Relating Online Word Predictions and Offline Probability Judgments: Partial Evidence for Shared Probabilistic Processes

Yutong Zhang (yutong.zhang@ling-phil.ox.ac.uk)[1] **& E. Matthew Husband**[1]
[1]Faculty of Linguistics, Philology and Phonetics, University of Oxford

Abstract

Probabilistic reasoning has been proposed to underlie both word prediction in language comprehension and probability judgment in decision-making, yet the connection between the two cognitive processes remains unclear. This study investigates whether word predictions and probability judgments rely on shared probabilistic processes by examining unpacking effects. An offline probability judgment task and an online self-paced reading task were used to test probability shifts in each task, and further analyses examined the relationship between measures across tasks. Results show that atypical unpacking reliably led to lower probability judgments and slower reading times, whereas typical unpacking didn't consistently lead to higher probability judgments or faster reading times. Interestingly, item-level changes in probability judgments were correlated with changes in reading times only for atypical unpackings. These findings provide partial evidence that word prediction and probability judgment draw on a shared foundation of probabilistic reasoning.

Keywords: Language Comprehension; Decision Making; Predictive Processing; Reasoning

Introduction

Probability plays a fundamental role in contemporary cognitive science, offering a way to characterize how minds work under uncertainty. In language comprehension, uncertainty arises naturally from the incremental nature of input. Unexpected continuations require more effort from comprehenders as a sentence unfolds, reflected in measures of processing difficulty. Probabilistic approaches account for these findings by proposing that comprehenders assign probabilities to possible continuations and use these probabilistic expectations to generate predictions (Kuperberg & Jaeger, 2016).

The idea that comprehenders allocate probabilities over upcoming continuations has gained much support over the last twenty years. Most prominently, Surprisal Theory proposes that word probability is a main determinant of processing difficulty, with the processing difficulty of a word being proportional to its **surprisal**, as given in (1) (Hale, 2001; Levy, 2008).[1] Empirical evidence for this predictive process has been observed both behaviorally and neurally. Lower-probability words are associated with slower reading (Shain, 2024; Smith & Levy, 2013) and increased brain activity (Michaelov et al., 2023; Shain et al., 2020).

$$RT \propto -\log P(W \mid Cxt) \tag{1}$$

[1] RT represents reading time, as a measure of processing difficulty, W the word, and Cxt the context.

This important connection between a word's probability and its processing difficulty raises the question of how word probabilities are obtained during real-time processing. At present, word probabilities are operationalized using human tasks like cloze completion or measured by Large Language Models (LLMs). In cloze, for example, humans are asked to generate possible continuations of a context, and word probabilities are estimated from the relative frequencies of the generated responses. However, given the size of the adult vocabulary (~30,000 to 50,000 words, Brysbaert et al., 2016), it is highly unlikely that human comprehenders implement this procedure by computing the probability of each possible word in real time. And while LLMs do compute the probability of each possible word, the way they do so is unlikely to reflect human processes. Therefore, how humans arrive at the probabilities that guide their language processing behavior remains to be determined.

Predicting upcoming words is not the only situation in which humans reason probabilistically. Probability judgment studies in the judgment and decision-making literature task people to explicitly estimate the probability of events. Decades of research in probability judgment have shown that people are systematically biased, with different statements about the same underlying event receiving different probability judgments.

