Title
Readers do not strongly rely on full-context information, but might utilize local word statistics, when ‘correcting’ word transposition errors in text

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors
Huang, Kuan-Jung
Staub, Adrian

Publication Date
2022

Peer reviewed

Readers do not strongly rely on full-context information, but might utilize local word statistics, when 'correcting' word transposition errors in text

Kuan-Jung Huang (kuanjunghuang@umass.edu) & Adrian Staub (astaub@umass.edu)
Department of Psychological and Brain Sciences, University of Massachusetts Amherst
135 Hicks Way, Amherst, MA 01003

Abstract

Rational inference over a noisy channel can potentially explain readers' occasional misreading. We tested if the prior probability of an intended message modulates the rate of misreading a transposed-word sentence as grammatical. In Experiment 1 we manipulated the cloze probability of a word given its full context (*Because my socks had holes, I bought a new* ***pair/pack***) but found no reliable effect on the rate of noticing word transpositions (*pair new* vs. *pack new*). In Experiment 2 we manipulated the 4-gram frequency of the sequence ending with the transposed words and again found no effect (*I always* ***know what mean they*** **vs.** ***love they***). We conclude readers do not effectively exploit full-context information to derive nonliteral messages. Despite the results of Experiment 2, comparison of error rates across conditions in several experiments suggests a role for local ngram statistics, though perhaps only in a restricted range of ngram frequency.

Keywords: rational inference; misreading; ngram statistics

Introduction

While traditional approaches to sentence comprehension during reading assume that the parsing system takes in noise-free perceptual input (e.g., Frazier & Fodor, 1978), recent frameworks view sentence comprehension as a rational process that takes into account potential noise during the communication process (Levy, 2008a; Futrell, Gibson, & Levy, 2020). This has been supported by empirical evidence of nonliteral reading in both behavioral and eye-tracking studies (Levy, Bicknell, Slattery, & Rayner, 2009; Gibson, Bergen, & Piantadosi, 2013; Staub, Dodge, & Cohen, 2019). This rational approach can be formalized as optimal Bayesian decoding of the intended, underlying message based on the perceived, surface message, as in the following equation:

$$P(s_i|s_p) \propto P(s_i)\, P(s_p|s_i)$$

Here s_i is a message intended by the producer, and s_p is the message perceived by the comprehender. $P(si)$ is the prior probability of an intended message based on the reader's language experience and world knowledge. $P(s_p|s_i)$ is the likelihood of the intended message being corrupted into the perceived message by noise during the communication process (e.g., the perceiver's errors, the producer's errors, environmental noise). The probabilities $P(si)$ and $P(s_p|s_i)$ jointly determine the probability of the perceived surface message ultimately being comprehended nonliterally. The more similar in form the intended and the perceived messages are, and the more plausible the intended message is, the more likely a nonliteral reading will be.

However, it is not known exactly what readers use as the prior ($P(s_i)$) for this Bayesian calculation. The current study addresses this question by separately manipulating two variables potentially associated with $P(s_i)$, in the empirical domain of misreading sentences containing word transpositions, repetitions, and omissions. In other words, we aim to identify what properties of the error-involving words determine the detectability of these errors.

Recent studies have reported readers' occasional failure to notice that two words in a sentence are actually transposed (Mirault, Grainger, & Snell, 2018; Huang & Staub, 2021b, see (1)), that a *the* is repeated, or that a *the* is omitted in a sentence (Staub et al., 2019).

(1) A clear sky blue is common here.

This misreading of such ungrammatical sentences as grammatical sentences could be evidence for reading as a rational process. The perceived transposition of the words *sky* and *blue* could, for example, be attributed to the combination of a possibility that the reader has not fixated the two words in the correct order, and the high probability of the intended message *a clear blue sky* (see Huang and Staub, 2021c, for review).

