Syntactically-guided Information Maintenance in Sentence Comprehension

Shinnosuke Isono
NINJAL
s-isono@ninjal.ac.jp

Kohei Kajikawa
Georgetown University
kk1571@georgetown.edu

Abstract

Maintaining information in context is essential in successful real-time language comprehension, but maintenance is cognitively costly and can slow processing. We hypothesize that rational language users selectively maintain information that is crucial for future prediction, guided by syntactic structure. Under this view, two factors affect maintenance cost: the number of predicted heads and the number of incomplete dependencies. Although these factors have been treated as competing hypotheses in the literature, our account predicts that they are not reducible to one another. We show this is the case in a naturalistic reading time dataset in Japanese, a language in which the two factors contrast particularly clearly. We further show that there is a tradeoff such that readers that slow down for maintenance tend to benefit more from predictability, providing additional support for the proposed account. These patterns are not evident in English, however, and we highlight some issues to be resolved to understand the contribution of syntax in memory-efficient processing of various languages.

1 Introduction

"Language is fleeting," as Christiansen and Chater (2016) put it. Language unfolds over time, and comprehension takes place incrementally (Tanenhaus et al., 1995), but a human language user cannot keep track of everything that is said or written. Yet words are related to each other, often at a distance, and the language user needs to work out their relations to fully understand the meaning of the expression. From a different viewpoint, the language user should seek to minimize the error in prediction using the past input as the context (Hale, 2001; Levy, 2008). How do humans manage to do that with their limited memory capacity?

We hypothesize that a rational language user saves memory resources by selectively maintaining words that are critical for predicting future inputs (Hahn et al., 2022; Xu and Futrell, 2026), and syntactic structure provides a reliable clue for such maintenance by limiting words that can be directly related to upcoming words (Figure 1).

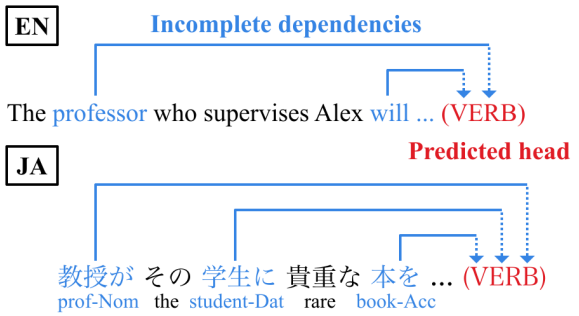

Figure 1: Incomplete dependencies and predicted heads in example sentence fragments in English (a verb-medial language) and Japanese (a verb-final language).

This view revives an old question about **maintenance (storage) cost**. Maintaining linguistic information should be cognitively costly (Yngve, 1960; Miller and Chomsky, 1963; Kimball, 1973; Gibson, 1991; Stabler, 1994; Lewis, 1996; Gibson, 1998, 2000), and can be observed as reading slowdown. In terms of dependency structure, the number of **predicted heads**[1] is known to be a major determinant of maintenance cost crosslinguistically (Chen et al., 2005; Nakatani and Gibson, 2010; Ristic et al., 2022; Isono et al., 2025). In the current account, this is because heads being predicted means some bits of information that are valuable for future prediction are carried.[2] Importantly, the current account also predicts that additional **incomplete dependencies** to be completed at the same predicted head further increase the maintenance cost (cf. Gib-

[1]Here, *heads* refers to atoms of dependency analysis. We do not mean to contrast heads with dependents. If one element triggers prediction of another element, that counts as a predicted head, regardless of the direction of the dependency.

[2]See also Kajikawa et al. (2026) for an information-theoretic formulation.

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 57–69

son, 1991), since they carry additional information. Previous studies, in contrast, treat predicted heads and incomplete dependencies as competing explanations for the maintenance cost, and empirical evidence has been taken to favor predicted heads (Gibson, 1998; Nakatani and Gibson, 2008).

We reinvestigate the contribution of predicted heads and incomplete dependencies to the maintenance cost during reading, using a large-scale naturalistic reading time dataset in Japanese. Japanese is suitable for the current purpose because these two factors contrast most clearly in a strictly head-final language. The use of large-scale dataset and the statistical control for predictability may reveal the maintenance cost of incomplete dependencies which could have been masked in the previous study by being too small or because of a confounding factor (see Section 6.2). We indeed find that both predicted heads and incomplete dependencies have a slowdown effect that is not reducible to the other, with the former being larger, as predicted by the current account.

There is another "rational" strategy that a reader can take. When there are too many pieces of information, one can give up maintaining everything and hasten to get out of the demanding structure (cf. Ferreira et al., 2002). The drawback is that the future input is less predictable since the context is forgotten. We show that such a tradeoff exists, adding further support for the view that maintenance is for better prediction.

While a strictly head-final language like Japanese is the most suitable to address the current question, it is important to see whether the pattern generalizes to typologically different languages, as reviewers pointed out. We additionally conducted the same analysis with an English dataset. The number of predictive heads turns out to be a significant predictor, as in Japanese, when content words are analyzed. Beyond that, however, we do not find evidence for the patterns observed in Japanese. There are some issues in implementing a syntactically-guided memory-efficient predictive processing in typologically diverse languages, and we discuss them in Limitations.

The rest of this paper is structured as follows. Section 2 introduces the background and describes the current hypothesis in more detail. Section 3 reports the experiment evaluating the two metrics of maintenance cost in Japanese. Section 4 reports the follow-up experiment investigating the tradeoff between maintenance and prediction. Section 5 reports the experiment in English. Section 6 discusses the implications of the results, related work, and limitations. The code is available online.[3]

2 Background

2.1 Syntax as a clue for rational maintenance

A key observation that motivates the current study is that a subset of words in the context is highly informative in predicting the upcoming words. Consider predicting the next word following (1).