Unpacking is one important case of bias, where introducing category members as a contextual anchor can systematically shift the probability judgment of the category. Much work has shown that the typicality of anchors modulates the direction of this shift (Dasgupta et al., 2017; Hadjichristidis et al., 2001; Sloman et al., 2004; Tversky & Koehler, 1994). For example, participants may be asked, *"What is the probability that a person dies from disease?"* Unpacking the category *disease* by listing category members does not change the underlying event, however, participants assigned higher probabilities when it was unpacked with typical anchors (e.g., *"What is the probability that a person dies from heart disease, cancer, stroke, or any other disease?"*) (Tversky & Koehler, 1994), and lower probabilities when it was unpacked with atypical anchors (e.g., *"What is the probability that a person dies from pneumonia, diabetes, cirrhosis, or any other disease?"*) (Sloman et al., 2004). In general, typical anchors shift probability judgments upward and atypical anchors shift them downward. The implicit unpacking effects can be formalized as follows[2]:

[2] P denotes the subjective probability, T represents the target cat-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

$$P(T \mid C) < P(T \mid C = C \cup \{A_{typical}\}) \quad (2)$$

$$P(T \mid C) > P(T \mid C = C \cup \{A_{atypical}\}) \quad (3)$$

Beyond anchor typicality, the target category also appears to modulate unpacking effects. Sloman et al. (2004) examined the effects using different types of target categories, including natural categories (e.g., *disease*), ad hoc categories (e.g., *geometric shapes corresponding to building top views*), and fuzzy categories (e.g., *guns sold in hardware stores*). Although introducing atypical anchors reliably reduced probability judgments across category types, introducing typical anchors did not consistently increase probability judgments, suggesting that the effects of typical anchors are less robust.

Notably, these target categories differ in their level of abstraction. According to Rosch et al. (1976), categories such as *gun* are basic-level and the most frequently used, whereas categories such as *disease* and *shape* are superordinate-level and more abstract. This pattern suggests that variability in unpacking effects, particularly with the typical unpacking effect, may be related to the category level of the target. However, this possibility has not yet been systematically tested.

There is a deep theoretical literature addressing both humans' capacity for effective probabilistic reasoning and their vulnerability to systematic biases. Influential accounts can be broadly divided into two lines of work. One line focuses on Marr's (1982) **computational level**, proposing that people do not compute classical probabilities, but instead rely on substitute mental shortcuts as proxies for probability (Kahneman, 2011; Kahneman & Tversky, 1972; Tversky & Kahneman, 1973). The heuristic model, for example, uses substitute measures to explain unpacking effects via availability. According to this framework, things that are more easily come to mind are judged as more probable. As typical anchors are common, they are readily retrievable from memory. In contrary, atypical anchors are rare, and thus harder to think of. As a result, when an category is unpacked, typical anchors make it easier to recall, leading to higher probability judgments, whereas atypical anchors make it harder to recall, leading to lower judgments. Other models such as Probability Plus Noise or Bayesian Sampler models (Chater et al., 2020; Costello & Watts, 2014; Zhu et al., 2020) are at Marr's (1982) **algorithmic level**, emphasizing the processes humans use to approximate probability, as exact probabilistic computation is intractable due to limited resources. Within Bayesian Samplers, for example, unpacking effects are understood as anchors influencing noisy samples taken from memory. When estimating the probability of a packed event, the sampler starts at a relatively random position. Introducing typical anchors places the sampler toward high-probability regions within the target category, resulting in a greater proportion of samples falling within high-probability regions. In contrast, atypical anchors place the sampler toward low-probability regions, resulting in fewer samples in high-probability regions. Because probability judgments are assumed to depend on the probabilities of the samples, typical anchors lead to higher probability judgments, whereas atypical anchors lead to lower probability judgments.

In this study, we ask whether the ways that humans judge probabilities offline is related to the probabilities that underlie online word prediction. Establishing such a link would connect the question of how comprehenders generate word probabilities to the well-studied question of how humans make explicit probability judgments, allowing us to use judgment and decision-making accounts to understand word prediction in language comprehension, while also using time-sensitive language-processing accounts to inform models of probability judgment.