In a post-hoc analysis, Huang and Staub (2021a) observed a high item-wise correlation between the rate of failure to notice transpositions and the bigram log frequency of the transposed words in their canonical order (e.g., *blue sky*), even though this variable was not systematically manipulated. As a follow-up investigation, we re-examined the data from Experiment 3 in Huang[1] (2021, unpublished master's thesis), and again found a strong correlation (Figure 1) between the rate of failure to notice a transposition and the trigram ending with the transposed words (e.g., *clear blue sky*[2]). Notably, while several factors were experimentally manipulated, differences in the error rate across the conditions appeared to be explainable also by differences in mean trigram log frequency (the bullseyes in Figure 1), and moreover, within each experimental condition, the item-wise variation also was well captured by trigram log frequency. This finding provides preliminary support for the rational

[1] The experimental paradigm in both Huang (2021) and Huang and Staub (2021a) was the same as in the current study. See *Methods* for detail.

[2] As in the example sentence (1), all transposed words were at the third and fourth positions in the sentence, in that experiment.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

approach, as the frequency of the intended sequence presumably reflects its prior probability, $P(s_i)$. The higher the frequency of the critical words in their untransposed order, the more likely it is that readers will interpret them as occurring in that order.

However, if reading is fully rational such that a reader makes use of all available information sources, the intended message that is considered should include all the preceding context, i.e., 4-gram log frequency in Huang's material, e.g., *a clear blue sky*. However, a notable problem with ngram models is the sparsity of exactly matching long sequences in any corpus. Indeed, among the 120 items used in Huang (2021), 97 had zero 4-gram occurrences in the 1-billon word COCA corpus (Davies, 2019). Experiment 1 of the current study formally tests the hypothesized role of $P(s_i)$, with s_i containing the full preceding context, while avoiding the sparsity problem of the ngram approach. We manipulated the cloze probability of the second word in its canonical order (e.g., *P(sky|a clear blue)*) as a proxy for the probability of the whole string: The conditional probability[3] of a word given its preceding context is proportional to the probability of the whole sequence, when the preceding context is held the same:

$$P(w\_1, w\_2, \ldots, w\_n) \propto P(w\_n|w\_1, w\_2, \ldots, w\_n\text{-}1)$$

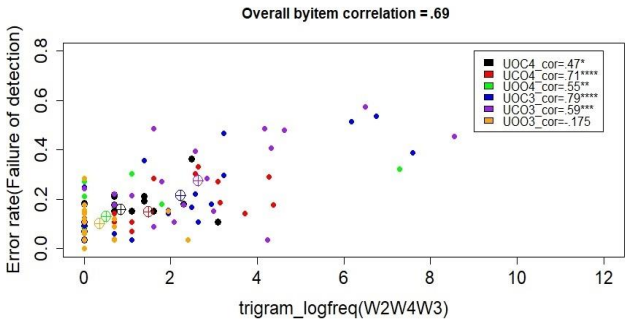

Figure 1. Exploratory analysis of correlation in Experiment 3 of Huang (2021). Bullseyes represent condition means; filled points represent individual items. Trigram log frequency was measured in the sequences' canonical order (hence Word2-Word4-Word3). See *Discussion* of Exp. 2 and Footnote 7 for details of the experimental conditions.

Experiment 1

Methods

Participants One hundred seventy participants with IP addresses in the United States were recruited online via Amazon Mechanical Turk, all self-reported to be native English speakers. Eighty-eight were excluded from analysis (final N = 82) as (1) they failed to provide human-like or native-speaker-like answers to our antibot questions (n=55) or (2) their accuracy on the comprehension questions following fillers or their accuracy in accepting well-formed sentences was below 70% (n=26) or (3) their accuracy in detecting transposition errors was significantly lower than chance level (i.e., answering correctly in fewer than 6 transposition trials—the 5th percentile of a binomial distribution with 20 trials and 50% probability of success, n =7).

Materials The within-item manipulations in Experiment 1 were the order of two critical words (*grammatical*, 2a-b, vs. *transposed*, 2c-d) and the cloze/conditional probability of the second transposed word in the words' grammatical order (*high*, 2a and 2c, vs. *low*, 2b and 2d).