(1) The professor who supervises Alex will

Will suggests that the next word is likely to be an infinitival verb like *be* and *teach*, rather than a finite verb, noun, adjective, etc. *Professor* suggests that the next word is likely to be an action that a professor often does, such as *teach* or *publish*, and is less likely to be actions like *cook* and *roar*. In contrast, other words are less informative: *the, who, supervises,* and *Alex* tell relatively little about what the next word is likely to be. The syntactic structure explains why. An analysis based on the Universal Dependency–style dependency grammar (de Marneffe et al., 2021) of (1) is shown in Figure 1. *Professor* and *will* are expected to depend on the same upcoming verb. Other words in the fragment are unlikely to form a dependency with upcoming words, as they have already found their head and there is a general ban on crossing dependencies. More generally, syntactic dependency structure tells which words in context are highly informative for future prediction. In fact, this is the source of the anti-locality effect documented in many languages (Konieczny, 2000; Konieczny and Döring, 2003; Vasishth and Lewis, 2006; Asahara et al., 2019; Isono, 2024), where preceding dependents facilitates the processing of the head. Though other words can be informative through semantics and world knowledge in some contexts, the contribution of syntactically related elements is more general and stable.

This observation directly suggests a rational strategy that a memory-limited language processor that seeks to minimize prediction error would employ: maintaining *professor* and *will* while neglecting others. More generally, we hypothesize that the human language processor rationally allocates memory resources to words that are particularly informative for future prediction (cf. Hahn

[3] https://osf.io/cvket/overview?view_only=0b8d2db223274da981f7ceea1d39b5f5

et al., 2022; Xu and Futrell, 2026), and syntax helps identify which words are. Words that are not maintained will be eventually forgotten by decay and/or interference (Lewis and Vasishth, 2005), unless there are other reasons to maintain them.

2.2 Syntactic metrics of maintenance cost

The syntax-based approach to resource-rational processing is also interesting since it brings a new perspective to an old question about the maintenance cost during processing. As mentioned in the Introduction, there are two possible ways to quantify maintenance cost under a dependency-style syntactic analysis: the number of predicted heads and the number of incomplete dependencies (Gibson, 1998). While these two are correlated since a predicted head implies an incomplete dependency, they often diverge in head-final structures.

Japanese, for example, is a strictly head-final language. As a result, multiple dependents often precede the clause-heading verb. It has been shown that speakers can predict the clause structure before the head verb using case markers on those dependents (Kamide and Mitchell, 1999; Miyamoto, 2002). In (2), the markers suggest that the three nouns can all depend on a single ditransitive verb.

(2) Kyoju-ga gakusei-ni hon-o / miseta.
prof-NOM student-DAT book-ACC showed
'The prof. showed the student the book.'

At the point before the verb, indicated by the slash, **one** head is predicted (the ditransitive verb), while **three** dependencies remain incomplete. Compare (2) with (3), where all the three nouns are marked as nominative:

(3) #Kyoju-ga gakusei-ga hon-ga /
Prof-NOM student-NOM book-NOM
omosiroi-to omotteiru-to itta.
interesting-that think-that said
'The professor said that the student thought that the book was fun.'

At the slash, **three** heads are predicted and **three** dependencies remain incomplete. Gibson (1998) takes the apparent difficulty of (3) compared to (2) as evidence that the number of predicted heads is a better metric of maintenance cost than the number of incomplete dependencies. Controlled experiments in English (Chen et al., 2005) and Japanese (Nakatani and Gibson, 2008) corroborate this view (we revisit the Japanese experiment in Section 6.2). The effect of predicted heads is also observed in a Japanese naturalistic reading time dataset (Isono et al., 2025), although that study does not compare predicted heads with incomplete dependencies.

Unlike previous work, the current resource-rational view sees predicted heads and incomplete dependencies as complementary, rather than competing, explanations of maintenance cost. For example, the dative-marked *gakusei-ni* in (2) and the nominative-marked *gakusei-ga* in (3) are both informative for future prediction, just differently. The former narrows down the type of the already predicted verb, while the latter implies another verb. The rational processor should maintain both types of information. Accordingly, both should affect the maintenance cost. The relative difficulty of (3) suggests that the latter type of information (predicted heads) is "heavier" than the former (incomplete dependencies).

3 Experiment 1

Our hypothesis predicts that the maintenance cost is affected by both the number of predicted heads and the number of incomplete dependencies, and the former has a greater impact on reading times. We test these predictions using a large-scale reading time dataset in Japanese.

3.1 Materials

We use the BCCWJ-SPR2 dataset (Asahara, 2022). It is currently the largest dataset of naturalistic reading times in Japanese to our knowledge. It is based on the BCCWJ corpus (Maekawa et al., 2014) and contains self-paced reading times (Just et al., 1982) from native speakers. Texts were presented one *bunsetsu* at a time, where bunsetsu is a word-like unit commonly used in Japanese linguistics. A bunsetsu often contains multiple morphemes, e.g., a noun and a case marker, or a verb and a series of auxiliaries. For readability, bunsetsu will be referred to as region in the rest of this paper. We use the PB subset, which is a collection of excerpts from various books, since Universal Dependencies (UD) annotation is available for this subset.

For the current analysis, we take the mean reading time for each region in the dataset, aggregating the actual reading time from individual participants, and use it as the dependent variable (e.g., Goodkind and Bicknell, 2018; Wilcox et al., 2023). As in many reading studies using naturalistic data, the sentence-final regions are excluded from analysis

since they often reflect wrap-up effects. The first two regions of each sentence are also excluded since they can be affected by spillover from the final region of the previous sentence. After the exclusion, the data consists of 83 documents, 8,050 sentences, or 50,285 regions, coming from 424 participants.

This dataset is already shown to exhibit the effect of maintenance cost quantified as the number of predicted heads (Isono et al., 2025). Our focus is therefore on the effect of incomplete dependencies and its relation to the effect of predicted heads.

3.2 Variables

Maintenance cost The variables of interest for the current study are the number of predicted heads and the number of incomplete dependencies at each region. They are computed from the UD–style syntactic annotation of the target dataset provided by Asahara et al. (2018). We only consider these gold parses, while in reality the incremental parser may also pursue different parses (the "perfect oracle" assumption in the sense of Brennan (2016)). The number of predicted heads is operationalized as the number of words that have a dependency with a word in the context but are yet to be seen. The number of incomplete dependencies is the number of dependencies that start with a word in the context and end with a word that is yet to be seen. These values are initially calculated at the morpheme level, following the UD annotation, and then aggregated at the region level by taking the minimal value. Taking the minimum virtually means that we only count dependencies that neither begin nor end in the region in question. This is necessary because the amount of region-internal dependencies would be strongly affected by the length of the region, and by the idiosyncrasy of UD annotation which attaches all the verbal suffixes to the stem.

