}^{1:T} & \lambda_{L_{1:N}}^{1:T} & P_{1:N}^{0:T} & A^{1:T} & \mu_A^{1:T} & \sigma_A^{1:T} & \lambda_{P_{1:N}}^{1:T} & C_{A_{1:N}}^{1:T} \\ G^{1:T} & S_{1:N}^{1:T} & \Delta I_{1:N}^{1:T} & I_{1:N}^{1:T} & & & & & \end{array} \right),$$

which is defined by the following decomposition:

$$JD_{Word}^T = P(W^0) \prod_{n=1}^N P(P_n^0) \prod_{t=1}^T \left[\begin{array}{l} P(W^t \mid W^{t-1}) \prod_{n=1}^N P(L_n^t \mid W^t) \\ P(\lambda_{L_n}^t \mid L_n^t P_n^t) \\ \prod_{n=1}^N P(P_n^t \mid P_n^{t-1}) \\ P(A^t \mid \mu_A^t \sigma_A^t) P(\mu_A^t) P(\sigma_A^t) \\ \prod_{n=1}^N [P(\lambda_{P_n}^t \mid P_n^t I_n^t C_{A_n}^t) P(C_{A_n}^t \mid A^t)] \\ P(G^t) \prod_{n=1}^N [P(S_n^t) P(\Delta I_n^t) P(I_n^t \mid S_{1:N}^t \Delta I_n^t G^t)] \end{array} \right].$$

In previous inferences, the model $JD_{Percept}$ could be interpreted as a HMM with a complex observation model. This is not the case here, as the JD_{Word} model contains two parallel dynamic models: one over the word space $P(W^t \mid W^{t-1})$, the other over the letter space $P(P_n^t \mid P_n^{t-1})$. This makes JD_{Word} a Hierarchical HMM (HHMM; Murphy, 2002). Of course, since JD_{Word} is nested in JD_{BRAID} , this also applies to our overall BRAID model.

Various techniques exist to perform inference in such structures, and guarantee that the parallel temporal dependencies are handled correctly. There are exact and approximate

methods. For instance, standard Bayesian Network (BN; Pearl, 1988) or Dynamic Bayesian Network (DBN; Dean & Kanazawa, 1989) exact inference techniques, such as clique tree methods, would amount to grouping the W^t and P_n^t variables into a single joint variable, in effect collapsing the two parallel dynamic models into a single one over a more complicated space. This obviously costs memory space and computation time. On the other hand, approximate techniques, in essence, perform iterations of information propagation in the model, until numerical convergence.

We chose a method inspired from such approximate inference methods, performing single-pass inference. Indeed, at each time step, we “cut” the temporal dependency between P_n^t and P_n^{t-1} . This allows to drastically simplify notations. Indeed, we can then consider that the whole “bottom portion” of the model (between stimulus $S_{1:N}^{1:T}$ and letter percepts $P_{1:N}^{0:T}$) is synthesized and replaced by the process of isolated letter recognition we previously described. Recall the notation shorthand for known variables in the right-hand side: $k^t = s_{1:N}^t g^t \mu_A^t \sigma_A^t$. The joint probability distribution then features $Q_{P_n^T} = P(P_n^T | k^{1:T} [\lambda_{P_n^{1:T}} = 1])$ as a term of its (simplified) decomposition¹. It also becomes a conditional joint distribution:

$$\begin{aligned} JD_{Word}^T &= P(W^{0:T} L_{1:N}^{1:T} \lambda_{L_{1:N}}^{1:T} P_{1:N}^{1:T} | k^{1:T} [\lambda_{P_{1:N}^{1:T}} = 1]) \\ &= \left[\begin{array}{c} P(W^0) \prod_{n=1}^N P(P_n^0) \\ \prod_{t=1}^T \left[\begin{array}{c} P(W^t | W^{t-1}) \\ \prod_{n=1}^N \left[\begin{array}{c} P(L_n^t | W^t) \\ P(\lambda_{L_n^t} | L_n^t P_n^t) \\ P(P_n^t | k^{1:T} [\lambda_{P_n^{1:T}} = 1]) \end{array} \right] \end{array} \right] \end{array} \right] \end{aligned}$$

The quality of this approximation, i.e., the discrepancy between the results we obtained and results that would be provided by an exact inference method, is an open question. However, since this approximation mostly concerns the speed of information accumulation in the letter and word spaces, we believe any information lost by our approximation could be counter-balanced by changing the “memory decay” parameters $Leak_P$ and $Leak_W$. Since we calibrated these parameters on experimentally observed dynamics from human participants, it is likely that replacing our approximate inference by an exact inference method would, after recalibration of the model, not change the overall dynamics of the inferences we perform, and thus of the simulation results we obtained.