(2a) Because my socks had holes, I bought a new pair…
(2b) Because my socks had holes, I bought a new pack…
(2c) *Because my socks had holes, I bought a pair new…
(2d) *Because my socks had holes, I bought a pack new…

We adapted 80 stimuli from a large-scale cloze norming study by Peelle et al. (2020), which elicited at least 100 cloze responses to each sentence. The second critical word in (2a) was highly predictable (modal response, ranging from 0.51 to 0.9, mean = 0.67), while the second critical word in (2b) was unpredictable, but did appear in the norms (cloze probability ranging from .01 to .06, mean = 0.03). Items were selected based on the following criteria: (a) the critical word and the immediately preceding word were both between 3 and 6 characters and did not differ in length by more than one character; (b) the two responses were not synonyms (e.g., *pig* and *hog*); and (c) the target and the word immediately preceding formed a determiner-noun or adjective-noun sequence. These two structures occurred in equal number and were treated as two conditions, along with the predictability manipulation, leading to a 2 (grammatical/transposed) x 2 (structure) x 2 (predictability) design. Contexts following the target words were created to fit both high-predictable and low-predictable targets. Each participant saw either the grammatical or transposed version and either the predictable or unpredictable version of a given item, for a total of 40 critical transposed sentences and 40 grammatical counterparts. Another 160 filler sentences, of similar length but varying in sentence structure, were included.

Procedure Participants were presented with one sentence at a time and instructed to read at their own speed. After finishing reading the sentence, they hit a button to proceed to a question screen. For the filler trials, the following question was a comprehension question. For the critical trials, the following question was an error-detection question: *Was*

[3] We note that cloze probability is only one way to quantify conditional probability and a cloze task might not capture subtle differences among items in conditional probability. However, as will be seen in the methods section, our manipulation is categorical with an extreme difference between the two groups (highly probable response vs. very unlikely response), leaving little concern about task sensitivity.

there anything wrong with the sentence? Filler and critical trials were randomly intermixed. Every trial started with a fixation mark on the left side of the screen lasting 1.25 seconds. The sentence then appeared several spaces to the right of the fixation mark.

Results

Prior to the calculation of error rates, trials with excessively long or short reading time or response time were excluded (reading time < 150 ms or > 15000 ms; RT < 100 ms or > 15000 ms), which accounted for 5.6% of the data.

Generalized linear mixed effect models (GLMMs) were run for the accuracy data with a logistic link function. We used the bobyqa optimizer with 200,000 iterations to improve convergence. All models were constructed with maximal random slopes unless there was a singularity or convergence issue (Barr, Levy, Scheepers, & Tily, 2013), in which case the highest-level random factors and/or correlation terms were removed.

Table 1 shows the mean error rates and by-subject standard error for each condition. There is only a very weak numerical trend (≈3%) indicative of a predictability effect on error rate in the transposition conditions.

Table 1: Condition means for error rate in detecting grammatical and transposed sentences (or error rate in comprehension questions, for fillers) and by-subject standard errors in Experiment 1.

	HPred, noun-det	LPred, noun-det	HPred, noun-adj	LPred, noun-adj
Gramm	5.8 (1.17)	7.3 (1.37)	4.9 (0.85)	4.9 (1.03)
Transp	25.0 (2.81)	22.1 (1.99)	22.1 (2.37)	18.5 (1.85)
Filler	3.7 (0.27)			

The model used sum-coded contrasts (0.5 vs. -0.5) to code the three factors. The full model (the three factors and all their interactions) indicated only a main effect of grammaticality and a marginal interaction between grammaticality and predictability. A reduced model (structure + grammaticality * predictability) showed again the interaction being marginal (z = -1.74) along with a marginal effect of structure (z = -1.72); higher predictability tended to lead both to fewer rejections of the grammatical sentences and to increased failure to notice the errors in transposed sentences. Critical to our main question, however, when looking at only the transposed-condition data, there was no significant effect of predictability (z = -1.28). Thus, we do not regard the experiment as confirming the hypothesis that a transposition is more easily overlooked when the second word is predictable, in the grammatical order.