Control variables The baseline model contains the following five control variables: the position of the sentence in the document, the position of the region in the sentence, the number of characters in the region, the sum of the unigram surprisal of the morphemes comprising the region, and the sum of the GPT-2 surprisal of the morphemes. We take surprisal (Hale, 2001; Levy, 2008) into account not only because it is known to strongly affect reading times (Wilcox et al., 2023; Xu et al., 2023), but because the confounding effect of expectation may be one reason that the cost from incomplete dependencies were not observed previously (see Section 6.2). The unigram surprisal is calculated using the frequency table of the NINJAL Web Japanese Corpus (Asahara et al., 2014). We obtain by-morpheme surprisal values and then summed them at the region level rather than directly assessing the by-region surprisal because the exact combination of morphemes can be hard to find in a corpus. GPT-2 surprisal, based on the GPT-2 large language model (Radford et al., 2019), is calculated using the `japanese-gpt2-xsmall` model (Sawada et al., 2024).[4] We use the `xsmall` model since larger models are known to fit worse to reading times both generally (Oh and Schuler, 2023) and with the current dataset (Isono et al., 2025).

Spillover In the self-paced reading paradigm, the processing load of one region often affects reading times of the subsequent regions. This is called the spillover effect (Mitchell, 1984). Spillover effects are taken into account in the current analysis by including variables from $(i-1)$th and $(i-2)$th regions in the regression model for the ith region. We do this for the number of characters, unigram surprisal, and GPT-2 surprisal. Spillover of the position effects is not considered because the position of the previous words is perfectly predictable from the current position. Spillover of the maintenance cost is not considered either, because the maintenance cost of adjacent regions should correlate strongly, and because when they differ, it is because some dependencies have initiated or terminated there, but we do not want effects of dependency initiation or termination to be mixed with the maintenance cost.

Correlations Correlations between the variables are shown in Figure 2. The two metrics of maintenance cost are highly correlated. This is because a predicted head by definition implies an incomplete dependency. This correlation can distort the estimate of the coefficients. We therefore use the number of *additional* incomplete dependencies, which is simply the number of incomplete dependencies minus the number of predicted heads, when estimating the coefficients (Section 3.4).

3.3 Analysis

Psychometric predictive power Following studies on naturalistic reading time datasets (Goodkind

[4] `https://huggingface.co/rinna/japanese-gpt2-xsmall`. The perplexity of the model for the current dataset is 61.1.

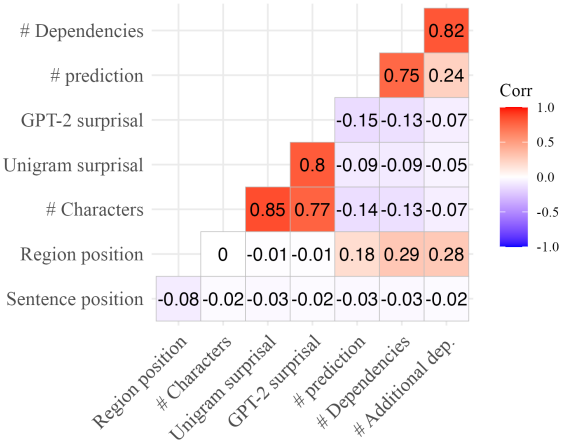

Figure 2: Correlations between variables used in the regression analysis. # = number of.

	ΔMSE
Sentence position	176.3
Region position	6.6
# Characters	3.4
Unigram surprisal	331.9
GPT-2 surprisal	2.4
# Heads	6.9
# Dependencies	5.9
Both	7.4

Table 1: ΔMSE of predictors.

and Bicknell, 2018; Wilcox et al., 2023, among others), we evaluate the psychometric predictive power of a maintenance cost variable by comparing linear regression models with and without it. Both models contain the control variables defined above. If the inclusion of a variable improves the model fit, that variable is considered as having psychometric predictive power. Specifically, we measure the decrease in the mean squared error (ΔMSE) of the linear regression model by the inclusion of the maintenance cost variable. The error for each region is calculated by 10-fold cross-validation repeated 50 times, and a permutation test is conducted to test whether ΔMSE is significantly above zero.

Coefficients We also check the regression coefficients assigned to the maintenance cost variables. If the psychometric predictive power of these variables reflects maintenance cost, the coefficients should be positive (i.e., slowdown effects). We also predict the coefficient for predicted heads to be larger (recall the comparison between (2) and (3)). For this analysis, we use the number of additional incomplete dependencies instead of the number of all incomplete dependencies. To obtain a reliable estimate of the effect, the data is randomly split into 10 folds and linear regression models are fit to each of them. This gives a distribution of estimated coefficients for each variable. We conduct permutation tests on these distributions to address the above questions.

Exclusion of potential hidden dependencies One potential concern for the current approach is that there are hidden dependencies that are not available in the UD annotation. Certain expressions in Japanese trigger prediction for a particular type of auxiliary or particle to be attached to the verb, but such relations are not expressed in the UD annotation. In the following example, *-sika* 'only' requires a negative auxiliary to be attached to the verb. In the UD annotation, however, *-sika* attaches to the noun, and the noun depends on the head verb as usual; the dependency between *-sika* and negation is not indicated.

(4) Kyoju-ga hon-**sika** kawa-**nakat**-ta.
Prof-NOM book-only buy-not-ed
'The professor only bought a book.'