To compute Q_W^T from JD_{Word}^T , we first marginalize over unknown variables. Then,

¹We note here that we chose to feature $Q_{P_n^T}$, i.e. isolated letter recognition without influence of lexical knowledge. The alternative choice would be to feature $Q'_{P_n^T}$, i.e. isolated letter recognition with influence from lexical knowledge. However, this would not simplify the model, as letter recognition would be influenced by word recognition in order to compute the influence of letter recognition over word recognition: this is another way to see the information propagation loop we chose to approximate.

we rearrange the terms of the summation so that the temporal recursion appears:

$$\begin{aligned}
 Q_W^T &= P(W^T \mid k^{1:T} [\lambda_{L_{1:N}}^{1:T} = 1] [\lambda_{P_{1:N}}^{1:T} = 1]) \\
 &\propto \sum_{W^{0:T-1}, L_{1:N}^{1:T}, P_{1:N}^{1:T}} JD_{Word}^T \\
 &\propto \left[\sum_{W^{T-1}} \left[\sum_{W^{0:T-2}, L_{1:N}^{1:T-1}, P_{1:N}^{1:T-1}} JD_{Word}^{T-1} \right] P(W^T \mid W^{T-1}) \right] \\
 &\propto \left[\prod_{n=1}^N \sum_{L_n^T, P_n^T} \left[\begin{array}{l} P(L_n^T \mid W^T) \\ P([\lambda_{L_n^T} = 1] \mid L_n^T, P_n^T) \\ P(P_n^T \mid k^{1:T} [\lambda_{P_n^{1:T}} = 1]) \end{array} \right] \right]
 \end{aligned}$$

Then, as $\lambda_{L_n^T} = 1$, we can reduce the double sum over L_n^T and P_n^T to a single sum over domain \mathcal{L} :

$$\propto \left[\sum_{W^{T-1}} \left[P(W^T \mid W^{T-1}) Q_{W^{T-1}} \right] \prod_{n=1}^N \sum_{p \in \mathcal{D}_L} \left[\begin{array}{l} P([L_n^T = p] \mid W^T) \\ P([P_n^T = p] \mid k^{1:T} [\lambda_{P_n^{1:T}} = 1]) \end{array} \right] \right]$$

Using $\langle \cdot, \cdot \rangle$ to denote the inner product:

$$Q_W^T \propto \left[\sum_{W^{T-1}} \left[P(W^T \mid W^{T-1}) Q_{W^{T-1}} \right] \prod_{n=1}^N \langle P(L_n^T \mid W^T), P(P_n^T \mid k^{1:T} [\lambda_{P_n^{1:T}} = 1]) \rangle \right]$$

We recognize, in the last expression, the term $P(P_n^T \mid k^{1:T} [\lambda_{P_n^{1:T}} = 1])$, which is $Q_{P_n^T}$: the process of word recognition includes a component which can be interpreted as letter recognition. Therefore, finally:

$$Q_W^T \propto \left[\sum_{W^{T-1}} \left[P(W^T \mid W^{T-1}) Q_{W^{T-1}} \right] \prod_{n=1}^N \langle P(L_n^T \mid W^T), Q_{P_n^T} \rangle \right]$$

4.3 Isolated letter identification with lexical influence: $Q'_{P_n^T}$

In a previous section, we have shown how the probabilistic question $Q_{P_n^T}$ was used to model isolated letter recognition without influence of lexical knowledge. In $Q_{P_n^T}$, the $\lambda_{L_{1:N}}^{1:T}$ coherence variables were left unspecified, i.e. they were “open” Bayesian switches, disconnecting in effect lexical knowledge from inference over variable P_n^T .

We now develop a variant in which we “close” the Bayesian switch between letter percepts P_n^T and lexical knowledge, to model isolated letter recognition with lexical influence. The corresponding probabilistic question is $Q'_{P_n^T} = P(P_n^T \mid k^{1:T} [\lambda_{P_n^{1:T}} = 1] [\lambda_{L_n^{1:T}} = 1])$:

given the stimulus $S_{1:N}^{1:T}$, given attention distribution parameters $\mu_A^{1:T}, \sigma_A^{1:T}$ and gaze position $G^{1:T}$, and given that information propagates in the model in a bottom-up manner from stimulus to the perceived letter at position n ($[\lambda_{P_n}^{1:T} = 1]$), but also in a top-down manner from lexical knowledge to the perceived letter at position n ($[\lambda_{L_n}^{1:T} = 1]$), what is the probability distribution over the perceived letter at position n , P_n^T ?

