Predictive Text Encourages Predictable Writing

Kenneth C. Arnold
Calvin University
Grand Rapids, Michigan
kcarnold@alum.mit.edu

Krysta Chauncey
Charles River Analytics
Cambridge, Massachusetts
krysta.chauncey@gmail.com

Krzysztof Z. Gajos
Harvard University
Cambridge, Massachusetts
kgajos@g.harvard.edu

ABSTRACT

Intelligent text entry systems, including the now-ubiquitous predictive keyboard, can make text entry more efficient, but little is known about how these systems affect the content that people write. To study how predictive text systems affect content, we compared image captions written with different kinds of predictive text suggestions. Our key findings were that captions written with suggestions were shorter and that they included fewer words that that the system did not predict. Suggestions also boosted text entry speed, but with diminishing benefit for faster typists. Our findings imply that text entry systems should be evaluated not just by speed and accuracy but also by their effect on the content written.

CCS CONCEPTS

• **Human-centered computing → Text input**; **Empirical studies in HCI**; • **Computing methodologies** → *Natural language generation.*

KEYWORDS

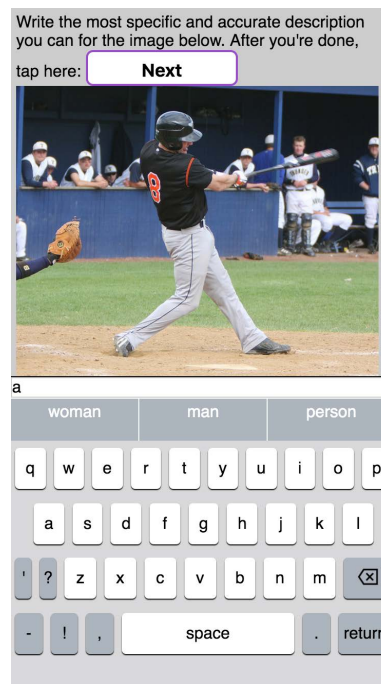predictive text, content effects of intelligent systems

1 INTRODUCTION

Predictive text suggestions are ubiquitous on touchscreen keyboards and are growing in popularity on desktop environments as well. For example, suggestions are enabled by default on both Android and iOS smartphones, and the widely used Gmail service offers phrase suggestions on both desktop and mobile [12]. The impacts of system design choices on typing speed, accuracy, and suggestion usage have been studied extensively [4, 8, 37]. However, relatively little is known about how text suggestions affect *what* people write. Yet suggestions are offered up to several times per second in the middle of an open-ended process of planning the structure and content of writing, so these suggestions have the potential to shape writing content.

Figure 1: **When writing with suggestions, people tended to choose shorter and more predictable wordings (such as 'man', in the screenshot above, instead of 'hitter' or 'baseball player'), than writing without such suggestions. (Image credit: https://flic.kr/p/6Yxc61)**

Most prior text entry studies have not been able to study the effects of suggestions on content because they either prescribed what text to enter (such as [8, 37]) or had no measures that were sensitive to changes in content. The few prior studies that did investigate the content effects of phrase suggestions did not have a no-suggestions baseline [3, 4], so the effects of single-word suggestions on content are still unknown.

Predictive systems are designed to offer suggestions that reduce writers' typing effort by offering shortcuts to enter one of a small number of words that the system predicts are most likely to be typed next. As such, the suggestions are, by construction, the words that are the most predictable in their context. Thus, writers who follow these suggestions may create writing that is more predictable than they would create without such suggestions.

To study what effects predictive text suggestions might have on content, we conducted a within-subjects study (N=109 participants, $109 \times 12 = 1308$ texts) where writers wrote captions for images while we varied characteristics of the predictive suggestions that

a keyboard offered. Our study compared always-available predictive suggestions against two alternatives. To measure the overall effect of suggestions on writing content, we compared to a baseline where suggestions were never shown. (We were able to make this comparison because we used measures of *content* that were insensitive to differences in *process*.) And to study the content effects of an intervention previously studied only for efficiency [37], we also compared always-available suggestions to a condition in which suggestions were hidden when the predictive system had low confidence.

Our key findings were that captions written with suggestions were shorter and that they included fewer words that that the system did not predict. Thus, predictive text led to more predictable writing. We also found that suggestions increased typing speed, but with diminishing returns for faster typists.

2 BACKGROUND

Our work builds on prior research on communication, text entry, language modeling, and human factors in automation.

2.1 Communication

Writing integrates processes at many different levels of cognition [16, 24, 42]. Technology has been developed to aid in some of the higher-level processes of writing, such as collaboration [2, 41], ideation [14], information gathering [5], and feedback [26].

Communication (written or spoken) is more than description of the current state of the world; it is goal-directed. Speech Act theory accounts for speakers' (and writers') communicative choices in terms of the effect that the communication has on the hearer (or reader). In particular, Rational Speech Act (RSA) theory posits that the choice of words depends on both the effect those words are expected to have on the listener and the cost of producing those words [22, 35].

Writing process theory suggests that considerations of word choice are not made *a priori* but rather during the course of text production [42]. Thus, these considerations may be influenced by events that happen during text production, such as a predictive text system offering a suggestion.

2.2 Text Entry

The low-level process of text entry has been the subject of extensive study and innovation. Predictive language modeling in text input interfaces were first developed to assist those with motor impairments and poor typists [15], but have seen much wider adoption today. They reduce motor effort and errors, but their overall effect on speed depends on system design choices [37] and individual characteristics [29]. Many different interactions have been evaluated for using predictive models in typing, such as spatial navigation [47], or dynamic adjustment of input parameters such as effective key sizes [6] or gaze dwell times [34]. Modern touchscreen keyboards use a flexible interface typically called the suggestion bar, which can be used to show completions, corrections (possibly automatically accepted) [8], alternative interpretations of ambiguous stroke gestures [40], and even emoji and other functionality. The predictions are usually generated by a language model that is

trained on a wide range of data [45, 46], though some implementations customize the predictions using the author's past writing or suggestions from conversation partners [18]. Recent systems have explored alternative interfaces, such as showing contextual phrase previews [3, 4] offering complete-sentence replies [27], and offering a single highly likely phrase continuation, like Google's Smart Compose [12].

Almost all evaluations of text input have relied on transcription studies: participants were given phrases to enter as quickly and accurately as possible [36]. In 2014, Kristensson and Veranen advocated composition tasks in text entry studies [30], but only a few studies published since then ([3, 4, 10, 45]) have heeded their advice. One of these studies presented a methodology to collect typing data in the wild, but the design consideration of participant privacy prevented collection of rich data about writing content [10]. Nevertheless, the study did find that people used suggestions extensively and that there were substantial differences in suggestion use between writers.