If the two tasks share probabilistic processes, the probability shifts described in (2) and (3) should also be reflected in surprisal during reading, assuming that C is related to the meaning of Cxt. Linking the shifts in (2) and (3) into (1) predicts the following relationships, with introducing typical anchors predicted to reduce surprisal (4) and introducing atypical anchors predicted to increase surprisal (5):

$$-\log P(W \mid Cxt) > -\log P(W \mid Cxt + A_{typical}) \quad (4)$$

$$-\log P(W \mid Cxt) < -\log P(W \mid Cxt + A_{atypical}) \quad (5)$$

Further, if probability shifts in the two tasks scale together, we predict a negative correlation[3] between changes in item-level probability judgments (ΔPJ) and word predictions (ΔRT), given in (8).

$$\Delta PJ = P(T \mid C \cup A) - P(T \mid C) \quad (6)$$

$$\Delta RT \propto (-\log P(W \mid Cxt + A)) - (-\log P(W \mid Cxt)) \quad (7)$$

$$\Delta PJ \sim -\Delta RT \quad (8)$$

To address these questions, we examine unpacking effects in both probability judgment and word prediction. In a probability judgment task, we test how judgments change when typical or atypical anchors are introduced into question frames, addressing whether (2) and (3) hold for more garden-variety statements compared to more targeted cases commonly found in the probability judgment literature. In a self-paced reading task, we test how reading times change when contexts are unpacked using the same methods, addressing whether (4) and (5) also hold for maximally similar sentences. Finally, as unpacking effects essentially concern changes induced by anchors, we also examine whether changes in reading times are related to changes in probability judgments, addressing whether (8) holds.

While additional differences in task between online reading and offline judgment may further influence behavioral measures, a shared processes account concerns that the same change in context will produce the same change in probability across the two tasks. We predicted that, if word prediction and probability judgment rely on a shared probabilistic

egory being evaluated (e.g., *disease*), C the category described by the question frame (e.g., for *"What is the probability that a person dies from ...?"*, $C = \{x :$ a person dies from $x\}$), and A the anchors.

[3]The correlation is negative because increases in surprisal reflect decreases in probability.

foundation, the probability shifts induced by unpacking effects will arise in both domains. Specifically, typical anchors should lead to higher probability judgments and faster reading times, whereas atypical anchors should lead to lower probability judgments and slower reading times. Furthermore, if judged probabilities in word prediction and probability judgment are related, increases in probability judgments should be correlated with decreases in reading times.

Experiments

Materials for Experiments 1 and 2 Experimental materials were designed to be as similar as possible across the probability judgment and self-paced reading tasks. An initial 66 contexts were selected from Peelle et al. (2020)'s cloze norming study. Each context had a moderate-cloze category term response, which was used as the target word completing the context (33 basic-level, 33 superordinate-level), modified to describe repeatable or habitual events. For each target, two anchors were selected from Van Overschelde et al. (2003)'s category norming study: a typical anchor, which was both representative of the category and highly plausible in the context, and an atypical anchor, which was a less representative category member but remained contextually plausible.

Anchor typicality was normed by 20 additional participants who did not participate in either of the other studies. Participants saw a packed sentence and one of its anchor words and were asked to decide whether that anchor word could be a possible member of the target word (bolded) given the sentence context, followed by a typicality judgment on a 7-point scale from 1 (very atypical) to 7 (very typical). Trials in which typical or atypical anchors were judged to be impossible referents of the target word were excluded from the analysis. Six items were removed because their typical anchor received lower typicality ratings than their atypical anchor. For the 60 remaining items, ratings for typical anchors ($M = 6.40, SD = 0.65$) were higher than atypical anchors ($M = 4.90, SD = 1.17$), suggesting that they were perceived as more typical category members of the target. This difference in rating scores was statistically significant ($t = 8.95, p < .001$).

The 60 experimental items followed a 3x2 design: three conditions based on the typicality of anchor and two groups based on the category level of target, as shown in Table 1. Anchors were conjoined to the target category word by "*or any other*" following traditional unpacking manipulations. In addition, 60 filler items were included. A complete set of items can be found on OSF.