Word recognition, with the probabilistic question Q_W^T , and letter recognition here with $Q'_{P_n^T}$ both involve the same portion of the model. As when solving Q_W^T , the issue arises to deal with the coupled Markov models (one over variables P_n^t , the other over variables W^t). In word recognition, we had chosen to “decouple” the Markov chains, resulting in a strict bottom-up flow of information, where percept variables P_n^t would inform about the word variable W^t , but not the other way around. Following the same strategy for our approximated inference would of course make impossible to have lexical knowledge inform letter identity. Indeed, a way to interpret exact inference would be to have a two-way exchange of probabilistic information between W^t and P_n^t , at each time step t , until convergence. Instead of performing this, which is costly in terms of computation costs, we introduce a one-pass top-down propagation of information, at the last time step considered.

To derive our solution for $Q'_{P_n^T}$, we first make the joint probability distribution JD_{Word}^T appear by marginalizing over missing variables. Then, we reorganize terms and sums to make the temporal recursion appear:

$$\begin{aligned} Q'_{P_n^T} &= P(P_n^T \mid k^{1:T} [\lambda_{L_{1:N}}^{1:T} = 1] [\lambda_{P_n}^{1:T} = 1]) \\ &\propto \sum_{\substack{w^{0:T}, \\ l_{1:N}^{1:T}, \\ p_{1:N}^{1:T-1}, p_{m \neq n}^T}} JD_{Word}^T \\ &\propto \sum_{\substack{w^{T-1:T}, \\ l_{1:N}^T, \\ p_{m \neq n}^T}} \left[\left[\begin{array}{c} P(W^T \mid w^{T-1}) \\ \sum_{\substack{w^{0:T-2}, \\ l_{1:N}^{1:T-1}, \\ p_{1:N}^{1:T-1}}} JD_{Word}^{T-1} \end{array} \right] \left[\begin{array}{c} \prod_{\substack{m=1, \\ m \neq n}}^N P(l_m^T \mid w^T) \\ P(\lambda_{L_m}^T \mid l_m^T p_m^T) \\ P(p_m^T \mid k^{1:T}) \end{array} \right] \left[\begin{array}{c} P(l_n^T \mid w^T) \\ P(\lambda_{L_n}^T \mid l_n^T P_n^T) \\ P(P_n^T \mid k^{1:T} [\lambda_{P_n}^{1:T} = 1]) \end{array} \right] \right] \end{aligned}$$

We can now make $Q_W^{T-1}(w^{T-1})$ appear, that corresponds to the second summation over JD_{Word}^{T-1} . Reorganizing and simplifying further:

$$\begin{aligned} Q'_{P_n^T} &\propto \sum_{w^{T-1:T}} \left[\left[\begin{array}{c} P(W^T \mid w^{T-1}) \\ Q_W^{T-1}(w^{T-1}) \end{array} \right] \prod_{\substack{m=1, \\ m \neq n}}^N \left[\begin{array}{c} P(l_m^T \mid w^T) \\ P(\lambda_{L_m}^T \mid l_m^T p_m^T) \\ P(p_m^T \mid k^{1:T}) \end{array} \right] \right] \sum_{l_n^T} \left[\begin{array}{c} P(l_n^T \mid w^T) \\ P(\lambda_{L_n}^T \mid l_n^T P_n^T) \\ P(P_n^T \mid k^{1:T} [\lambda_{P_n}^{1:T} = 1]) \end{array} \right] \\ &\propto \sum_{w^T} \left[\left[\begin{array}{c} \sum_{w^{T-1}} [Q_W^{T-1}(w^{T-1}) P(W^T \mid w^{T-1})] \\ \prod_{\substack{m=1, \\ m \neq n}}^N \langle P(L_m^T \mid w^T), P(P_m^T \mid k^{1:T}) \rangle \end{array} \right] \left[\begin{array}{c} \sum_{l_n^T} P(l_n^T \mid w^T) P(\lambda_{L_n}^T \mid l_n^T P_n^T) P(P_n^T \mid k^{1:T} [\lambda_{P_n}^{1:T} = 1]) \end{array} \right] \right] \end{aligned}$$

The first two lines correspond to $Q_W^T(w^T)$, without considering a similarity between letter percepts and words. This is where we temporarily ignore the top-down influence of lexical knowledge on letter recognition.