The most detailed studies of the effects of text entry on writing content have been by Arnold et al. [3, 4], who studied phrase suggestions on touchscreen keyboards. They found that phrase suggestions influence writing content much more strongly than single words [4], and that those influences can extend to the sentiment of writing in a review setting [3]. However, neither of their studies compared with a baseline of no suggestions; their findings leave open the possibility that single-word suggestions have no more influence on content than no suggestions at all.

2.3 Human Factors in Automation

The decision of if and when to proactively offer assistance is one of the key choices in the design of mixed-initiative user interfaces [25]. Predictive text systems are no exception: some systems offer suggestions almost constantly (e.g., the "suggestion bar" in most touchscreen keyboards), while other systems, such as Smart Compose [12], offer suggestions only in situations where they are highly confident that the suggestion will be accepted. One study found that this confidence thresholding intervention made text entry faster when compared with always-visible suggestions [37], but the effect of this intervention on content is unknown.

The diagnostic automation literature has found that confidence thresholding can influence human attitudes and behaviors. Diagnostic automation systems such as alarms can use a confidence threshold to trade off misses and false alarms, as described by signal detection theory. Varying the threshold can result in different human responses to the system; frequently studied dimensions include reliance, compliance [33], and trust [11]. These different responses can, in turn, affect how well people perform at detection tasks when using these systems [48].

Although these studies typically focused on binary decisions in repetitive contexts, they may have implications for open-ended tasks like text composition. For example, the confidence threshold may affect the degree to which people attend to predictive text suggestions, rely on them, or comply with them. If the system shows suggestions rarely but those suggestions are often useful (corresponding to a low false alarm rate), the writer may perceive system as being more useful, and thus pay more attention to it.

Perceptual considerations may also be relevant to how confidence thresholding affects writers. When the confidence level of the predictive system transitions from below-threshold to above-threshold, a suggestion will appear suddenly. Since the appearance of new content captures attention [32, 39, 49], the writer may pay more attention to the new suggestion than if suggestions had been always available. If the new suggestion is irrelevant, it may interfere with the writer's working memory [13]; if it is relevant, it risks out-competing the word that the writer would have generated [38].

3 RESEARCH QUESTIONS

Our main research question is: how do predictive suggestions affect what is written? Since we expect that the primary mechanism of this effect might be that people *accept* suggestions that are offered, our more specific question is:

RQ1: To what degree do people choose the words that the system suggests?

We also wondered how suggestions might affect the length of text entered. Since suggestions reduce the physical effort (number of taps) required to enter texts, we wondered if writers would choose longer texts when suggestions were available. So our second research question is:

RQ2: How do suggestions affect text length?

Predictive systems using confidence thresholding have been widely deployed [12] and have been studied for speed effects [37], but their effects on content are unknown. Since applying a threshold reduces the frequency at which at which the system presents suggestions, its suggestions may have overall less effect on the content written than when suggestions are always available. But subjective and perceptual factors (discussed in Section 2.3) may cause a relatively larger effect when the suggestions *do* appear. Since even the direction of any content effects is unclear, we ask:

RQ3: How does the effect of suggestions on writing content differ if only high-confidence suggestions are shown?

Finally, the literature is currently divided on the impact of intelligent text entry technology on speed. Some studies found speed and error rate benefits [1], especially for systems that permit ambiguous input gestures such as swipes [40] and imprecise tapping [45, 46]. But other studies failed to find speed benefits for predictive suggestions [4, 37]. The authors of those studies conjectured that the time required to attend to suggestions more than made up for the speed benefit of avoiding extra taps. However, the temporal costs of these two activities may have substantial individual differences [29], so there may be some people for whom predictive suggestions are indeed helpful (such as people with motor impairments) that may not be observed in small studies. Also, the speed impact of predictive suggestions depends on their accuracy, as pointed out by authors as early as [29]; recent advances in language modeling technology have substantially improved predictive accuracy. Finally, transcription studies require the additional task load of attending to the text to transcribe, so study design may have a large impact on text entry performance [36]. Few studies have measured the speed of text entry outside of transcription tasks (exceptions include [4, 45]). So our final question is:

RQ4: How does suggestion visibility affect text entry speed?

4 STUDY

To evaluate the effect of predictive text on writing content, we conducted a within-subjects experiment in which participants wrote captions for images while we varied the visibility of predictive suggestions that the keyboard offered.

4.1 Task

An open-ended writing task allowed us to measure the effect of suggestions on content. We chose image captioning as our task because it was short, controlled, and repeatable: many different images can be captioned in a short time, so a within-subjects design was feasible. The range of possible captions for a single image was wide enough to observe differences in content but narrow enough that the variance in content characteristics between writers was not too large.

In each trial, participants were instructed to write a "specific and accurate" caption for a given image, by typing on a simplified touchscreen keyboard. Figure 1 shows an example of the task.

4.2 Design

We manipulated a single factor, the VISIBILITY of suggestions presented by the touchscreen keyboard, with three levels:

ALWAYS The keyboard showed three predicted words above the keyboard, using the familiar "suggestion bar" interface.
NEVER No suggestions were shown (the suggestion bar was hidden)
ONLYCONFIDENT Like ALWAYS, except the keyboard only showed suggestions when the confidence of the predictive model exceeded a threshold.

The study was a within-subjects design: each participant wrote twelve captions, four with each level of VISIBILITY (NEVER, ALWAYS, and ONLYCONFIDENT). The order of conditions was counterbalanced across participants, but the images were presented in a fixed order, resulting in a counterbalanced assignment of images to VISIBILITY conditions.

4.3 Measures

4.3.1 Word Predictability. Imagine typing a given sentence while a predictive text system offers suggestions. Sometimes a word to be entered will appear as one of the three suggestions before even its first letter is entered; we refer to such words as *predictable*. We refer to all other words as *unpredictable* (even if it is later suggested after more letters are entered). The (un)predictability of a word is a property of the *text*, not the manner in what that text was *entered*: we count a word as predictable even if the suggestions were *disabled* at the point that it was actually entered. This contrast is crucial to be able to compare the content written between different types of suggestion systems.

Texts differ in predictability: on one extreme are texts that use only suggested words, on the other would be texts that are generated by *avoiding* initially-suggested words. This observation motivates our primary measure.

Our primary measure is the number of *predictable* words. Since the length of text written could differ between conditions, we also measure the total length of captions in words and the number of

words that were *not* predictable words (predictable + unpredictable = total length). To simplify analysis, we stripped all punctuation except for mid-word. (Almost all captions written were a single sentence, so punctuation was not needed to separate sentences.) These measures allow us to answer RQ1–RQ3.