Experiment 1: Probability Judgment Task

Participants 36 participants were recruited on Prolific. Participants were paid at a rate of £6 per hour for their participation. All participants were native British English speakers and provided explicit consent before participating.

Materials The 60 experimental and 60 filler items were modified into explicit probability judgment questions by adding "*What is the probability that*".

Procedure Participants were presented probability questions and indicated their probability judgments by dragging a slider continuously ranging from 0% to 100%. One-third of the items were followed by a comprehension question, which tested whether participants paid full attention.

Analysis Probability judgment (PJ) data was analyzed using linear mixed-effects models implemented in `lme4` (version 1.1–35.2; Bates et al., 2015). Typicality was coded into two sum-contrasts: Typical (0.5) vs. Packed (-0.5) and Atypical (0.5) vs. Packed (-0.5). CatLevel was also sum-coded, Basic (0.5) vs. Superordinate (-0.5).

- *PJ ~ Typicality * CatLevel + (1 + Typicality * CatLevel | Participant) + (1 + Typicality | Item)*

Results Table 2 reports the linear mixed-effects model results for all items. As the experiment was also designed to examine whether unpacking effects vary across category levels, and a marginal interaction was observed between the typical unpacked condition and category level, we further conducted separate linear mixed-effects models for the basic- and superordinate-level target groups.

As shown in Figure 1, for basic-level targets, introducing anchors replicated both unpacking effects: PJs were higher in the typical unpacked condition ($Est = 6.88; t = 3.39; p = .002$) and lower in the atypical unpacked condition ($Est = -7.90; t = -3.55; p = .002$) compared to the packed baseline. For superordinate-level targets, by contrast, PJs in the typical unpacked condition did not significantly differ from the packed baseline ($Est = 1.36; t = 0.57; p = .576$), whereas atypical unpacking reliably resulted in lower PJs ($Est = -6.87; t = -2.46; p = .035$).

The inconsistency suggests that the typical unpacking effect is sensitive to the hierarchical level at which unpacking occurs. Specifically, the effect emerged when unpacking basic-level targets to subordinate-level anchors, whereas it failed to be observed when unpacking superordinate-level targets to basic-level anchors.

Experiment 2: Self-paced Reading Task

Participants 60 participants were recruited on Prolific. Participants were paid at a rate of £6 per hour for their participation. All participants were native British English speakers and provided explicit consent before participating.

Materials The 120 sets of probability questions (60 experimental + 60 filler) were adapted into sentences by removing the probability question frame, and a spillover region of at least five words was added following the target word to capture spillover effects.

Procedure Participants read sentences word by word by pressing the space bar to move forward, and intervals between key presses were recorded as reading times. The same one-third of items were followed by a comprehension question.

Table 1: Example item in the probability judgment task (Question) and self-paced reading task (Sentence). Anchors are underlined and target words are in bold.

Condition	Question (Experiment 1)	Sentence (Experiment 2)
Basic-level Target		
Packed	*What is the probability that the timid boy would scream loudly when he saw a **snake**?*	The timid boy would scream loudly when he saw a **snake** because he was terrified of reptiles.
Typical Unpacked	*What is the probability that the timid boy would scream loudly when he saw a cobra or any other **snake**?*	The timid boy would scream loudly when he saw a cobra or any other **snake** because he was terrified of reptiles.
Atypical Unpacked	*What is the probability that the timid boy would scream loudly when he saw a cottonmouth or any other **snake**?*	The timid boy would scream loudly when he saw a cottonmouth or any other **snake** because he was terrified of reptiles.
Superordinate-level Target		
Packed	*What is the probability that Joey wanted to learn to play the **instrument**?*	Joey wanted to learn to play the **instrument** before he went away to college.
Typical Unpacked	*What is the probability that Joey wanted to learn to play the piano or any other **instrument**?*	Joey wanted to learn to play the piano or any other **instrument** before he went away to college.
Atypical Unpacked	*What is the probability that Joey wanted to learn to play the banjo or any other **instrument**?*	Joey wanted to learn to play the banjo or any other **instrument** before he went away to college.