Noticing that $\langle P(L_n^T | W^T), P(P_n^T | k^{1:T} [\lambda_{P_n^{1:T}} = 1]) \rangle$ is never 0, we write:

$$Q_{P_n}^T \propto \sum_{w^T} \left[\frac{Q_W^T(w^T) / \langle P(L_n^T | W^T), P(P_n^T | k^{1:T} [\lambda_{P_n^{1:T}} = 1]) \rangle}{\sum_{l_n^T} P(l_n^T | w^T) P(\lambda_{L_n^T} | l_n^T P_n^T) P(P_n^T | k^{1:T} [\lambda_{P_n^{1:T}} = 1])} \right]$$

We now approximate our solution by multiplying the first line by $\langle P(L_n^T | W^T), P(P_n^T | k^{1:T} [\lambda_{P_n^{1:T}} = 1]) \rangle$, thus simplifying it to $Q_W^T(w^T)$. This choice amounts to considering the word recognition process terminated, before “sending” the probability distribution over words to inform letter identity. This approximation allows to write:

$$Q'_{P_n}{}^T \approx \sum_{w^T} \left[Q_W^T(w^T) \sum_{l_n^T} P(l_n^T | w^T) P(\lambda_{L_n^T} | l_n^T P_n^T) P(P_n^T | k^{1:T} [\lambda_{P_n^{1:T}} = 1]) \right]. \quad (3)$$

Recognizing that $P([P_n^T = p] | k^{1:T} [\lambda_{P_n^{1:T}} = 1])$ is letter identification without lexical influence, $Q_{P_n}^T$:

$$Q'_{P=p_n}{}^T \approx \sum_{w^T} \left[Q_{W=w^T}{}^T P([L_n^T = p] | w^T) Q_{P=p_n}{}^T \right]. \quad (4)$$

Compared with letter recognition without lexical influence, the inference for $Q'_{P_n}{}^T$ features a marginalization over the word space w^T , so that information about words influence the probability distribution over letter percepts, $P(P_n^T | k^{1:T} [\lambda_{P_n^{1:T}} = 1]) = Q_{P_n}^T$, computed without lexical influence. To understand intuitively how lexical information combines with information about letters, consider distribution $P(l_n^t | w^t)$: this is a constant distribution, that is to say, independent of time index t . This distribution is almost 0 everywhere, except for the correct letter, that is to say, the letter at position n of the spelling of word w^t . The second part of Equation (3) can thus be interpreted as the comparison between letters predicted by word w^t , and letters perceived from the stimulus, $P(P_n^T | k^{1:T} [\lambda_{P_n^{1:T}} = 1])$, for all positions. This comparison is not performed for a single recognized word, but for all words (\sum_{w^T}) according to their probability $Q_W^T(w^T)$ to be recognized in the stimulus. This is where we obtain a two-way exchange of information: stimulus processing informs percepts, which inform words, and which, in a one-pass top-down influence, inform percepts.

4.4 Lexical decision: Q_D^T

The fourth and final cognitive task we consider is lexical decision, mathematically modeled by the probabilistic question $Q_D^T = P(D^T | S_{1:N}^{1:T} [\lambda_{D_{1:N}}^{1:T} = 1] [\lambda_{P_{1:N}}^{1:T} = 1] \mu_A^{1:T} \sigma_A^{1:T} G^{1:T})$: given the stimulus $S_{1:N}^{1:T}$, given attention distribution parameters $\mu_A^{1:T}, \sigma_A^{1:T}$ and gaze position $G^{1:T}$, and given that information propagates throughout the whole model, what is the probability that $D^T = T$, i.e., that the input is a known word?

In this question, observe that the $\lambda_{L_{1:N}}^{1:T}$ coherence variables are not specified. Contrary to the precedent cases, this does not allow us to simplify portions of the model, as the $\lambda_{L_{1:N}}^{1:T}$ variables do not play the role of Bayesian switches here. Indeed, in Q_D^T , the $\lambda_{D_{1:N}}^{1:T}$ coherence variables are, for the first time in this manuscript, closed, and they connect the $\lambda_{L_{1:N}}^{1:T}$ and $C_{D_{1:N}}^{1:T}$ (i.e. they appear on the right-hand side of the terms $P(\lambda_{D_n}^t | \lambda_{L_n}^t C_{D_n}^t)$ in JD_{BRAID}). This means that probability distributions over $\lambda_{L_{1:N}}^{1:T}$ and $C_{D_{1:N}}^{1:T}$ are (implicitly) computed during inference, and their correspondence evaluated by the coherence term.