We also measured the *uncorrected error rate* as the number of low-level errors (typos, misspelling, etc.) that were present in the caption that the writer submitted. (Since our primary interest was how system design affected *content*, we ignored errors that the writer corrected before submission.) Since most captions had zero or one typos, we simplified our analysis to consider only whether or not a submitted caption included a typo.

To label typos, one of the authors inspected all writing, blind to condition, with the help of the Microsoft Word contextual spelling checker, and corrected typos and spelling mistakes including mismatched articles ('a' vs 'an'). Grammatical and factual errors, which occurred rarely, were left uncorrected. Any caption that was corrected in this way was labeled as having a typo.

Since typos would artificially reduce the number of predictable words, leading to inflated estimates of content effects, we computed that measure on typo-corrected text also.

4.3.2 Process Measures. To answer RQ4, we used logged data to compute *typing speed*. We compute speed by dividing the final text length in characters (including spaces) by the interval between the first and last input action on the typing screen.

We used the participant's mean typing speed in the Never condition as a baseline to control for individual differences (which could stem from prior touchscreen typing experience, effort invested, device characteristics, and many other factors). Our main measure was the *ratio* of the participant's typing speed to this baseline speed.

4.3.3 Subjective Measures. We collected both block-level and overall subjective measures. Surveys after each keyboard block collected task load data using all six NASA TLX items on a 7-point scale [23]. We analyze the sum of these measures, but also individually examine the "physical" and "mental" load items, as has been done in prior work [37].

The final survey asked participants to pick which of the three keyboard designs they experienced were "most helpful" for three goals: accuracy, specificity, and speed. Keyboard designs were indicated by number, and participants could see all of their captions for reference. We analyzed the total number of times that participants picked each keyboard design.

4.4 Analysis

We applied statistical estimation methods for our primary outcomes [17]. Except where indicated, we estimated means and confidence intervals by non-parametric bootstrapping. Since we expected substantial individual differences, we bootstrapped grouped by participant: Each of the 10,000 bootstrap iterations resampled participants with replacement; we used the complete data for each participant chosen.

Since we expected substantial variance across both participants and images for all measures, we used lme4 [7] to estimate linear mixed-effects models at each bootstrap iteration with both participant and image as random effects. (The random effects structure

mitigates the pseudoreplication that would otherwise occur from analyzing trial-level data.) We report the bootsrapped estimates of the means and pairwise contrasts for the Visibility fixed effect.[1]

4.5 Procedure

4.5.1 Images. We used 12 images selected from the Microsoft COCO (Common Objects in Context) dataset [31]. Most images showed people doing outdoor activities (surfing, flying kites, etc.), or familiar scenes such as a train station or a bus on a street. Our selection process was motivated by potential use in a different (unpublished) experiment. We found the twelve pairs of images in the validation set of 2014 COCO release where the two images had the most similar captions. We defined similarity as the tf-idf similarity of unigrams in the concatenation of all five of the captions that crowd workers had originally entered for each image. We randomly picked one image from each pair to be a prompt for caption writing.

4.5.2 Predictive Keyboard. We implemented a custom touchscreen keyboard modeled on commercial keyboards but where we could manipulate the content and visibility of the suggestions. Compared with commercial keyboards, our keyboard was simplified in several ways; the instructions explicitly pointed out the first three:

- the keyboard had a single layer (lowercase only, minimal symbols, and no numbers)
- no ability to edit past text except for backspacing and retyping (and delete key did not automatically repeat), so editing was more cumbersome than people may have been used to
- no auto-correct (the instructions encouraged participants to manually correct typos)
- no automatic insertion of suggestions or corrections; ignoring the suggestion bar produced the same results as if it were not present
- no key target resizing; the mapping from screen location to key was fixed

The UI showed word predictions in the familiar "suggestion bar" interface used in contemporary mobile phone keyboards [4, 8, 37]. When the writer entered a partial word, the suggestions offered completions of that word, otherwise the suggestions showed likely next words. The writer could choose to tap a suggestion, tap a key, or tap the backspace key (which deleted a single letter at a time). The system updated the suggestions after each user action.

Figure 1 shows the task as it appeared on a participant's device, including the image, caption written so far, and suggestions offered by the system. The screen layout ensured that the participant's complete writing was always fully visible and visually close to the keyboard and suggestions (if applicable); participants may have had to scroll to see the complete image.

The keyboard showed the top 3 most likely predictions from the language model as suggestions, subject to the constraint that if the cursor was in the middle of a word, all predictions must have the characters typed so far as a prefix.

Our keyboard generated predictions using an LSTM language model using OpenNMT [28], trained on image captions. For this study we did *not* give the system access to visual features from the

---

[1]The overall analysis approach was planned and clearly indicated content effects of predictive suggestions, but the analyses reported here reflect refinements and simplifications performed after seeing the initial results.

image being captioned (i.e., the system offered the same predictions regardless of image). Models ran on a cloud VM, providing predictions to the client with a typical latency of under 300ms from tap to prediction visibility.

The language model was a single-layer LSTM, with hidden state dimension of 2048.[2] The model was trained on the COCO training set captions using the Adam optimizer with the "Noam" learning rate schedule [44], with a base learning rate of 2, 8000 warm-up steps, $\beta_2 = 0.998$, and parameters initialized using the Xavier uniform scheme [21]. The batch size was 128. If the norm of the gradient for any batch exceeded 2, it was re-normalized to have a norm of 2. After 10 epochs, the model achieved a perplexity of 16.32 and a top-1 accuracy of 46.00%.

We constructed the OnlyConfident system by modifying the Always system to hide all three suggestions when the predicted likelihood of the words was less than a threshold. We chose the thresholding method and value by generating predictions at 1000 randomly chosen beginning-of-word locations in the COCO validation set and logging whether the word that followed was one of the three predicted. We considered thresholding based on the maximum, mean, or minimum likelihood of each of the three predictions, and chose to use the maximum because it obtained the highest AUC. We then chose the threshold value that would have resulted in suggestions being displayed 50% of the time. At this threshold value, the false positive rate was 25.7%. When the maximum confidence dropped below the threshold, the keyboard showed a blank suggestion bar.

4.5.3 Participants. The study was carried out remotely as a mobile web application that participants accessed using their own touchscreen devices.[3] We recruited 111 participants (61 male, ages 19–61) from Amazon Mechanical Turk. Participants received $5 for completing the study. Since the experiment required a low-latency connection to our US-based server, we limited participants to those in the US and Canada. We required participants to have a 99% approval rating on at least 1000 HITs. The study was conducted in English; all participants reported "native" or "fluent" English proficiency.