There are many other negation-triggering expressions in Japanese. A similar problem arises with the UD analysis of a *wh*-expression (e.g., *dare* 'who') and an obligatory question particle attached to the verb (usually *ka*). Psycholinguistic evidence shows online processing is sensitive to such correspondences (Ono and Nakatani, 2014; Nakatani, 2023). Such correspondences can result in a superficial effect of incomplete dependencies above and beyond predicted heads, because expressions like *hon-sika* 'book-only' in (4) increase the number of incomplete dependencies but not the number of predicted heads. For this reason, we also run an analysis using only sentences without any negation or question particles.[5] This excluded 1,908 sentences (23.8%).

3.4 Results

Psychometric predictive power We first compare the number of predicted heads and the number of incomplete dependencies by their psychometric predictive power. The results are summarized in Figure 3 (the orange bars). Both the number of predicted heads and the number of incomplete dependencies show a significantly positive ΔMSE,

[5] Particles are identified using the lemma column of the UD annotation.

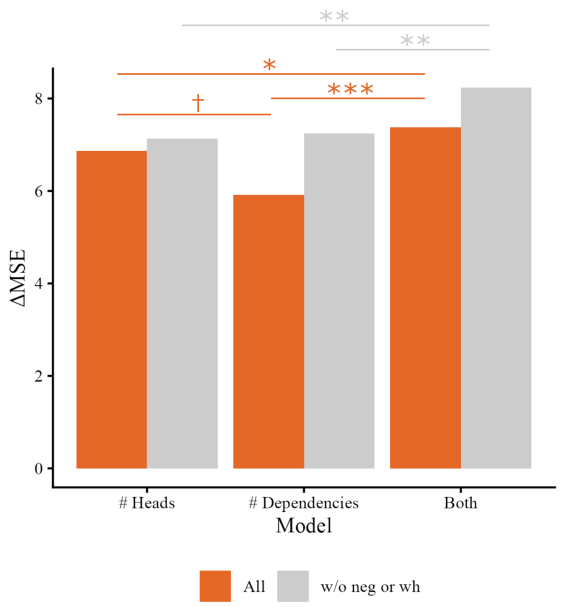

Figure 3: ΔMSE per word by model. The stars indicate the significance of permutation tests: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, † for $p < 0.1$

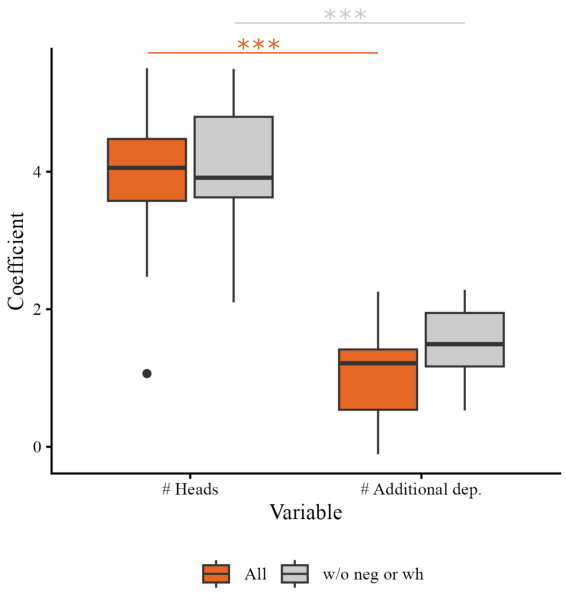

Figure 4: Distribution of estimated coefficients of the metrics of maintenance cost. The stars indicate the significance of permutation tests: *** for $p < 0.001$

meaning that both are predictive of human reading times. Predicted heads have a higher ΔMSE than incomplete dependencies, though the difference is only marginally significant ($p = 0.058$). The model that contains both predicted heads and incomplete dependencies performs significantly better than the models that contain only one of them ($p = 0.023$ when compared against predicted heads, and $p < 0.001$ when compared against incomplete dependencies). These results indicate that both predicted heads and incomplete dependencies have psychometric predictive power that cannot be reduced to that of the other.

The magnitude of the psychometric predictive power of the maintenance cost predictors is compared with baseline predictors in Table 1. The ΔMSE values for baseline predictors are calculated by comparing baseline models with and without that predictor. While the power is much smaller than those of some low-level predictors, it is larger than GPT-2 surprisal, which is supposed to capture high-level linguistic structure.

Coefficients The distribution of estimated coefficients for the maintenance cost variables are summarized in Figure 4. As can be seen, both variables show a consistent positive effect across the folds. The coefficient for the predicted heads is significantly larger than the coefficient for the incomplete dependencies ($p < 0.001$). These results are consistent with the view that both predicted heads and incomplete dependencies impose maintenance cost, but the former is heavier.

Exclusion of potential hidden dependencies The above analyses are repeated with the subset of the original data that do not contain any negation or question particle. The results after exclusion are shown as gray bars in Figures 3 and 4. The exclusion does not change the key results: both of the two maintenance cost variables have psychometric predictive power, which cannot be reduced to that of the other, and the coefficient is larger for the number of predicted heads.

4 Experiment 2

So far we have analyzed the general tendency among readers. However, different readers can take different strategies of maintenance. Recall that a high maintenance cost is symptomatic of a center-embedding structure, like (3). Given the difficulty of such structures, some readers may give up mid-sentence maintaining the required information; then the maintenance cost would disappear. Some may even speed up in order to get out of the center-embedding structure, to at least comprehend the message carried by the matrix structure. These can be seen as instances of good-enough processing (Ferreira et al., 2002). By giving up maintenance, readers are relieved from memory load, but they

are sacrificing accurate parsing and prediction.

Below, we show such a tradeoff exists. First, we show that participants respond differently to the need of maintenance: many slow down but some speed up. Second, we show that this difference in maintenance strategy affects how well they predict.

4.1 Materials

We use the same BCCWJ-SPR2 dataset as in Experiment 1. Unlike in Experiment 1, we do not collapse data across participants. Instead, we model raw reading times from participants with 1,000 or more data points. 353 participants are included in this analysis, with 6,370,767 data points in total.