Therefore, we deal here with the complete model, with its three dynamic models (over letter percepts P_n^t , over words W^t and over lexical membership D^t) that evolve in a coupled manner. To solve Q_D^T , we follow the same strategy as before, by isolating letter recognition, and answering first $Q_{P_n^T}$ at all positions n . Then, in the general model, we replace the temporal dynamic model over P_n^T by the answer to Q_{P^T} . As before, this “decouples” the letter dynamic model from the other two. This yields the following rewriting of the model joint probability distribution:

$$\begin{aligned} JD_{LD}^T &= P(W^{0:T} L_{1:N}^{1:T} \lambda_{L_{1:N}}^{1:T} P_{1:N}^{0:T} \lambda_{D_{1:N}}^{1:T} C_{D_{1:N}}^{1:T} D^{0:T} | k^{1:T} [\lambda_{P_n}^{1:T} = 1]) \\ &= P(W^0) \prod_{t=1}^T \left[\begin{array}{l} P(W^t | W^{t-1}) \\ \prod_{n=1}^N \left[\begin{array}{l} P(L_n^t | W^t) \\ P(\lambda_{L_n}^t | L_n^t P_n^t) \\ P(P_n^t | k^{1:T} [\lambda_{P_n}^{1:T} = 1]) \end{array} \right] \\ \prod_{n=1}^N P(\lambda_{D_n}^t | \lambda_{L_n}^t C_{D_n}^t) \\ P(D^t | D^{t-1}) P(C_{D_{1:N}}^t | D^t) \end{array} \right] \end{aligned}$$

We also decouple the dynamic model over words in a similar manner, replacing it by the answer to word recognition Q_W^T . This yields the final form of the model that we consider for lexical decision:

$$JD_{LD}^T = \prod_{t=1}^T \left[\begin{array}{l} P(W^t | K^{1:t-1} \lambda_{L_{1:N}}^{1:t-1}) \\ \prod_{n=1}^N \left[\begin{array}{l} P(L_n^t | W^t) \\ P(\lambda_{L_n}^t | L_n^t P_n^t) \\ P(P_n^t | k^{1:t} [\lambda_{P_n}^{1:T} = 1]) \end{array} \right] \\ \prod_{n=1}^N P(\lambda_{D_n}^t | \lambda_{L_n}^t C_{D_n}^t) \\ P(D^t | D^{t-1}) P(C_{D_{1:N}}^t | D^t) \end{array} \right] \quad (5)$$

To solve Q_D^T , we first introduce missing variables by marginalizations, in order to

make the joint probability distribution appear:

$$\begin{aligned}
 Q_D^T &= P(D^T \mid k^{1:T} [\lambda_{D_{1:N}}^{1:T} = 1] [\lambda_{P_{1:N}}^{1:T} = 1]) \\
 &\propto \sum_{\substack{w^{1:T} \\ l_{1:N}^{1:T} \\ p_{1:N}^{1:T}}} \sum_{\substack{d^{0:T-1} \\ c_{D_{1:N}}^{1:T}}} JD_{LD}^T \\
 &\propto \left[\sum_{d^{T-1}} \left[\left[\sum_{\substack{w^{1:T-1} \\ l_{1:N}^{1:T-1} \\ p_{1:N}^{1:T-1}}} \sum_{\substack{d^{0:T-2} \\ c_{D_{1:N}}^{1:T-1}}} JD_{LD}^{T-1} \right] P(D^T \mid d^{T-1}) \right] \right] \\
 &\propto \left[\sum_{\lambda_{L_{1:N}}^T} \left[\sum_{\substack{w^T \\ l_{1:N}^T \\ p_{1:N}^T}} \left[\begin{aligned} &P(w^T \mid k^{1:T-1} [\lambda_{L_{1:N}}^{1:T-1} = 1] [\lambda_{P_{1:N}}^{1:T} = 1]) \\ &\prod_{n=1}^N \left[\begin{aligned} &P(l_n^T \mid w^T) \\ &P(\lambda_{L_n}^T \mid l_n^T p_n^T) \\ &P(p_n^T \mid k^{1:T} [\lambda_{P_n}^{1:T} = 1]) \end{aligned} \right] \end{aligned} \right] \right] \\
 &\quad \left[\begin{aligned} &P(c_{D_{1:N}}^T \mid D^T) \\ &\prod_{n=1}^N P(\lambda_{D_n}^T \mid \lambda_{L_n}^T c_{D_n}^T) \end{aligned} \right] \right] \right]
 \end{aligned}$$