The landing page described the study as using various mobile phone keyboards to type descriptions of images, with an expected time of about 30 minutes. After a statement of informed consent, participants read a description of the task, which promised a $0.50 bonus for the most specific and accurate captions. They then read a brief overview of the flow of the experiment, which emphasized that they would be using 3 different keyboard designs and they should attempt to remember their experiences with each.

Before any of the writing tasks, participants completed a task tutorial with the overall instruction to write the most specific and accurate caption they could for each image. The tutorial included examples of captions that differed in specificity and accuracy. Some pilot participants seemed to think that we simply meant for them to write *long* captions, so we revised the instructions to encourage writers to be concise. Examples were provided, based on different images than those used in the experiment. We did not prevent

writers from writing multiple sentences, but all examples provided were a single sentence (as were most captions that participants wrote).

Each participant wrote captions for twelve images. The body of the experiment consisted of three blocks, one for each condition (which we referred to as "keyboard design"). Each block began with a page prominently displaying the number of the keyboard design they were about to use (e.g., "Keyboard Design 3"). Next, participants completed a "practice round" with that keyboard, in which they were given a sentence to transcribe (a caption written for an image, not shown, that was not one of the images to caption). If they did not use suggestions, they were encouraged to complete the transcription task again, in case they had been too fixated on the text to transcribe that they failed to notice the suggestions. Then they typed captions for 4 images, followed by a survey about their experience with that keyboard. We chose to keep the same keyboard design within each block of trials so that participants could become accustomed to the behavior of each keyboard. The experiment closed with a survey asking for comparisons between their experiences with each of the three keyboard designs, as well as demographics (all questions optional).

The experiment enforced that participants typed at least one word before a caption could be marked as completed, but otherwise no restrictions were enforced on the length or time taken for writing captions. Notably, we did not require participants to use suggestions while writing their captions.

We excluded two participants who visibly violated our instructions to write captions that were specific and accurate. Both wrote captions that averaged less than 5 words, such as "there is teniss" and "people flying kites." Other than those written by these participants, all captions seemed generally appropriate and grammatical.

All participants used a suggestion at least once when typing captions, and no participant accepted every suggestion, so we did not need to exclude participants based on those criteria.

5 RESULTS

We collected a total of 1308 captions (109 participants after exclusion; each wrote captions for 12 images).
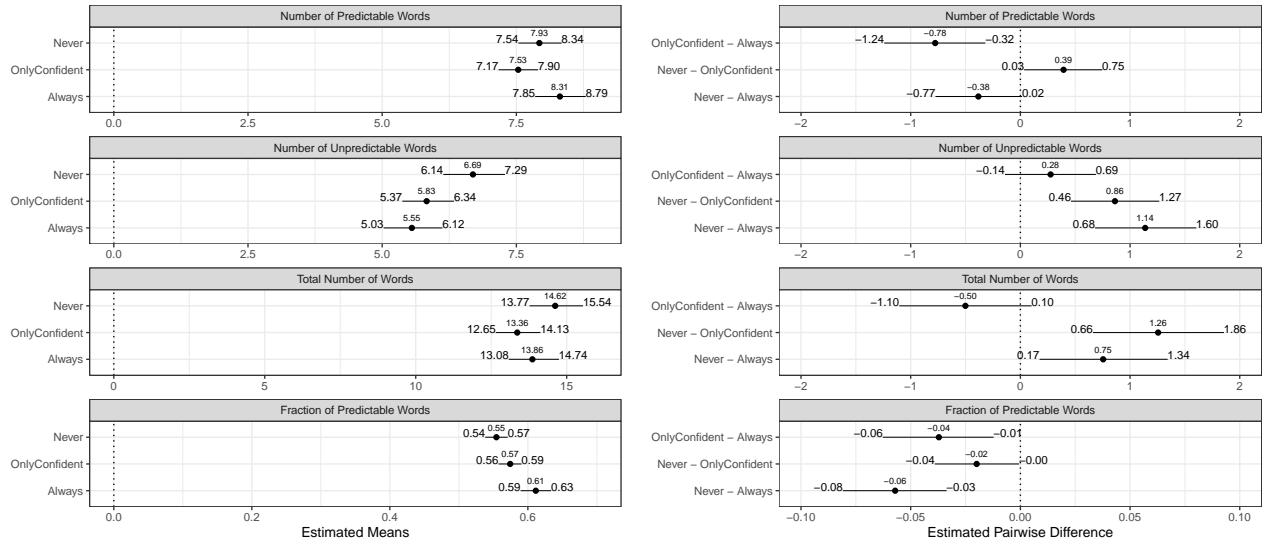
5.1 Content Effects

Predictability. Figure 2 shows the estimated means and pairwise comparisons between suggestion Visibility conditions for the main content measures. The strongest contrast that emerged was that an average of about one additional *unpredictable* word was used when suggestions were Never visible compared to the Always (CI: [0.68, 1.60]) or OnlyConfident (CI: [0.46, 1.27]) conditions.[4] The data also indicate (albeit less clearly) that captions written in Always had around 0.78 (CI: [0.32, 1.24]) more *predictable* words than OnlyConfident.

Figure 2 also shows two measures derived from the above measures, length and fraction predictable, which convey no new statistical information but may be useful for interpretation. Captions written with Never-visible suggestions were longer (14.6 words) than those written in the other two conditions (Always: 13.9 words,

---

[2]For historical reasons, we actually used a "sequence-to-sequence" model but with the input set to a constant token; this does not affect our results.
[3]The study procedure was approved by our institutional review board.

[4]A pairwise difference that is statistically significant at the $\alpha = 0.05$ level (in a null-hypothesis test setting) will have a 95% CI that does not contain 0.

Figure 2: Estimated means (left) and pairwise differences between VISIBILITY levels (right) of predictability measures for captions written. Error bars show 95% confidence intervals computed by non-parametric bootstrap by participant. Note that the visualization contains redundancy, e.g., the bottom two plots can be computed from the top two.

ONLYCONFIDENT: 13.4 words), with a clear difference of about 1.26 (CI: [0.66, 1.86]) more words with NEVER than with ONLYCONFIDENT suggestions. The difference between NEVER and ALWAYS was in the same direction but did not seem to be as strong (.76, CI: [0.18, 1.36]), and there did not seem to be a substantial difference in caption length between ONLYCONFIDENT and ALWAYS. The fraction of predictable words was about 6% (CI: [3%, 8%]) higher for ALWAYS-visible than NEVER-visible suggestions and about 4% (CI: [1%, 6%]) for ONLYCONFIDENT than NEVER.