Table 2: Linear mixed-effects model results for all PJs

	Est	SE	t	Pr(>|t|)
(Intercept)	59.49	2.80	21.27	$< 2e^{-16}$
Typical Unpacked	4.12	1.58	2.61	0.013
Atypical Unpacked	-7.38	1.87	-3.95	<0.001
Category Level	0.97	3.91	0.25	0.805
Typical:CatLevel	-5.54	3.21	-1.73	0.092
Atypical:CatLevel	1.04	3.41	0.31	0.761

Table 3: Linear mixed-effects model results for all RTs

	Est	SE	t	Pr(>|t|)
(Intercept)	344.54	12.02	28.66	$< 2e^{-16}$
Typical Unpacked	8.22	9.78	0.84	0.404
Atypical Unpacked	27.89	11.94	2.34	0.023
Category Level	14.76	6.10	2.42	0.019
Typical:CatLevel	-12.16	12.80	-0.95	0.346
Atypical:CatLevel	-5.38	16.49	-0.33	0.746

Analysis As indicated in Figure 1, reading time (RT) data on target words (TW) showed no significant difference across conditions, potentially due to spillover effects. We therefore analyzed RTs at the first post-target word (TW+1) using linear mixed-effects models.

- *RT ~ Typicality * CatLevel + (1 + Typicality * CatLevel | Participant) + (1 + Typicality | Item)*

Results Table 3 reports the linear mixed-effects model results for all items, and we also conducted separate linear mixed-effects models for basic- and superordinate-level targets.

For both basic- and superordinate-level targets, RTs in the typical unpacked condition did not differ from the packed baseline (Basic: $Est = 8.15; t = 0.83; p = .413$; Superordinate: $Est = -3.87; t = -0.45; p = .653$). However, the atypical unpacked condition showed significantly slower RTs than the packed baseline for basic-level targets ($Est = 27.94; t = 2.29; p = .027$), and showed a trend toward slower RTs for superordinate-level targets ($Est = 22.33; t = 2.00; p = .053$).

Cross-Task Correlations

To relate the two tasks, we conducted linear regression analyses to examine item-level relationships between ΔPJ and ΔRT when introducing different types of anchors.

As shown in Figure 2, for both types of targets, no correlation was observed when introducing typical anchors (Basic: $\beta = -0.10, p = .858$; Superordiante: $\beta = -0.06, p = .910$), but significant correlations emerged when introducing atypical anchors (Basic: $\beta = -1.24, p = .039$; Superordiante: $\beta = -1.41, p = .012$). In summary, changes in PJs and RTs were related only in the presence of atypical unpacking.

Discussion

Online word prediction and offline probability judgment have both been characterized as relying on probability, raising the possibility that the two tasks share a common foundation of probabilistic processes. This study tested this shared processes hypothesis by examining how each task responds to the unpacking manipulation. Specifically, we asked whether introducing typical or atypical anchors produces probability shifts in both tasks, and whether the shifts they generate are related. To address these questions, our experiments investi-