Reorganizing terms to make appear the temporal recursion term Q_D^{T-1} , we obtain:

$$\begin{aligned}
 Q_D^T &\propto \left[\sum_{d^{T-1}} \left[Q_D^{T-1} P(D^T \mid d^{T-1}) \right] \right] \\
 &\propto \left[\sum_{\lambda_{L_{1:N}}^T} \left[\sum_{w^T} \left[\prod_{n=1}^N \sum_{\substack{l_n^T \\ p_n^T}} \left[\begin{aligned} &P(w^T \mid k^{1:T-1} [\lambda_{L_{1:N}}^{1:T-1} = 1] [\lambda_{P_{1:N}}^{1:T} = 1]) \\ &P(l_n^T \mid w^T) \\ &P(\lambda_{L_n}^T \mid l_n^T p_n^T) \\ &P(p_n^T \mid k^{1:T} [\lambda_{P_n}^{1:T} = 1]) \end{aligned} \right] \right] \right] \\
 &\quad \left[\begin{aligned} &P(c_{D_{1:N}}^T \mid D^T) \\ &\prod_{n=1}^N P(\lambda_{D_n}^T \mid \lambda_{L_n}^T c_{D_n}^T) \end{aligned} \right] \right] \right]
 \end{aligned}$$

We now consider separately the two cases for question Q_D^T , that is to say, since variable D^T is Boolean, the case $D^T = \text{True}$ first, and $D^T = \text{False}$ second.

Assuming that $D^T = \text{True}$ amounts to consider that all coherence variables $\lambda_{L_{1:N}}^T$ are “closed”, because they are controlled by control variables $C_{D_{1:N}}^T$, which are all to 1 when $D^T = \text{True}$. The summation over $C_{D_{1:N}}^T$ thus collapses to a unique value. This propagates in the model, to variables $\lambda_{L_{1:N}}^T$, which are also “closed”, thus also collapsing

their summation. This yields:

$$\begin{aligned}
 Q_{D=True}^T &= P([D^T = True] | k^{1:T} [\lambda_{D_{1:N}}^{1:T} = 1] [\lambda_{P_{1:N}}^{1:T} = 1]) \\
 &\propto \left[\begin{aligned} &\sum_{d^{T-1}} [P([D^T = True] | d^{T-1}) Q_{D=d^{T-1}}^{T-1}] \\ &\sum_{w^T} \left[\begin{aligned} &P(w^T | k^{1:T-1} [\lambda_{L_{1:N}}^{1:T-1} = 1] [\lambda_{P_{1:N}}^{1:T} = 1]) \\ &\prod_{n=1}^N \langle P(L_n^T | w^T), Q_{P_n^T} \rangle \end{aligned} \right] \end{aligned} \right] \quad (6)
 \end{aligned}$$

The case $Q_{D=False}^T$ is more complicated. Indeed, variables $C_{D_{1:N}}^T$ now can take several values (enumerating all possible positions of a single error in the stimulus, that is, all $C_{D_{1:N}}^T$ are 1 except one, which is 0). This yields N configurations to consider, which we consider equally probable ($P(C_{D_{1:N}}^T | [D^T = False]) = \frac{1}{N}$ when $C_{D_n}^T = 1 \forall n \neq i$, and $C_{D_i}^T = 0$, and $P(C_{D_{1:N}}^T | [D^T = False]) = 0$ otherwise).

$$\begin{aligned}
 Q_{D=False}^T &= P([D^T = False] | k^{1:T} [\lambda_{D_{1:N}}^{1:T} = 1] [\lambda_{P_{1:N}}^{1:T} = 1]) \\
 &\propto \left[\begin{aligned} &\sum_{d^{T-1}} [P([D^T = False] | d^{T-1}) Q_{D=d^{T-1}}^{T-1}] \\ &\frac{1}{N} \sum_{i=1}^N \left[\begin{aligned} &\sum_{w^T} \left[\begin{aligned} &P(w^T | k^{1:T-1} [\lambda_{L_i}^{1:T-1} = 0] [\lambda_{L_{n \neq i}}^{1:T-1} = 1] [\lambda_{P_{1:N}}^{1:T-1} = 1]) \\ &\prod_{\substack{n=1 \\ n \neq i}}^N \langle P(L_n^T | w^T), Q_{P_n^T} \rangle \\ &\langle P(L_i^T | w^T), 1 - Q_{P_i^T} \rangle \end{aligned} \right] \end{aligned} \right] \end{aligned} \right] \quad (7)
 \end{aligned}$$

Once $Q_{D=True}^T$ and $Q_{D=False}^T$ are computed, we normalize them, and thus obtain a probability distribution over variable D^T . Therefore, when used at the next time step, this ensures that probability values are used during the combination with the dynamic term $P(D^{T+1} | D^T)$.