Typos. Suggestions seemed to reduce the number of typos that participants left uncorrected in their captions. Of the 124 captions that had typos, 73 (59%) were written with NEVER suggestion visibility, 27 (22%) with ONLYCONFIDENT, and 24 (19%) with ALWAYS. Comparing the two conditions with suggestions visible (ALWAYS and ONLYCONFIDENT) jointly against the NEVER condition, Fisher's Exact Test found that the odds ratio for a caption having a typo was 0.31 (CI: [.21, .45]) in favor of fewer typos for suggestion conditions.

5.2 Process Effects

We found that baseline typing rate was a strong predictor of the ratio between typing speed with suggestions (either ALWAYS or ON-LYCONFIDENT) and baseline typing speed[5]. We used a linear mixed model to predict the block-wise mean ratio of speed to baseline speed: speed ratio to baseline $= a \times$ baseline speed $+ b + \epsilon_{\text{participant}}$, where ϵ represents the participant-level random effect. The 95% confidence interval for b was [1.35, 1.66], indicating that suggestions increased typing speed overall. But the 95% confidence interval for a was [-0.29, -0.14], indicating that as baseline speed increased, the benefit of suggestions decreased. As Figure 3 shows, some of the

fastest typists in our experiment wrote slower when suggestions were visible, but since our participants included few such typists, we lack evidence to determine whether suggestions would slow down fast typists in general. The figure also shows that we did not observe a significant difference between the two levels of suggestion VISIBILITY (ALWAYS and ONLYCONFIDENT) in terms of speed. To quantify this observation, we fit a separate model including a term for VISIBILITY; the confidence intervals for both VISIBILITY ([-0.08, 0.08]) and its interaction with baseline speed ([-0.05, 0.03] were nearly symmetric around 0.
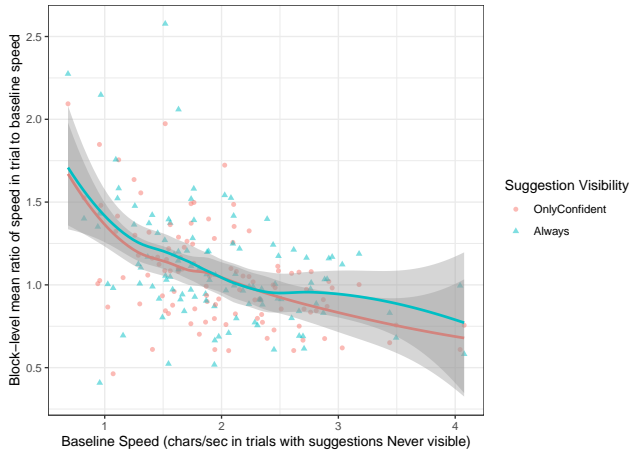
5.3 Subjective Experience

Ranking results from the closing survey suggest that participants strongly preferred visible suggestions over NEVER and generally preferred ALWAYS over ONLYCONFIDENT visibility. Participants picked the ALWAYS condition as most helpful 206 times, ONLYCONFIDENT condition 101 times, and NEVER condition 20 times. A χ^2 goodness-of-fit test finds that this result would be highly unexpected under the null hypothesis that all three VISIBILITY conditions are equally helpful ($\chi^2_2 = 159.6$, $p < .0001$).
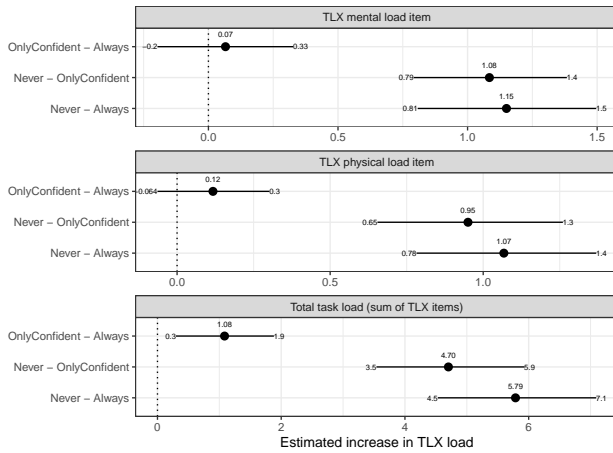
When suggestions were hidden (VISIBILITY=NEVER), participants reported higher task load overall as well as for both the physical and mental effort items individually. Figure 4 shows that the pairwise difference was approximately 1 point on a 7-point scale for both the physical and mental items, for a difference of about 5.5 points overall.

6 SUPPLEMENTAL ANALYSES

Since we observed that captions written with suggestions were *shorter* than those written without suggestions, we conducted supplemental analysis to explore potential explanations for this result.

---

[5]Since NEVER forms the baseline, analyses in this paragraph consider only ALWAYS and ONLYCONFIDENT. Analyses in this paragraph use 1000 iterations of parametric bootstrapping.

Figure 3: Predictive suggestions had a positive effect on typing speed overall, but with less benefit for faster typists. Scatterplot points show block-level average speed ratios, solid lines show loess curves for each Visibility condition, and shaded areas show approximate confidence intervals for each curve.
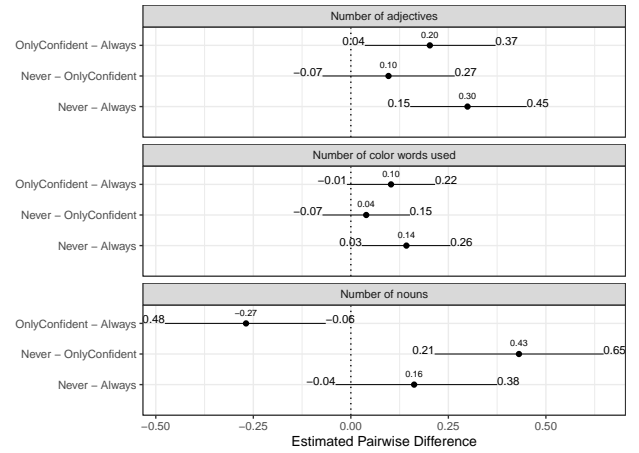

Figure 4: Bootstrap estimates of pairwise differences in task load (overall, physical, and mental)

Since the analyses in this section were conceptualized after seeing the data, they should be treated as exploratory.