4.2 Variables

Along with the variables used in Experiment 1, we use the number of dependencies completed at each region.[6] This value should predict the anti-locality effect (Konieczny, 2000; Konieczny and Döring, 2003; Vasishth and Lewis, 2006), i.e., the processing facilitation due to preceding elements that are related to the current input. In fact, previous studies on Japanese naturalistic texts show robust facilitatory effects of dependency completion (Asahara et al., 2019; Isono et al., 2025). We predict that this facilitatory effect is weaker for participants who speed up in the face of a high maintenance cost, since these participants are sacrificing prediction.

4.3 Analysis

We first identify whether each participant slows down or speeds up when faced with structures that impose a high maintenance cost. We model each participant's reading time separately using two regression models, one containing the number of predicted heads in addition to control variables, and another containing the number of incomplete dependencies in addition to control variables.[7] Since we are interested in the sign of the coefficients, we split data from each participant into 10 folds, fit models for each fold separately, and conduct permutation tests testing whether the mean of the coefficients is significantly above or below zero. This gives a typology of participants: for each of the two maintenance cost variables, each participant faced with a high maintenance cost slows down significantly, speeds up significantly, or shows no significant evidence of either.[8]

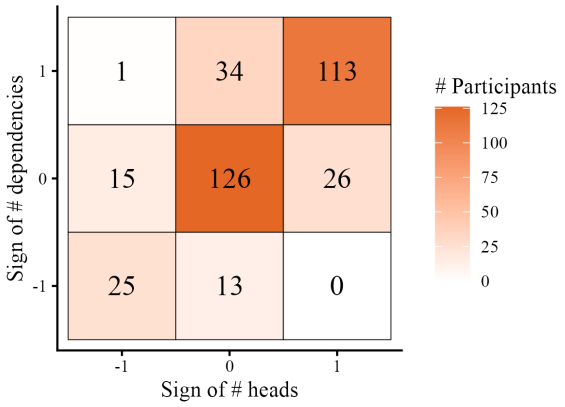

Figure 5: Classification of participants by their response to high maintenance cost. 1 indicates significant slowdown, -1 indicates significant speedup, and 0 indicates significance in neither direction.

We then investigate whether this typology affects prediction. An estimate of the anti-locality effect for each participant is obtained by fitting another model that contains all the target and control variables plus the number of dependency completions. We then fit a linear model that predicts this anti-locality effect by each of the typology for the two maintenance cost variables.

4.4 Results

The typology of participants is summarized in Figure 5. When the number of predicted heads is high, 139 participants (39.4%) show a significant slowdown while 41 (11.6%) show a significant speedup. The trend is similar when analyzed in terms of the number of incomplete dependencies.

The relation between this typology and the anti-locality effect is summarized in Figure 6. As expected, participants who slow down in the face of a high maintenance cost show a significantly stronger anti-locality effect, both when the typology is in terms of predicted heads ($p = 0.001$) and incomplete dependencies ($p = 0.049$).

5 Experiment 3

So far we have analyzed a Japanese dataset, but a natural question to ask is whether the observed

[6] We count dependencies completed at each UD token and then take the maximal value for each region. Thus we avoid counting most completions of region-internal dependencies.

[7] We fit two separate models for each type of maintenance cost since putting them together in one model can destabilize the coefficient estimates by the high correlation between them.

[8] We use a categorial typology rather than coefficients themselves since the absolute value of the coefficients can reflect the participant's general sensitivity to linguistic features.

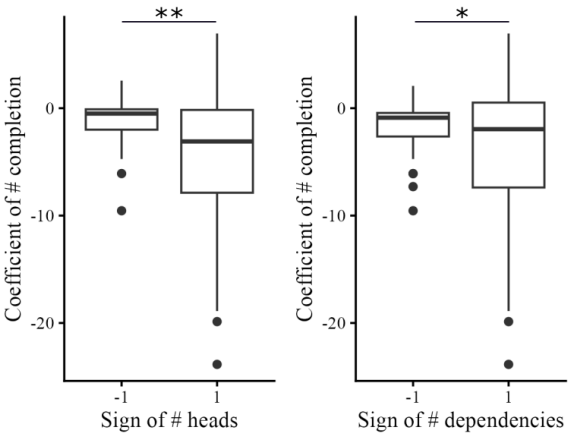

Figure 6: Distribution of anti-locality effect (y-axis) by participants' maintenance strategy (x-axis). The stars indicate the significance of the permutation test: ** for $p < 0.01$, * for $p < 0.05$

patterns generalize to typologically different languages, as reviewers pointed out. We therefore additionally analyzed the Natural Stories dataset (Futrell et al., 2021), which contains self-paced reading time by 181 native speakers of English reading 10 stories (10,245 words). The methods are the same as Experiments 1 and 2 unless otherwise specified. The figures are in the Appendix.

5.1 Variables

Maintenance cost English often uses adjuncts that adjoin from the right, unlike in Japanese. We do not count those right adjuncts as predicted or forming an incomplete dependency.[9] This is because right adjuncts are generally unpredictable. For example, in *the professor read the book carefully*, the adverbial *carefully* is optional. It is not sensible to assume that *read* triggers different number of predictions depending on whether an optional adjunct follows in the actual continuation.

Control variables The following control variables are used: the position of the region in the story and in the sentence, the number of characters in the region, the negative log frequency of the word, and GPT-2 surprisal of the word.

5.2 Results

In the initial analysis, neither the number of predicted heads or the number of incomplete dependencies showed significant predictive power. We then limited the analysis to content words,[10] since the heavy use of functional *words* is a major difference from Japanese, which mainly use functional *morphemes*. The results reported below are therefore exploratory.

Psychometric predictive power The number of predicted heads shows a significantly positive ΔMSE ($p = 0.03$), while the number of incomplete dependencies is only marginally significant ($p = 0.07$). The difference is not significant. Adding one to the other does not significantly improve the model fit (Figure 8 and Table 2).

Coefficients The number of predicted heads shows a consistently positive effect, while the number of incomplete dependencies does not. The difference is significant ($p = 0.01$) (Figure 9).

Variation between participants. 22.2% of participants consistently slow down when the number of predicted head is high, while 8.9% speed up (Figure 10). The tradeoff between the antilocality effect and the slowdown due to high maintenance cost is not observed: the coefficient for the antilocality effect is numerically higher for participants who slow down. The difference is marginally significant when the number of dependencies is analyzed ($p = 0.06$), but not significant when the number of predicted heads is analyzed.