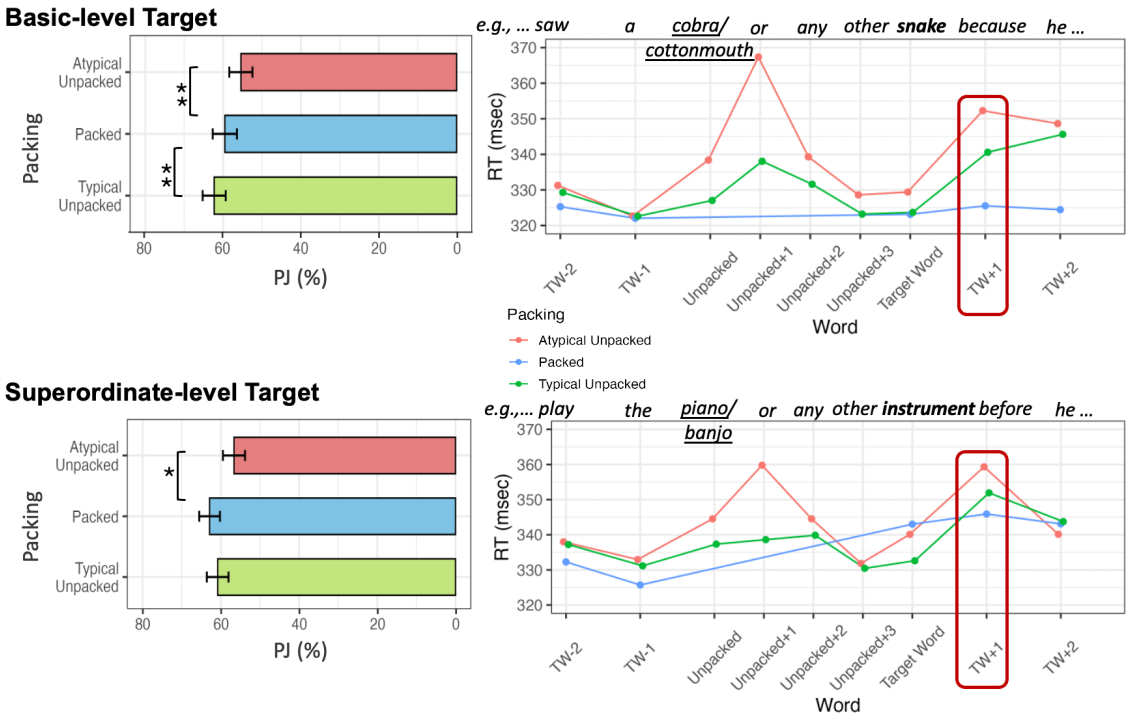

Figure 1: Average offline probability judgments (PJs) and average online reading times (RTs) by word.

gated unpacking effects **within each task**, and the relationship between these effects **across tasks**.

In the probability judgment task, atypical anchors lowered probability judgments, replicating unpacking effects and extending them from the prototypical examples and small item sets typically used in judgment research to a wider set of sentences more commonly used in sentence processing research. In self-paced reading, atypical anchors slowed target spillover reading times, suggesting that the same probability shifts also arise in real-time sentence processing, with the caveat that these reading times also reflect other systematic effects of the unpacking manipulation, including the presence of an additional conjoined noun phrase. These within-task results show that introducing atypical anchors produces probability shifts in both offline probability judgment and online word prediction. More importantly, cross-task analyses reveal that item-level changes in probability judgments and reading times were negatively correlated when atypical anchors were introduced, for both basic and superordinate categories. Together, these convergent patterns with atypical anchors provides evidence for a shared probabilistic foundation underlying probability judgment and word prediction.

The remaining question is why typical anchors failed to influence reading times despite producing measurable shifts in probability judgments for basic-level targets. We suggest that this asymmetry reflects properties of our materials and of the reading task itself, rather than a failure of the hypothesis.

One reason for the absence of typical unpacking effects in reading times may lie in the materials themselves. In classic unpacking manipulations, typical anchors are often more available than the target category. For example, in the question "*I see a table. What is the probability that I also see a chair, computer, curtain, or any other object starting with a C?*", typical anchors *chair, computer, and curtain* are basic-level lexical items that more easily come to mind than the ad hoc target *objects starting with a C* (Dasgupta et al., 2017). In the present experiments, however, the target word was a moderate-cloze continuation of the context, whereas anchors were not cloze responses, but category members of the target. As a result, typical anchors were unlikely to be more probable continuations of the context than target words. This interpretation is supported by item measures of GPT-2 surprisal. Typical anchors ($M = 10.52, SD = 4.07$) have significantly higher surprisal values than target words ($M = 8.10, SD = 3.54$), $t = -3.49, p < 0.001$. Consequently, typical anchors may not have been informative enough to facilitate online sentence processing. If the contexts were modified such that typical anchors were more likely to occur next than target words, typical unpacking effects might instead emerge.