Because our post-hoc analysis of the earlier experiment (Huang, 2021, Exp 3) revealed a correlation between failure to notice transpositions and trigram log frequency (Figure 1), we explored the effect of this variable in the current experiment. When treating trigram log frequency as a continuous variable and adding it into the models as a covariate, there was a consistent effect of trigram log frequency, with or without the two experimentally manipulated factors in the model (all *p*s <= .01), while there was still no significant predictability effect. As shown in Figure 2, however, the trigram frequency effect appeared to be weaker than in the earlier experiment.

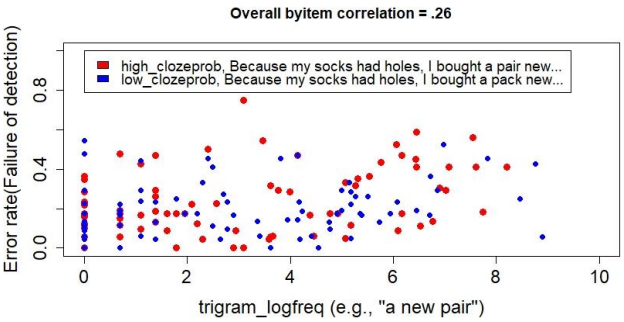

Figure 2. Scatterplot illustrating relationship between trigram log frequency and failure to notice transposition, by item, in Experiment 1.

Discussion

Experiment 1 was motivated by the observation, in a previous experiment, of a strong role for trigram frequency in explaining both between-condition and within-condition variance in the prevalence of failure to notice word transpositions, a pattern consistent with the rational inference framework. We hypothesized that the trigram frequency reflected $P(s_i)$, even though it does not capture full contextual information. We assumed that a fully rational processor should utilize all available contextual information in gauging how probable a message is, and as such, predicted that cloze probability, when the contexts are held the same, should be a good proxy for $P(s_i)$.

Contrary to our prediction, we failed to observe a reliable effect of predictability. This result suggests that readers may not be fully rational processors that constantly update information (cf. Levy, 2008b) or that readers' representation of the full context is lossy (Futrell et al., 2020). Instead, the correlation with trigram log frequency observed in the previous experiments, and in the present experiment, might simply reflect the utilization of local statistical information. That readers might prioritize local word information has been proposed in the literature (e.g., Tabor, Galantucci, & Richardson, 2004). Recent computational work also has shown that adding trigram and bigram frequency provided better fit to eye-tracking corpus data on top of effects of conditional word probability, word length, and unigram frequency (Goodkind & Bicknell, 2021; see also Duan and Bicknell, 2020). Given the null effect in Experiment 1, Experiment 2 was conducted to formally test

the hypothesis that readers utilize local word statistics to derive nonliteral, intended messages.

Experiment 2

Experiment 2 experimentally manipulated local ngram statistics while controlling other variables. Arnon and Snider (2010) suggested that language users keep track of the frequency of multi-word phrases, which in turn affects online processing of these phrases. In Arnon and Snider's study, the last word of a four-word sequence was manipulated such that stimulus pairs differed in their 4-gram frequency, but did not differ significantly in unigram (Word 4), bigram (Words 3 & 4), or trigram (Words 2, 3, & 4) frequency. In two phrasal-decision tasks, response time was faster when the four-word expressions had higher 4-gram frequency. The authors argued that linguistic units larger than words can be stored, represented, and retrieved as a whole, and accumulate as our experience grows. Recent work, however, proposed that the observed 4-gram frequency effect can be subsumed to forward and backward conditional predictability (Onnis & Huettig, 2021), which the authors viewed as supporting incremental compositional processing instead of precompiled chunk retrieval.