In summary, the significant effect of the number of predicted heads is consistent with the view that syntax helps memory-efficient processing. Beyond that, we do not observe the interesting results observed in Japanese. We consider this as pointing to issues to be resolved to generalize the current theory to languages like English (see Limitations).

6 Discussion

We showed that the reading slowdown due to information maintenance in a head-final language is affected by both the number of predicted heads and incomplete dependencies. The number of predicted heads is associated with a larger impact on reading times, though the two factors cannot be reduced to the other. We further showed that readers have different strategies for maintenance, and their choice affects how much they benefit from contextual prediction. These results are consistent with the view that the information maintenance of

[9] The following dependency types are regarded as adjunctions: `appos`, `acl`, `acl:relcl`, `advcl`, `advmod`, `amod`, `cc`, `compound`, `compound:prt`, `conj`, `dep`, `det:predet`, `discourse`, `nmod`, `nmod:npmod`, `nmod:poss`, `nmod:tmod`, `nummod`, `parataxis`, `punct`.

[10] Words of the following POS are included (`x` stands for any suffix): `CD`, `JJx`, `NNx`, `NP`, `RBx`, `VBx`.

a rational reader can be modeled by a syntactically-informed strategy aiming at better prediction or integration of future inputs.

6.1 Syntax is a good guide

We analyzed syntax-based metrics of maintenance cost. To be clear, we are not strongly committed to the mental reality of syntactic structure (cf. Chomsky, 1965). Rather, our point is that, if syntactic structures exist in language at least as "real patterns" (Futrell and Mahowald, 2025), the maintenance strategy of a rational reader should closely align with it.

It is interesting in this regard that the hierarchical syntactic structure by design offers an efficient maintenance strategy. A crucial feature of syntax exploited here is that only the top element of a dependency (sub)structure can in principle interact with the outside. If human language was less restrictive, freely allowing crossing and circular dependencies, the processor would have to maintain every word it has recognized, and would be quickly overburdened. Cognitive limitations on information maintenance can thus provide an explanation of why language has hierarchical syntax (Christiansen and Chater, 2016).

6.2 Related work

The current study shares with the resource-rational lossy-context surprisal (RR-LCS) model (Hahn et al., 2022) the view that human language users make a rational use of memory to minimize prediction error (also see Xu and Futrell, 2026). We believe that the current work is complementary to that line of work. Like ours, the RR-LCS model assumes that the processor minimizes prediction error (surprisal) by retaining important words in the context. The key difference is that the RR-LCS model implements this strategy using machine learning techniques. The prediction error is operationalized as GPT-2 surprisal, and the retention probability for each word is optimized using a dedicated neural network. Admittedly, a machine learning approach may give a more precise estimation of the rational retention strategy than a syntax-based approach, taking lexical properties of the words into account. An advantage of the syntax-based approach proposed here is that it can be transparently embedded in an algorithmic-level theory of resource-rational processing (Marr, 1982). It can also be connected to theoretical investigations about why language has structure that it has, as argued above.

The current results show that both predicted heads and incomplete dependencies slow processing. This is in a direct opposition to the observation of Nakatani and Gibson (2008). In a controlled experiment in Japanese, they observed that additional incomplete dependencies result in *faster* reading time. They used sentences like (5) (in the actual stimuli they are embedded within a matrix clause) and manipulated the presence of additional nouns preceding the verb, indicated in parentheses in (5):

(5) (zimusyo-no) sinnyusyain-ga (zimusyo-de)
(office-GEN) freshman-NOM (office-at)
(kokyaku-ni) tyumonsyo-o hassosita
(client-DAT) order.sheet-ACC sent
'the freshman (in the office) sent the order-sheet (to the client) (in the office)'

They observed that the accusative noun (*tyuumonsyo-o*) was read faster when a locative or dative noun precedes it. They take this result as evidence against incomplete dependencies as a determinant of maintenance cost.

The faster reading time could possibly be a predictability effect, however. The presence of a dative noun, for example, reduces the probability of another dative noun to follow. This can result in an increase in the relative probability of an accusative noun, making an accusative noun more predictable than when a dative noun does not precede (see Levy, 2008, for a similar argument regarding English PPs). Such a facilitation effect can mask the underlying maintenance cost.[11] Unlike their study, the current study controls for predictability by including GPT-2 surprisal in the regression model. The significantly positive coefficient found in the current analysis suggests that incomplete dependencies do impose maintenance cost. A factorial experiment that controls for predictability would be supplementary to the current corpus-based work in understanding the contribution of the two factors of maintenance cost.

7 Conclusion

Faced with fleeting language, a rational language user should use their limited memory wisely. We hypothesized that syntax offers a memory-efficient strategy of memory use, and our results indeed suggest that many readers of a head-final language can be modeled as adopting such a strategy.

[11] This remains a speculation since the raw data is not publicly available.

Limitations

The patterns we observed in Japanese (independent effects of the two types of maintenance cost; tradeoff between antilocality and slowdown for maintenance) were not replicated in English. We do not take this result as immediately undermining the notion of syntactically-guided memory-efficient processing, but as highlighting some issues in implementing the idea in typologically diverse languages. One obstacle for a non-head-final language like English is the argument/adjunct distinction. When a head requires an argument to follow, it counts as a predicted head; when an adjunct follows a head, it need not be predicted. We relied on the UD annotation to detect right adjuncts and excluded them from predicted heads, but the distinction is not clear-cut (MacDonald et al., 1994; Manning, 2003), causing a problem for estimating maintenance cost. Another difference that might account for the different results with Japanese and English is that functional morphemes are suffixes in the former but words in the latter. In the current formulation, only the latter counts as triggering prediction. But functional words carry relatively little information about the likely continuation compared to content words (e.g., *the* versus *delicious*). The English results may particularly be affected by the fact that the current approach does not take such differences into account.