Here is a technical surprise: Equations (6) and (7) do not exactly match the implementation of the BRAID model. Indeed, it was found that the resulting dynamics of these equations were not satisfying. During Thierry Phénix's PhD thesis, an empirical solution was found, which consisted in dividing the inner products that appear in these equations by the cardinal $|\mathcal{D}_L|$ of the letter space (i.e., in practice, 27); this solution was applied without studying its theoretical justification. This has been the focus of a later paper (Steinhilber et al., 2022), which suggests that dividing by the L2-norm of one of the probability distributions (instead of the cardinal of the set) provides a variant of the similarity operator (the inner products) that accounts and corrects for the entropy of the compared distributions. Since the cardinal and L2-norm were numerically close, the current implementation involves dividing by the L2-norm, instead.

5. Control variables

Control variables are an additional tool to coherence variables for expressing modularity in Bayesian algorithmic models, due to Jacques Droulez (Droulez, 2015). Both are

binary variables that link sub-models in probabilistic dependency graphs, that are used for expressing whether sub-models they connect share information or not. Control variables, as coherence variables, can be interpreted as ‘‘Bayesian switches’’. However, whereas coherence variables allow controlling the switch state using probabilistic questions to the overall model, control variables explicitly set, by their value, the switch state.

This allows a more natural semantic of the variable: using coherence variable λ , sub-models A and B are connected, i.e., the Bayesian switch is closed, by setting the constraint $\lambda = 1$ in the right-hand side of a given probabilistic question, and A and B are disconnected by leaving variable λ unconstrained in the probabilistic question. In contrast, using control variable ξ , sub-models A and B are connected whenever $\xi = 1$, and disconnected whenever $\xi = 0$.

Control variables act in conjunction with coherence variables. We now provide the mathematical definition of control variables and coherence variables as Bayesian switches, and demonstrate their properties.

Let λ and ξ be two binary variables, A and B be probabilistic variables with the same, arbitrary domain (although this can be generalized easily). ξ is a control variable if it appears, in the decomposition of a joint probability distribution, in a term of the form $P(\lambda \mid A B \xi)$, defined by:

$$\begin{aligned} P([\lambda = 1] \mid [A = a] [B = b] [\xi = c]) \\ = \begin{cases} 1 & \text{if } c = 1 \text{ and } a = b \\ 0 & \text{if } c = 1 \text{ and } a \neq b \\ \theta_\xi & \text{if } c = 0. \end{cases} \end{aligned}$$

To demonstrate how the value of a control variable ξ connects or disconnects sub-models A and B , we take the example of a joint probability distribution $P(A B \lambda \xi)$ defined by:

$$P(A B \lambda \xi) = P(A)P(B)P(\xi)P(\lambda \mid A B \xi) .$$

This is without loss of generality; in the general case, A and B represent ‘‘gateways’’ to arbitrarily complex probabilistic models, with additional variables and probabilistic terms. However, they do not affect the local property of control variables.

We now demonstrate that setting $\xi = 1$ connects sub-models A and B , i.e., computing $P(A \mid [\lambda = 1] [\xi = 1])$ involves $P(B)$. Bayesian inference yields:

$$\begin{aligned} P([A = a] \mid [\lambda = 1] [\xi = 1]) \\ = \frac{P([A = a] [\lambda = 1] [\xi = 1])}{P([\lambda = 1] [\xi = 1])} \\ = \frac{\sum_B P(A)P(B)P([\xi = 1])P([\lambda = 1] \mid A B [\xi = 1])}{\sum_{A,B} P(A)P(B)P([\xi = 1])P([\lambda = 1] \mid A B [\xi = 1])} \\ = \frac{P([\xi = 1])P([A = a])P([B = a])}{P([\xi = 1]) \sum_{a' \in A} P([A = a'])P([B = a'])} \\ = \frac{P([A = a])P([B = a])}{\sum_{a' \in A} P([A = a'])P([B = a'])} . \end{aligned}$$