In Experiment 2 we manipulated local word statistics up to the 4-gram window. If readers' ability to detect transpositions is influenced by 4-gram frequency, this will provide even stronger evidence for language users' mental storage of large chunks (Arnon & Snider, 2010). On the other hand, this finding would be hard to explain by means of incremental compositional processing, as both the forward and backward conditional probabilities are near zero when sequences are presented in their transposed order, for both high and low 4-gram sequences (e.g., forward *P*(they | know what mean) = *P*(they | know what love) = 0; backward *P*(know what mean | they) = P(know what love | they) = 0).

Finally, Experiment 2 differed from Experiment 1 in that it involved not only word transpositions but also word repetitions and omissions. Staub et al. (2019) reported that readers had a strong tendency to fail to notice repetitions or omissions of *the* in a sentence, but rarely failed to notice a repeated content word. Huang and Staub (2021c) further argued that failure to notice word transpositions, repetitions, and omissions can potentially be explained with a unified account, namely rational inference over a noisy channel (see *Introduction*). Indeed, in yet another exploratory analysis with the data in Staub et al. (2019), we again found a significant correlation, across items, between error detection rates and trigram log frequency (Figure 3). For the current experiment, we thus included all three types of errors within an experiment and tested if local word statistics influenced detection of them in similar ways.

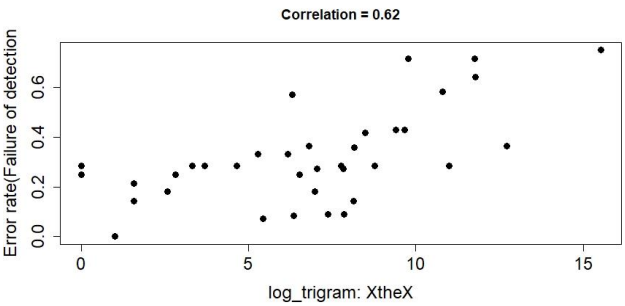

Figure 3. Exploratory analysis of *omitted-the* data from Staub et al. (2019): The more frequent a three-word sequence with *the* in the middle position, the more frequently the omission of *the* was not detected.

Methods

Participants A hundred sixty-six participants from the same pool as Experiment 1 initially participated; 72 were excluded from analysis using the same exclusion criteria as Experiment 1 (forty-three for Criterion 1, twenty-three for Criterion 2, and six for Criterion 3, final N = 94).

Materials We adopted about half of the items used in Arnon and Snider (2010). We found that for many of Arnon and Snider's item pairs, 4-gram frequency differences calculated from COCA (Davies, 2019) did not support Arnon and Snider's original categorization. These items were excluded, and new items were developed, for a total of 12 pairs targeting each of the three error types. The ungrammatical versions of the items are illustrated in (3); the error type manipulation was a between-item factor. Versions 3a, 3c, and 3e are high 4-gram frequency.

(3a) *I always **know what mean they** even though we are from two different generations. (T)
(3b) * I always **know what love they** even though we are from two different generations. (T)
(3c) *They have looked **all over country** to find you. (O)
(3d) *They have looked **all over house** to find you. (O)
(3e) ***We're here to help help** the people in the airport. (R)
(3f) ***We're here to keep keep** the people in the airport. (R)

Different constraints were applied for the different error types. For transpositions, the two transposed words could not differ in length by more than one letter. For omissions, the omitted word was always at the third position in the 4-gram, and was always a function word; *the* is the ommitted word in (3c-d), but a range of function words was used. For repetitions, the repeated word was always the fourth word of the 4-gram and was always a short word with high unigram frequency[4].

Table 2 presents the ngram frequency statistics for each condition. While we aimed to manipulate only 4-gram log

[4] There was one exception each among the transposition and repetition items, where a target word was a rather long word.

frequency while matching trigram, bigram, and unigram log frequency between the high and low groups, it was empirically difficult to do so. The resulting items thus were only matched on bigram and unigram log frequency. The difference of group means in trigram log frequency was in the same direction as in 4-gram log frequency. Among the 36 pairs, all except for one item in the high-4-gram group had higher trigram frequency than its counterpart. That is, the high and low groups differed similarly in both trigram log frequency and 4-gram log frequency.