6.1 What types of words were predictable?

Table 1 gives examples of captions at different levels of predictability. Inspection of examples like those shown suggested that the difference in the fraction of predictable words might express itself in terms of a difference in use of words of different parts of speech. Of particular interest seemed to be nouns and adjectives. Also, since we noticed that descriptions of *color* were sometimes missing in high-predictability image captions, we looked at the number of

Figure 5: Exploratory analysis suggested that the Always-visible suggestions may lead to the use of fewer adjectives, especially color adjectives, and that OnlyConfident visibility resulted in fewer nouns.

color adjectives used. Figure 5 shows that suggestions may have resulted in fewer adjectives.

6.2 Why were captions shorter with suggestions?

We conjecture that the suggestions offered may nudge the writer to skip a word. For example, suppose someone is typing typing "a tennis player is swinging his racket on a green tennis court". As they are about to type "green," the system instead suggests "tennis," encouraging the writer to skip "green." To describe this scenario we will say that "green" was *skip-nudged*: one of the words suggested at the beginning of a word matched the *following* word.

We analyzed the 436 texts written in the Never condition, thus not affected by suggestions, to identify potential skip-nudges. Of these texts, 299 (69%) had at least one skip-nudge. There were a total of 488 skip-nudges, 202 (41%) of which were predictable (i.e., the skip-nudged word was *also* one of the predictions). (If we consider only those suggestions that would be still presented in OnlyConfident, there are only 228 skip-nudges, of which 120 (53%) are predictable.) The top 10 *predictable* skip-nudged words were: a, wedding, tennis, is, to, tree, at, train, of, baseball; the top 10 *unpredictable* skip-nudged words were: red, white, desktop, is, sits, sitting, computer, on, small, bathroom.

7 DISCUSSION

Captions that people wrote when presented with predictive suggestions differed from what they wrote without suggestions. The differences that were most clearly supported by our data are:

(1) captions written with suggestions visible were *shorter* and used fewer words that were *unpredictable*, both by a magnitude of about 1 word, than when suggestions were not visible (RQ1, RQ2),

image	P	U	%	corrected text
	11	11	50	people are on sand with five kites in the air as a man in a red shirt and two children hold kites
	7	3	70	a family of three on the beach flying kites together
	7	9	44	an old brown train pulling away from a small train station by a baby blue building
	5	3	62	a train pulling into a quaint train station
	11	12	48	a man in a black shirt with the number eight and grey pants swinging a baseball bat with many players in the background
	11	5	69	a baseball player with the number eight jersey has just hit the ball with the bat

Table 1: Examples of captions with varying percentages of predictable words (%). P = number of predictable words, U = number of unpredictable words. (The captions for each image were sorted by percentage predictable and an example was taken at the first quartile and third quartile for each image.) Image credits: https://flic.kr/p/GyXLw, https://flic.kr/p/fkybX6, https://flic.kr/p/6Yxc61

(2) captions written with low-confidence suggestions hidden had fewer *predictable* words than those written with suggestions were always shown (RQ3), and

(3) predictive suggestions had a positive effect on typing speed overall, but with decreasing benefit for faster typists (RQ4).

Supplemental analysis enables us to conjecture a two-part explanation for these observations. However, further study is needed to determine whether these explanations are accurate and sufficient.

First, suggestions may have sometimes encouraged people to *skip* a word that they would have entered. Since our analysis found that both predictable and unpredictable words could be skip-nudged at similar rates, this encouragement would lead to reduced numbers of both unpredictable and predictable words, resulting in shorter captions overall.

Second, perhaps people would have entered an unpredictable word but the appearance of a prediction caused them to *substitute* a predictable word instead (see, e.g., the caption of Figure 1). This substitution would increase the number of predictable words and reduce the number of unpredictable words by the same amount, so length would be unaffected.

Together, skipping and substitution imply that the number of unpredictable words would be reduced, which could account for the first observed difference.

Confidence thresholding reduced the number of times that predictable words were suggested, thus reducing the likelihood of substitution. This account could explain the difference in predictable word count between the two conditions where suggestions were shown.

Our speed findings agree with the AAC literature (surveyed in [43]) that predictions often improve communication rate but with substantial individual differences [29].

Writers overall preferred conditions where suggestions were always available (as indicated by lower task load and explicit preference rankings). However, the finding that captions entered using suggestions tended to be shorter suggests that minimizing physical effort does not fully account for the differences in word choice that we observed. If participants were simply minimizing their physical effort, the captions entered with Never-visible suggestions would have been shortest, since that condition requires participants to type each character. Other participants typed shorter captions for

the same images in conditions where suggestions were available, which indicates that an easier-to-enter utterance was available and acceptable. This finding underscores that the *content* of the suggestions influences text content.

7.1 Limitations

Several limitations of our study lead us to urge caution against overgeneralizing its results: we do not claim that commercially deployed predictive systems have the kind and degree of content effects that we found in our study. However, we conjecture that they do already influence content and that this influence will grow as prediction generation and interaction technology improves. We urge follow-up study of deployed systems to evaluate these content effects.

Experimenter Demand Effects. Even though the instructions and recruitment never mentioned predictions (or synonyms such as suggestions or recommendations), the design of this experiment was vulnerable to experimenter demand effects in other ways. For example, the opening survey asked about participants' ordinary use of the suggestion bar, the consent form indicated the purpose of the research, and the suggestions constituted the main and salient difference between experiment blocks, which indicates to participants that their use is interesting to the researcher [50]. Moreover, if the participant did not use any suggestions whatsoever, even completely unambiguous completions of a word, during a practice transcription task in which relevant suggestions were available, the system encouraged them to repeat the practice round and use the suggestions; this intervention may have created a carry-over demand effect in the captioning tasks. This happened for 48 participants, many of whom reported that they use the suggestion bar on their own phones "often" or "almost always". So we suspect that participants did not use suggestions during practice rounds for more mundane reasons specific to the transcription task, such as having to switch attention between the text to transcribe, the text written so far, a potentially unfamiliar keyboard, and the suggestions offered.

Our findings are about the *effects* of suggestion use, not the *degree* to which they are used, so the presence of demand effects does not challenge the validity of our conclusions.

Generalization to other writing tasks. While the task we used was more representative of real-world writing tasks than transcription tasks used in most writing studies, captioning is still not a common task. We would expect our findings to generalize to other tasks where the main goal is describing concrete things (e.g., video description, reviewing of products and services, or describing real estate). But our findings may not generalize to other types of tasks, such as those involving conceptual exposition or persuasion, or even to writing descriptions longer than a sentence. Our findings may also be influenced by the specific prompt we provided, which asked participants to write captions that were "specific," "accurate," and "concise." Finally, participants wrote as part of a paid task on MTurk; predictive text could have different effects on writing by other groups of people or for different objectives.