Another problem, common to both Japanese and English, is that the we only considered the gold parse. If a reader holds multiple parses at a time, with a probability assigned to each of them, there is no obvious way to calculate the aggregate maintenance cost for all the parses. Weighting the maintenance cost for each parse by its probability is not necessarily appropriate, since there is no a priori reason to assume that a parse with lower probability occupies less memory space. How to calculate maintenance cost under parallel parses is a topic for future work.

These problems can be addressed if we give up a syntax-based theory of storage and move to an information-theoretic one, using a large language model to estimate predictive information. Our attempt in this direction has already proved to be promising, but the number of predicted heads remained independently significant (Kajikawa et al., 2026). Future work should probably integrate syntactic and information-theoretic perspectives for a deeper understanding of memory-efficient processing in humans.

An alternative account of storage cost we did not pursue here is that holding multiple heads or dependencies of the same kind is costly (Stabler, 1994; Lewis, 1996). Although the intuition of this idea is partially covered by counting the number of predicted heads, it is beyond the scope of the current study to develop a specific definition of the same kind of heads or dependencies that can be tested against naturalistic data.

We focused on self-paced reading data in this study since the notion of storage is most straightforward in a paradigm that does not allow regression. In the eye-tracking-during-reading paradigm, readers may put less effort in storage since they can freely look back on the past context. A comparison with eye-tracking data is an interesting venue for future study.

Acknowledgments

This study is dedicated to the late Akiyo Fukatsu, a cherished and warm-hearted colleague whose insatiable curiosity about language will continue to inspire and drive us.

We would like to thank the reviewers for their insightful suggestions. This work is supported by JSPS KAKENHI Grant number JP25K22996.

References

Masayuki Asahara. 2022. Reading time and vocabulary rating in the Japanese language: Large-scale Japanese reading time data collection using crowdsourcing. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*, pages 5178–5187, Marseille, France. European Language Resources Association.

Masayuki. Asahara, Hiroshi. Kanayama, Takaaki. Tanaka, Yusuke. Miyao, Sumire. Uematsu, Shinsuke. Mori, Yuji. Matsumoto, Mai. Omura, and Yugo. Murawaki. 2018. Universal Dependencies version 2 for Japanese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. 2014. Archiving and Analysing Techniques of the Ultra-Large-Scale Web-Based Corpus Project of NINJAL, Japan. *Alexandria*, 25(1-2):129–148.

Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. 2019. BCCWJ-EyeTrack. *Gengo Kenkyu*, 156:67–96.

Jonathan Brennan. 2016. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

Evan Chen, Edward Gibson, and Florian Wolf. 2005. Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, 52(1):144–169.

Noam Chomsky. 1965. *Aspects of the theory of syntax*. MIT Press.

Morten H. Christiansen and Nick Chater. 2016. The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39:e62.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Fernanda Ferreira, Karl G. D. Bailey, and Vittoria Ferraro. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1).

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, page 1–98.

Edward Gibson. 1991. *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown*. Ph.D. thesis, Carnegie Mellon University.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec P. Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, Brain*, pages 95–126. MIT Press, Cambridge, MA.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *PNAS*, 119(43).

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL 2001*.

Shinnosuke Isono. 2024. Category Locality Theory: A unified account of locality effects in sentence comprehension. *Cognition*, 247:105766.

Shinnosuke Isono, Kohei Kajikawa, Yohei Oseki, and Masayuki Asahara. 2025. Modeling memory effects in a head-final language with category locality. *PsyArXiv*.

Marcel Adam Just, Patricia A. Carpenter, and Jacqueline D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2):228–238.

Kohei Kajikawa, Shinnosuke Isono, and Ethan Gotlieb Wilcox. 2026. Information-theoretic storage cost in sentence comprehension. *arXiv preprint arXiv:2602.18217*.

Yuki Kamide and Don Mitchell. 1999. Incremental prehead attachment in Japanese parsing. *Language and Cognitive Processes*, 14(5-6):631–662.

John Kimball. 1973. Seven principles of surface structure parsing in natural language. *Cognition*, 2(1):15–47.

Lars Konieczny. 2000. Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29(6):627–645.

Lars Konieczny and Philipp Döring. 2003. Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In *Proceedings of the 4th International Conference on Cognitive Science*, pages 330–335. University of New South Wales.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Richard L. Lewis. 1996. Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, 25:93–115.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Maryellen C. MacDonald, Neal J. Pearlmutter, and Mark S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48:345–371.

Christopher D. Manning. 2003. Probabilistic Syntax. In *Probabilistic Linguistics*. The MIT Press.

David Marr. 1982. *Vision*. WH Freeman, San Fransisco, CA.

George A. Miller and Noam Chomsky. 1963. Finitary models of language users. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, editors, *Handbook of mathematical psychology*, volume 2, pages 419–491. Wiley.

Don C. Mitchell. 1984. An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In David E. Kieras and Marcel Adam Just, editors, *New methods in reading comprehension research*, pages 69–89. Erlbaum, Hillsdale, NJ.

Edson T. Miyamoto. 2002. Case markers as clause boundary inducers in Japanese. *Journal of Psycholinguistic Research*, 31(4):307–347.

Kentaro Nakatani. 2023. Locality-based retrieval effects are dependent on dependency type: A case study of a negative polarity dependency in Japanese. In Masatoshi Koizumi, editor, *Issues in Japanese Psycholinguistics from Comparative Perspectives (vol. 2)*, pages 31–54. De Gruyter Mouton.

Kentaro Nakatani and Edward Gibson. 2008. Distinguishing theories of syntactic expectation cost in sentence comprehension: Evidence from Japanese. *Linguistics*, 46(1):63–87.

Kentaro Nakatani and Edward Gibson. 2010. An online study of Japanese nesting complexity. *Cognitive Science*, 34(1):94–112.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Hajime Ono and Kentaro Nakatani. 2014. Integration costs in the processing of Japanese *wh*-interrogative sentences. *Studies in Language Sciences*, 13:13–31.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Bojana Ristic, Simona Mancini, Nicola Molinaro, and Adrian Staub. 2022. Maintenance cost in the processing of subject–verb dependencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(6):829–838.