Table 2: Mean log-transformed frequency of ngram in each condition and the t-test result between each high and low condition. T: transposition; O: omission; R: repetition. *p*-value: one-tailed paired t-test. ***: < .001; NS: > .05

Log frequency	T	*p*	O	*p*	R	*p*
High (4-gram)	6.68	***	7.24	***	6.48	***
Low (4-gram)	4.89		4.31		3.77	
High (trigram)	7.87	***	8.33	***	7.69	***
Low (trigram)	6.59		6.28		5.78	
High (bigram)	10.15	NS	10.64	NS	10.08	NS
Low (bigram)	9.73		10.65		9.76	
High (unigram)	12.67	NS	12.60	NS	12.72	NS
Low (unigram)	12.81		12.52		12.49	

Each participant saw either the grammatical or ungrammatical version and either the high or low 4-gram frequency version of a given item, for a total of 18 critical ungrammatical sentences and 18 grammatical counterparts. Another 72 filler sentences, of similar length but varying in sentence structure, were included.

Procedure The paradigm was the same as in Experiment 1.

Results

Analysis of data was the same as in Experiment 1, applying the same exclusion criteria for trials with excessively long or short reading or response time (4.8%). Table 3 presents the mean error rates and by-subject standard errors for each group.

Table 3: Condition means for error rate in detecting grammatical and transposed sentences (or error rate in comprehension questions, for fillers) and by-subject standard errors in Experiment 2. H: high tri/4-gram frequency; L: low tri/4-gram frequency; T: transposition; O: omission; R: repetition

	HT	LT	HO	LO	HR	LR
G	7.2 (1.67)	4.3 (1.37)	3.6 (1.19)	7.6 (1.45)	5.1 (1.43)	6.8 (1.48)
U	34.6 (2.81)	36.4 (2.94)	47.4 (3.53)	40.1 (3.45)	11.6 (2.12)	11.9 (2.18)
F	2.0 (0.34)					

As it was not our main goal here to compare the prevalence of failure of detection of different error types, we analyzed the data separately for transpositions, omissions, and repetitions. For each dataset, we first created a model with grammaticality and frequency and their interaction, using sum-coded contrasts. For transposition, there was only a significant effect of grammaticality ($p < .001$), without a significant interaction nor any hint of an effect of tri/4-gram frequency. For omission, there was a significant effect of grammaticality ($p < .001$) along with an interaction ($p < .05$); high tri/4-gram frequency reduced rejections of the grammatical sentences and increased failures to notice the omission error. However, when only looking at the omission-condition data, there was no significant effect of tri/4-gram frequency ($z = -0.78$). For repetition, none of the effects were significant (all $|z|$s < 0.89).

Discussion

Experiment 2 directly tested the effect of tri/4-gram log frequency on failure to notice errors in sentences. To our surprise, for none of the three error types was there compelling evidence that tri/4-gram log frequency influenced failure to notice errors. For repetition errors, this was possibly due to a floor effect, as readers rarely missed this type of error, consistent with the data from repeated content words in Staub et al. (2019). For omission errors, there was a numeric trend in the expected direction for tri/4-gram log frequency. For transposition errors, there was no hint at all of a difference between the high and low conditions[5].

While we failed to obtain the predicted effects, two results are worth noting. First, we found that omissions of function words were very easily missed. Our error rates were even higher than the 32.5% reported by Staub et al. (2019) where the omitted word was always *the*. Whether a strong tendency to overlook an omitted word is only limited to short function words should be further explored. Second, the rate of failing to detect word transpositions was also high. In fact, these might be the most illusory transposition items of any reported in the literature (Huang, 2021; Huang & Staub, 2021a; 2021b; Mirault et al., 2018; Wen, Mirault, & Grainger, 2021a; 2021b, ranging 10-27.7%). This indicates that there

[5] We note that while the failure to obtain the expected effect in the omission condition could have been due to insufficient power, the trend in error rate in the transposition condition is in the opposite direction from our prediction.

might have been some idiosyncratic features in the current experiment underlying the particularly high error rates.