Generalization to other predictive text systems. The predictive keyboard that we used in our experiments differed from commonly deployed predictive keyboards in two ways that may affect the generalizability of our findings. First, the keyboard did not offer automatic corrections of mistyped words. The lack of corrections may have caused writers to increase their propensity to consider suggestions because entering a word without using completion suggestions incurs the greater cost of potentially having to backspace and correct a typo. (On the other hand, writers may have also needed to pay more attention to the text that they have just entered, rather than looking at suggestions, which would decrease their propensity to consider suggestions.) Second, our interface did not allow participants to edit past words without backspacing over every character in between, so writers may have typed more carefully.

The suggestion generation system may also affect generalizablitiy, since its suggestions were very strongly adapted to the domain of image captioning. As such, our findings could be viewed as a peek into the future: as predictive text systems gain access to more contextual data (e.g., Google's Smart Compose [12] uses context from the writer and current email thread), they will likely be able to make predictions that are even more strongly adapted to the task (and also to the writer) than ours were.

Experience with System. Participants wrote only four captions (plus one practice) with each system. Writers may behave differently after more exposure to a predictive system; if that exposure leads them to trust the system more, the effects of the system on the content of their writing may be larger than what our short study observed.

8 CONCLUSION

Predictive text systems help many people write more efficiently, but by their nature these systems only make certain content efficient to enter. Our study found that writers are sensitive to these differences: when presented with predictive text suggestions, people wrote shorter and more predictable language. In short, predictive text suggestions—even when presented as single words—are taken as suggestions of what to write.

Our findings underscore the general call that evaluations of intelligent interactive systems be based on authentic tasks [9], and specifically the call of Kristensson and Veranen in 2014 that text

entry studies should include composition tasks [30]. We further request that text entry studies study *content effects* of their systems and have sufficient statistical power to notice effects of comparable sizes to those we reported here.

8.1 Implications for Deployed Systems

The content that people write using predictive systems will become part of the corpora used to train language models used by future predictive systems. Even a small bias in word choice could have a feedback effect.

Assistance through error avoidance [6, 34], correction [8, 45], and disambiguation [40] may better preserve writer autonomy than word or phrase suggestion. These systems do still make certain texts easier to enter than others (e.g., it becomes more difficult to enter creative misspellings or made-up words), but the system's biases are less salient, so we expect that they would impact writing content less.

Platforms providing predictive text functionality, such as Smart Compose [12], should be accountable for the effects that their systems have on writing content.

8.2 Future Work

Future work could further characterize how predictive typing affects writing content, such as by using even more sensitive measures to study other tasks (such as those involving persuasion), languages, and suggestion interaction designs. Future work should also explore individual differences in how suggestions affect people: both situational affect [20] and stable traits [19] have been shown to modulate how people use predictive systems.

Future work could also explore ways in which suggestions may be designed to have desirable effects on content. For example, predictive scaffolding could be used to help second-language writers write more fluently and naturally. Could suggestions be designed to help writers come up with ideas or express those ideas creatively? Initial studies have yielded mixed results [14], but the challenge is promising.

ONLINE APPENDIX

Data and analysis code is available at https://osf.io/w7zpa/.

ACKNOWLEDGMENTS

REFERENCES

[1] Ohoud Alharbi, Ahmed Sabbir Arif, Wolfgang Stuerzlinger, Mark D. Dunlop, and Andreas Komninos. 2019. WiseType: A Tablet Keyboard with Color-Coded Visualization and Various Editing Options for Error Correction. In *Proceedings of Graphics Interface 2019 (GI 2019)*. Canadian Information Processing Society, 10. https://doi.org/10.20380/GI2019.04

[2] Ofra Amir, Barbara J. Grosz, Krzysztof Z. Gajos, and Limor Gultchin. 2019. Personalized change awareness: Reducing information overload in loosely-coupled teamwork. *Artificial Intelligence* 275 (2019), 204–233. https://doi.org/10.1016/j.artint.2019.05.005

[3] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2018. Sentiment Bias in Predictive Text Recommendations Results in Biased Writing. In *Graphics*

Interface 2018. Toronto, Ontario, Canada, 8–11. http://graphicsinterface.org/wp-content/uploads/gi2018-7.pdf

[4] Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16).* Association for Computing Machinery, New York, NY, USA, 603âĂŞ608. https://doi.org/10.1145/2984511.2984584

[5] Tamara Babaian, Barbara J Grosz, and Stuart M Shieber. 2002. A writer's collaborative assistant. In *Proceedings of the 7th international conference on Intelligent user interfaces - IUI '02.* ACM Press, New York, New York, USA, 7. https://doi.org/10.1145/502716.502722

[6] Tyler Baldwin and Joyce Chai. 2012. Towards online adaptation and personalization of key-target resizing for mobile devices. (2012), 11. https://doi.org/10.1145/2166966.2166969

[7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015). https://doi.org/10.18637/jss.v067.i01

[8] Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2014. Both complete and correct? Multi-Objective Optimization of Touchscreen Keyboard. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14.* ACM Press, New York, New York, USA, 2297–2306. https://doi.org/10.1145/2556288.2557414

[9] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20).* ACM, New York, NY, USA.

[10] Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* Association for Computing Machinery, New York, NY, USA, Article Paper 255, 14 pages. https://doi.org/10.1145/3173574.3173829

[11] Eric T. Chancey, James P. Bliss, Yusuke Yamani, and Holly A.H. Handley. 2017. Trust and the Compliance-Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence. *Human Factors* 59, 3 (2017), 333–345. https://doi.org/10.1177/0018720816682648

[12] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, and et al. 2019. Gmail Smart Compose: Real-Time Assisted Writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19).* Association for Computing Machinery, New York, NY, USA, 2287–2295. https://doi.org/10.1145/3292500.3330723

[13] N. Ann Chenoweth and John R. Hayes. 2003. The Inner Voice in Writing. *Written Communication* 20, 1 (jan 2003), 99–118. https://doi.org/10.1177/0741088303253572

[14] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces (IUI '18).* Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[15] John J Darragh, Ian H Witten, and Mark L James. 1990. The Reactive Keyboard: A Predictive Typing Aid. *Computer* 23, 11 (1990), 41–49.