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Edward. P. Stabler. 1994. The finite connectivity of linguistic structures. In C. Clifton Jr., L. Frazier, and K. Rayner, editors, *Perspectives on sentence processing*, pages 303–336. Erlbaum.

M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Shravan Vasishth and Richard L. Lewis. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 82(4):767–794.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721, Singapore. Association for Computational Linguistics.

Weijie Xu and Richard Futrell. 2026. Strategic resource allocation in memory encoding: An efficiency principle shaping language processing. *Journal of Memory and Language*, 146:104706.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

Appendix: Figures and Table for Experiment 3

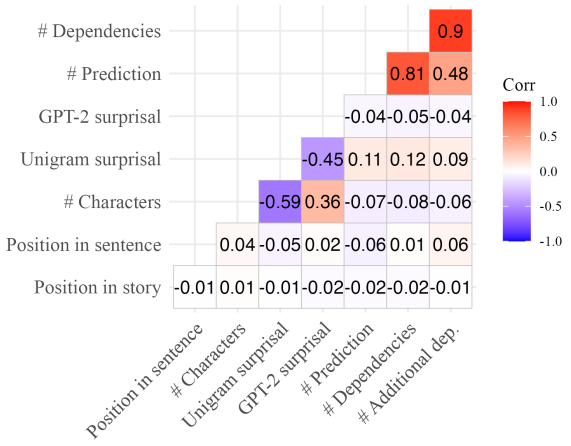

Figure 7: Correlations between variables used in the regression analysis (English). # = number of.

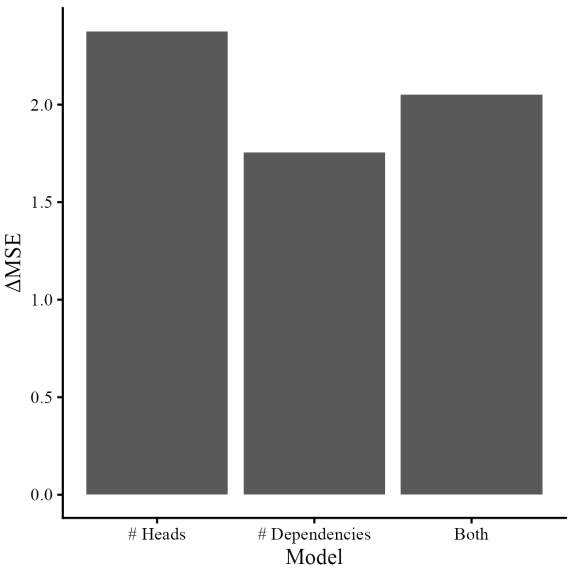

Figure 8: ΔMSE per word by model (English). None of the comparisons were significant.

	ΔMSE
Position in the story	290.1
Position in the sentence	0.1
# Characters	61.4
Unigram surprisal	32.8
GPT-2 surprisal	74.0
# Heads	2.4
# Dependencies	1.8
Both	2.1

Table 2: ΔMSE of predictors.

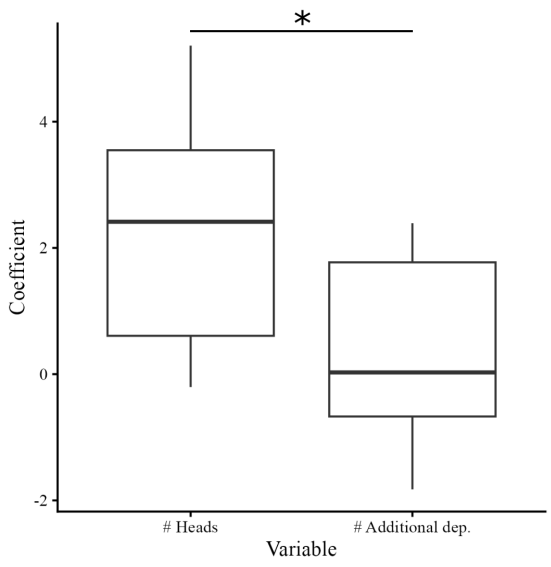

Figure 9: Distribution of estimated coefficients of the metrics of maintenance cost (English). The star indicates the significance of permutation test: * for $p < 0.05$

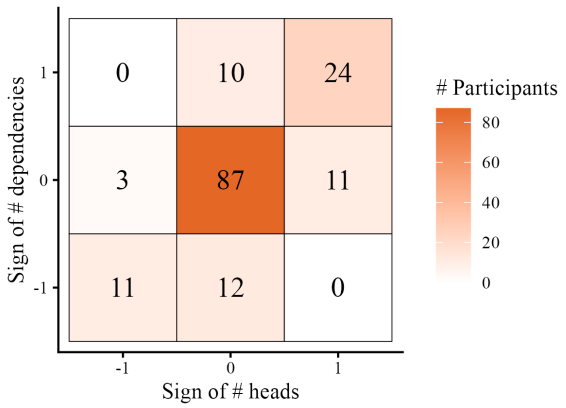

Figure 10: Classification of participants by their response to high maintenance cost (English). 1 indicates significant slowdown, -1 indicates significant speedup, and 0 indicates significance in neither direction.

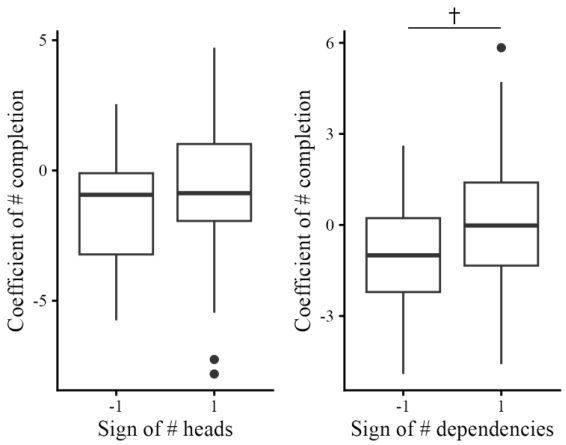

Figure 11: Distribution of anti-locality effect (y-axis) by participants' maintenance strategy (x-axis) (English). The star indicates the significance of the permutation test: † for $p < 0.1$