Previous studies have investigated several potential features of two transposed words that may influence the prevalence of failure to notice the transposition. These include word length, word length difference, word class, ungrammaticality point, and ngram frequency. Most of our stimuli in Experiment 2 would be categorized as UOC3, in reference to the experiment in Huang (2021; see Figure 1 above). That is, the two words in their transposed order were an open-class word followed by a closed-class word, and the ungrammaticality emerged at the third word position in the sentence, i.e., the first tranposed word (e.g., *I have **say** to …*). In addition, as in that experiment, the two transposed words were almost always 3-4 letters long. Therefore, it appears that the most obvious difference between the current items and the corresponding UOC3 items in Huang (2021) is their ngram log frequency (4-gram, trigram, and bigram), which is on average much higher in both conditions of the present study than in Huang (2021).

The items in the two experiments in the present study differed on several other dimensions, and thus the difference in failure to notice transpositions across these two experiments could be attributed to multiple causes. Nevertheless, as shown in Figure 4, one of the differential dimensions is again their trigram log frequency[6]. This figure plots the mean rate of failure to notice transposition errors in a condition against the condition's mean trigram log frequency, across both of the experiments in the present study and the two previous experiments we have mentioned (Huang & Staub, 2021a; Huang, 2021, Experiment 3), and reveals a very clear, if possibly nonlinear, relationship between the two variables.

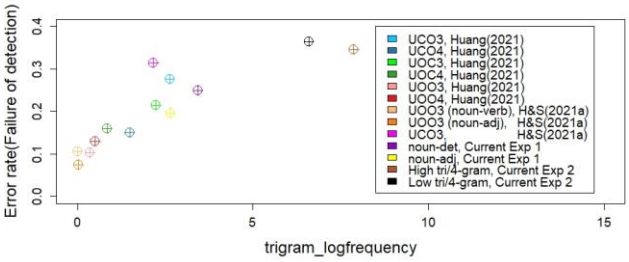

Figure 4. Scatterplot of condition means of trigram log frequency and detection error rate, combining transposition data from four experiments.[7]

To reconcile the lack of tri/4-gram log frequency effect under our well-controlled experimental manipulation in Experiment 2 and the recurring correlational hints within and across multiple experiments, our tentative explanation is that the items we used in Experiment 2 were in a suboptimal location on the frequency scale. That is, while the two experimental conditions significantly differed in trigram and 4-gram log frequency, in both conditions the means of these variables were quite high. This is visually obvious from Figure 4 (the right two points vs. the rest). We therefore plan a follow-up confirmatory experiment with within-item pairs whose trigram log frequency falls at a lower point on the log scale, for example at around 1 and 4.

Conclusions

The current study, in two experiments, tested the effects of conditional probability of the second transposed word and tri/4-gram log frequency surrounding the transposed words on readers' tendency to overlook the transposition error. Both variables presumably reflect $P(s_i)$, the prior probability of an intended message, a factor that should play a role in determining the probability of nonliteral reading under a rational inference account. The difference between conditional probability and tri/4-gram frequency is in the size of the window over which s_i is evaluated. Readers' use of conditional probability as $P(s_i)$ is expected if reading is a fully rational process, while the use of local ngram frequency as $P(s_i)$ suggests a processor with a limited attentional window. In neither experiment did we find the predicted effects, however. We believe that it is safe to conclude that there is a null or extremely small effect of conditional probability, given our extreme manipulation and the large number of items and participants in Experiment 1. On the other hand, we tentatively suggest that the null effect of tri/4-gram log frequency in Experiment 2 may have been due to the specific range of this variable that our manipulation targeted.

With respect to Arnon and Snider (2010), our results may be seen as failing to support their account, as our 4-gram manipulation fell in the same range as their mid-frequency bin in terms of occurrence per million words, but resulted in a null effect. This could be because of the fundamentally different phenomena investigated in the two studies, i.e., phrasal decision RT vs. misreading probability.