Finally, we demonstrate that setting $\xi = 0$ disconnects sub-models A and B . i.e., computing $P(A \mid [\lambda = 1] [\xi = 0])$ does not involve $P(B)$. Bayesian inference yields:

$$\begin{aligned}
 & P([A = a] \mid [\lambda = 1] [\xi = 0]) \\
 &= \frac{P([A = a] [\lambda = 1] [\xi = 0])}{P([\lambda = 1] [\xi = 0])} \\
 &= \frac{\sum_B P(A)P(B)P([\xi = 0])P([\lambda = 1] \mid A B [\xi = 0])}{\sum_{A,B} P(A)P(B)P([\xi = 0])P([\lambda = 1] \mid A B [\xi = 0])} \\
 &= \frac{\theta_\xi P([\xi = 0])P([A = a])}{\theta_\xi P([\xi = 0]) \sum_{A,B} P(A)P(B)} \\
 &= P([A = a]) .
 \end{aligned}$$

We note that, during this inference, the probability value θ_ξ assigned to $P([\lambda = 1] \mid A B [\xi = 0])$ is irrelevant, as it can always be factored out and simplified. However, when marginalizing over ξ , this value has an effect. Indeed, consider computing:

$$\begin{aligned}
 & P([A = a] \mid [\lambda = 1]) \\
 &\propto \sum_{\xi, B} P([A = a] B [\lambda = 1] \xi) \\
 &\propto \sum_{\xi, B} P([A = a])P(B)P(\xi)P([\lambda = 1] \mid [A = a] B \xi) \\
 &\propto P([A = a]) \left(\frac{P([\xi = 0]) \sum_B P(B)P([\lambda = 1] \mid [A = a] B [\xi = 0])}{+P([\xi = 1]) \sum_B P(B)P([\lambda = 1] \mid [A = a] B [\xi = 1])} \right) \\
 &\propto P([A = a]) (P([\xi = 0])\theta_\xi + P([\xi = 1])P([B = a])) .
 \end{aligned}$$

In that result, the prior probability distribution $P(\xi)$ serves as a weighting factor between an “open” ($\xi = 0$) and a “closed” ($\xi = 1$) mode of information transfer, which are combined by the summation. On the one hand, with factor $P([\xi = 1])$, submodels are connected, and $P(A)$ is multiplied by $P(B)$. On the other hand, with factor $P([\xi = 0])$, submodels are disconnected, and $P(A)$ is multiplied by a constant value θ_ξ . This can also be interpreted as a connected mode, where A would be connected to B , but B would be replaced by a uniform model. In that case, the “uniform” model over variable B should provide a numeric value consistent with a uniform distribution over B , i.e., $1/|B|$.

Indeed, in the context of the BRAID model, control variables $C_{A_n}^t$ are involved in the attention model, which connects variables I_n^t and P_n^t by coherence variables $\lambda_{P_n}^t$. Variables I_n^t and P_n^t are defined over the letter domain \mathcal{D}_L , and so we have defined $P([\lambda_{P_n}^t = 1] \mid [P_n^t = l_P] [I_n^t = l_I] [C_{A_n}^t = c])$ to be $1/|\mathcal{D}_L|$ when $c = 0$.

References

- Bessi ere, P., Mazer, E., Ahuactzin, J. M., & Mekhnacha, K. (2013). *Bayesian programming*. CRC Press.
- Dean, T., & Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3), 142–150.
- Droulez, J. (2015). *Coherence variables (suite)* [personal communication, 2015].

- Gilet, E., Diard, J., & Bessière, P. (2011). Bayesian action-perception computational model: Interaction of production and recognition of cursive letters. *PLoS ONE*, *6*(6), e20387.
- Lebeltel, O., Bessière, P., Diard, J., & Mazer, E. (2004). Bayesian robot programming. *Autonomous Robots*, *16*(1), 49–79.
- Murphy, K. (2002, July). *Dynamic Bayesian networks: Representation, inference and learning* [Ph.D. thesis]. University of California, Berkeley.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.
- Phénix, T., Ginestet, É., Valdois, S., & Diard, J. (2025). Visual attention matters during word recognition: A Bayesian modeling approach. *Psychonomic Bulletin & Review*.
- Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice Hall.
- Steinhilber, A., Valdois, S., & Diard, J. (2022). Bayesian comparators: A probabilistic modeling tool for similarity evaluation between predicted and perceived patterns. *44th Annual Meeting of the Cognitive Science Society*, 2264–2270.