[16] Paul Deane, Nora Odendahl, Thomas Quinlan, Mary Fowles, Cyndi Welsh, and Jennifer Bivens-tatum. 2008. Cognitive Models of Writing : Writing Proficiency as a Complex Integrated Skill. *Language* October (2008), 128. https://doi.org/ETSRR-08-55

[17] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI.* Springer, 291 – 330. https://doi.org/10.1007/978-3-319-26633-6_13

[18] A. Fiannaca, A. Paradiso, M. Shah, and M.R. Morris. 2017. AACrobat: Using mobile devices to lower communication barriers and provide autonomy with Gaze-based AAC. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2017). https://doi.org/10.1145/2998181.2998215

[19] Sara Garver, Caroline Harriott, Krysta Chauncey, and Meredith Cunha. 2017. Co-adaptive Relationships with Creative Tasks. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17* (2017), 123–124. https://doi.org/10.1145/3029798.3038357

[20] Surjya Ghosh, Kaustubh Hiware, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Does emotion influence the use of auto-suggest during smartphone typing?. In *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19.* ACM Press, New York, New York, USA, 144–149. https://doi.org/10.1145/3301275.3302329

[21] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. *Pmlr* 9 (2010), 249–256. https://doi.org/10.1.1.207.2059 arXiv:arXiv:1011.1669v3

[22] Noah D Goodman and Michael C Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences* 20, 11 (nov 2016), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

[23] Sandra G Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (oct 2006), 904–908. https://doi.org/10.1177/154193120605000909

[24] John R Hayes and N Ann Chenoweth. 2006. Is Working Memory Involved in the Transcribing and Editing of Texts? *Written Communication* 23, 2 (2006), 135–149. https://doi.org/10.1177/0741088306286283

[25] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99* (1999), 159–166. https://doi.org/10.1145/302979.303030

[26] Julie S. Hui, Darren Gergle, and Elizabeth M. Gerber. 2018. IntroAssist. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (2018), 1–13. https://doi.org/10.1145/3173574.3173596

[27] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *KDD.* https://doi.org/10.475/123 arXiv:1606.04870

[28] Guillaume Klein, Yoon Kim, Yuntian Deng, Josep Crego, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source Toolkit for Neural Machine Translation. (2017). https://doi.org/10.18653/v1/P17-4012 arXiv:1709.03815

[29] Heidi Horstmann Koester and Simon P. Levine. 1994. Modeling the Speed of Text Entry with a Word Prediction Interface. *IEEE Transactions on Rehabilitation Engineering* 2, 3 (1994), 177–187. https://doi.org/10.1109/86.331567

[30] Per Ola Kristensson and Keith Vertanen. 2014. The Inviscid Text Entry Rate and its Application as a Grand Goal for Mobile Text Entry. *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services - MobileHCI '14* (2014), 335–338. https://doi.org/10.1145/2628363.2628405

[31] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8693 LNCS, PART 5 (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48 arXiv:1405.0312

[32] Peter A. McCormick. 1997. Orienting attention without awareness. *Journal of experimental psychology. Human perception and performance* 23, 1 (1997), 168–180. https://doi.org/10.1037/0096-1523.23.1.168

[33] Joachim Meyer. 2004. Conceptual issues in the study of dynamic hazard warnings. *Human Factors* 46, 2 (2004), 196–204. https://doi.org/10.1518/hfes.46.2.196.37335

[34] Martez E. Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. 2017. Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* (2017), 2558–2570. https://doi.org/10.1145/3025453.3025517

[35] Naho Orita, Eliana Vornov, Naomi Feldman, and Hal Daumé III. 2015. Why discourse affects speakers' choice of referring expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Association for Computational Linguistics, Stroudsburg, PA, USA, 1639–1649. https://doi.org/10.3115/v1/P15-1158

[36] Ondrej Polacek, Adam J. Sporka, and Brandon Butler. 2013. Improving the methodology of text entry experiments. *4th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2013 - Proceedings* (2013), 155–160. https://doi.org/10.1109/CogInfoCom.2013.6719232

[37] Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 83–88. https://doi.org/10.1145/2858036.2858305

[38] Jeroen G W Raaijmakers and Emoke Jakab. 2013. Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language* 68, 2 (2013), 98–122. https://doi.org/10.1016/j.jml.2012.10.002

[39] Robert Rauschenberger. 2003. Attentional capture by auto- and allo-cues. *Psychonomic Bulletin & Review* 10, 4 (2003), 814–842.

[40] Shyam Reyal, Shumin Zhai, and Per Ola Kristensson. 2015. Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (2015), 679–688. https://doi.org/10.1145/2702123.2702597

[41] Jaime Teevan, Harmanpreet Kaur, Alex C. Williams, Shamsi T. Iqbal, Anne Loomis Thompson, and Walter S. Lasecki. 2018. Creating Better Action Plans for Writing Tasks via Vocabulary-Based Planning. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22. https://doi.org/10.1145/3274355

[42] Mark Torrance and David Galbraith. 2006. The processing demands of writing. *Handbook of Writing Research* (2006), 67–82.

[43] Keith Trnka, John McCaw, Debra Yarrington, Kathleen F. McCoy, and Christopher Pennington. 2009. User Interaction with Word Prediction: The Effects of Prediction Quality. *ACM Trans. Access. Comput.* 1, 3, Article Article 17 (Feb. 2009), 34 pages. https://doi.org/10.1145/1497302.1497307

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Nips* (2017). https://doi.org/10.1017/S0140525X16001837 arXiv:1706.03762

[45] Keith Vertanen, Crystal Fletcher, Dylan Gaines, Jacob Gould, and Per Ola Kristensson. 2018. The Impact of Word, Multiple Word, and Sentence Input on Virtual Keyboard Decoding Performance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 626, 12 pages. https://doi.org/10.1145/3173574.3174200

[46] Keith Vertanen, Haythem Memmi, Justin Emge, Shyam Reyal, and Per Ola Kristensson. 2015. VelociTap: Investigating Fast Mobile Text Entry using Sentence-Based Decoding of Touchscreen Keyboard Input. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. 659–668. https://doi.org/10.1145/2702123.2702135

[47] David J Ward, Alan F Blackwell, and David J C MacKay. 2000. Dasher - A Data Entry Interface Using Continuous Gestures and Language Models. *Proceedings*

of the 13th annual ACM symposium on User interface software and technology 2 (2000), 129–137. https://doi.org/10.1145/354401.354427

[48] Christopher D. Wickens, Stephen R. Dixon, and Nicholas Johnson. 2006. Imperfect diagnostic automation: An experimental examination of priorities and threshold setting. *Proceedings of the Human Factors and Ergonomics Society* October 2006 (2006), 210–214. https://doi.org/10.1177/154193120605000301

[49] Steven Yantis and John Jonides. 1984. Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance* 10, 5 (1984), 601–621. https://doi.org/10.1037/0096-1523.10.5.601

[50] Daniel John Zizzo. 2010. Experimenter demand effects in economic experiments. *Experimental Economics* 13, 1 (2010), 75–98. https://doi.org/10.1007/s10683-009-9230-z