Finally, it is clear that even if local ngram frequency is ultimately found to be a determinant of failure to notice transposition errors, it appears unlikely to be the sole determinant, as substantial variance exists even for items with trigram frequency of zero. The rational approach currently adopted here is on the computational level. Other processing factors might also play an important role (e.g., variability in integration time and integrability of a word into the context, Huang & Staub, 2021b; see also Wen et al., 2021b) in explaining readers' failure to notice different errors in written sentences.

[6] We also point out that the mean bigram log difference between the *noun-determiner* and *noun-adjective* conditions in Experiment 1 was extremely large (means = 7.53 vs. 3.68) but did not lead to a significant difference in error rates, a potential piece of evidence against bigram being the relevant ngram.

[7] For additional explanation of condition labels: UOO3 is an open-class-open-class sequence where ungrammaticality arises at the third word (e.g., *An awfully **lamp dim** …*); UCO4 a closed-class-open-class sequence with ungrammaticality at the fourth word (e.g., *The boy **on sat** …*).

Acknowledgments

We thank Emily Kaye for her help with item development for Experiment 1 and Chuck Clifton for his comments on an early version of the manuscript. The study was supported in part by a grant from the National Science Foundation BCS 1732008 to Adrian Staub.

Open Practice Statement

All materials, data, and analysis scripts are available on https://osf.io/ft72e/

References

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of memory and language, 62*(1), 67-82.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), 255-278.

Davies, M. (2019) *The Corpus of Contemporary American English (COCA)*. https://corpus.byu.edu/coca/

Duan, Y., & Bicknell, K. (2020). A rational model of word skipping in reading: ideal integration of visual and linguistic information. *Topics in cognitive science, 12*(1), 387-401.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition, 6*(4), 291-325.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science, 44*(3).

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences, 110*(20), 8051-8056.

Goodkind, A., & Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. *arXiv preprint arXiv:2103.04469*.

Huang, K. J. (2021). *Visual, lexical, and syntactic effects on failure to notice word transpositions: evidence from behavioral and eye movement data.* Master's Thesis, Department of Psychological and Brain Sciences, University of Massachusetts Amherst.

Huang, K. J., & Staub, A. (2021a). Limits on failure to notice word transpositions during sentence reading. Poster presented at the 34th Annual CUNY Sentence Processing Conference, Philadelphia.

Huang, K. J., & Staub, A. (2021b). Using eye tracking to investigate failure to notice word transpositions in reading. *Cognition, 216*, 104846.

Huang, K. J., & Staub, A. (2021c). Why do readers fail to notice word transpositions, omissions, and repetitions? A review of recent evidence and theory. *Language and Linguistics Compass, 15*(7).

Levy, R. (2008a). A noisy-channel model of rational human sentence comprehension under uncertain input. *Proceedings of the conference on empirical methods in natural language processing*, 234– 243. Association for Computational Linguistics.

Levy, R. (2008b). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences, 106*(50), 21086-21090.

Mirault, J., Snell, J., & Grainger, J. (2018). You that read wrong again! A transposed-word effect in grammaticality judgments. *Psychological Science, 29*(12), 1922-1929.

Onnis, L., & Huettig, F. (2021). Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled' from memory or compositionally constructed on the fly?. *Brain Research, 1772*, 147674.

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.

Staub, A., Dodge, S., & Cohen, A. L. (2019). Failure to detect function word repetitions and omissions in reading: Are eye movements to blame?. *Psychonomic bulletin & review, 26*(1), 340-346.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language, 50*(4), 355-370.

Wen, Y., Mirault, J., & Grainger, J. (2021a). Fast syntax in the brain: Electrophysiological evidence from the rapid parallel visual presentation paradigm (RPVP). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(1), 99.

Wen, Y., Mirault, J., & Grainger, J. (2021b). The transposed-word effect revisited: the role of syntax in word position coding. Language, Cognition and Neuroscience, 36(5), 668-673.