The nature of self-paced reading task itself may also have obscured typical unpacking effects. First, probability judgments scale linearly with probability, but reading times are proportional to surprisal, which is the negative logarithm of probability. The log transformation may compress probability shifts: a shift large enough to be reliably detected as a change in probability judgment may translate into a much smaller change in surprisal, and therefore only a small change in reading time. Second, probability judgment allows unlimited deliberation before response, whereas reading is incremental and proceeds under natural time pressure, which has

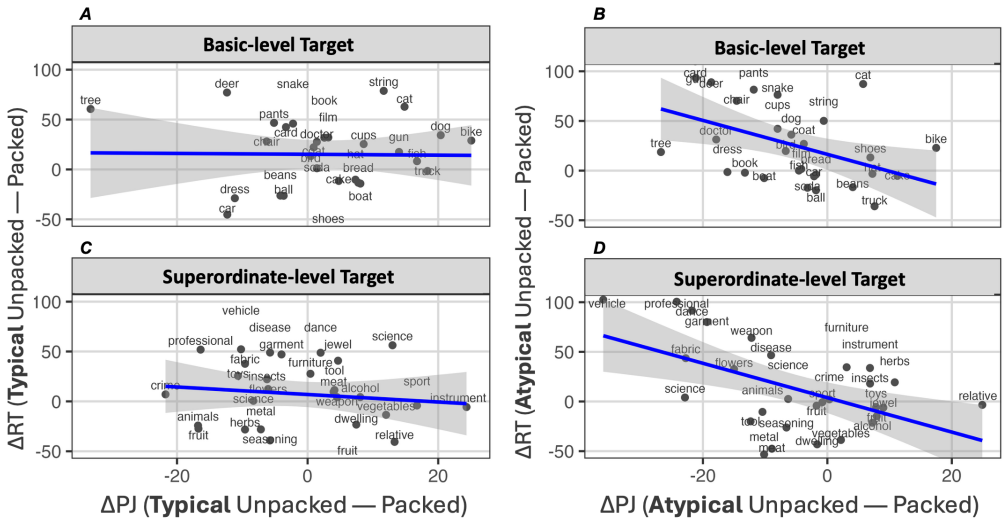

Figure 2: Item-level interaction between ΔPJ and TW+1 ΔRT. In each panel, the x-axis shows item-level changes in PJs, and the y-axis shows item-level changes in RTs. Panels A and B show unpacking basic-level targets to typical (A) and atypical (B) anchors, while Panels C and D show unpacking superordinate-level targets to typical (C) and atypical (D) anchors. Each point represents a single item, labeled by its target word. Solid lines indicate linear regression fits, with shaded areas representing 95% confidence intervals.

been argued to modulate unpacking effects (Dasgupta et al., 2017). This time pressure may particularly disadvantage typical anchors, because facilitation requires active integration of information, whereas inhibition arises more automatically from disrupted expectations. As a result, participants may be less able to respond to typical anchors during self-paced reading. These factors suggest that the absence of typical unpacking effects in reading times need not indicate that the underlying probabilistic processes are unaffected by typical anchors, only that the resulting shift may be hard to detect in reading. Future studies using more naturalistic eye-tracking measures will further investigate the time course of unpacking in sentence processing.

The shared probabilistic processes have important implications for both psycholinguistics and judgments and decision-making. From the **psycholinguistic** side, the mechanisms that account for probability judgment can be used to understand word prediction. Algorithmic-level accounts nowadays propose that humans are constrained by cognitive resources and therefore do not compute exact probabilities. Instead, people approximate probabilities by drawing samples from memory or from mental simulations (Sanborn & Chater, 2016; Zhu et al., 2020). These Bayesian mechanisms in probability judgment are consistent with the arising hypotheses of Bayesian brain in language processing, which argues that comprehenders rely on probabilistic representation and inference to process language (Chater & Manning, 2006; Smith & Levy, 2013). As human language processing also faces computational resource limitations, the proposed sampling framework in probability judgment may open a new avenue to understand how humans effectively process language under such constraints (Hoover et al., 2023). From the **judgment and decision-making** side, language processing theories, given their sensitivity to time, may offer insights into how judgment mechanisms integrate response times. For example, Autocorrelated Bayesian Sampler (Zhu et al., 2023) in judgment literature extends the original Bayesian Sampler (Zhu et al., 2020) to jointly model choices and response times by introducing a more realistic stopping rule for the sampling process. Whereas the original Bayesian Sampler assumes that sampling terminates after a fixed number of samples, which implicitly implies fixed response times, the Autocorrelated Bayesian Sampler instead asserts that individuals track the difference in sample counts between competing options and cease sampling once an option surpasses a decision threshold. This stopping condition closely parallels the race model in word prediction, in which candidate words compete towards a threshold level of activation, and the first candidate that reaches the threshold is produced by individuals (Staub et al., 2015). The convergence of these architectures suggests that mechanisms in judgment may capture not only the time course of explicit probability judgments, but also the time course of implicit probabilistic processes such as word prediction during reading.

Conclusion

This work takes the first step of relating online word prediction and offline probability judgment, highlighting the shared foundation of probabilistic reasoning on which these processes depend. As probabilistic approaches increasingly shape theories of cognition across domains, bridging these fields presents important and exciting challenges for future research. Understanding how various cognitive processes emerge from shared probabilistic principles may therefore be key to explaining both their commonalities and their differences.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*(1116). https://doi.org/10.3389/fpsyg.2016.01116

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344. https://doi.org/10.1016/j.tics.2006.05.006

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. N. (2020). Probabilistic biases meet the bayesian brain. *Current Directions in Psychological Science*, *29*(5), 506–512. https://doi.org/10.1177/0963721420954801

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480. https://doi.org/10.1037/a0037010

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25. https://doi.org/10.1016/j.cogpsych.2017.05.001

Hadjichristidis, C., Sloman, S., & Wisniewski, E. (2001). A probabilistic earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 1–8.

Hale, J. (2001). Judging the probability of representative and unrepresentative unpackings. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 1–5.

Hoover, J. L., Sonderegger, M., Piantadosi, S. T., & O'Donnell, T. J. (2023). The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, *7*, 1–42. https://doi.org/10.1162/opmi_a_00086

Kahneman, D. (2011). *Thinking, fast and slow*. NY: Macmillan.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*, *5*(1), 107–135. https://doi.org/10.1162/nol_a_00105

Peelle, J. E., Miller, R. L., Rogers, C. S., Spehar, B., Sommers, M. S., & Van Engen, K. J. (2020). Completion norms for 3085 english sentence contexts. *Behavior Research Methods*, *52*(4), 1795–1799. https://doi.org/10.3758/s13428-020-01351-1

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-x

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Shain, C. (2024). Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, *8*, 177–201. https://doi.org/10.1162/opmi_a_00119

Shain, C., Blank, I. A., Van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). Fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 573–582. https://doi.org/10.1037/0278-7393.30.3.573

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17. https://doi.org/10.1016/j.jml.2015.02.004

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567. https://doi.org/10.1037/0033-295X.101.4.547

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2003). Category norms: An updated and expanded version of the battig and montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719–748. https://doi.org/10.1037/rev0000190

Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2023). The autocorrelated bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychological Review*, *131*(2), 456–493. https://doi.org/10.1037/rev0000427