On Suggesting Phrases vs. Predicting Words for Mobile Text Composition

Kenneth C. Arnold[1] **Krzysztof Z. Gajos**[1] **Adam T. Kalai**[2]

[1]Harvard SEAS
Cambridge, MA USA
{kcarnold,kgajos}@seas.harvard.edu

[2]Microsoft Research
Cambridge, MA USA
adam.kalai@microsoft.com

ABSTRACT

A system capable of suggesting multi-word phrases while someone is writing could supply ideas about content and phrasing and allow those ideas to be inserted efficiently. Meanwhile, statistical language modeling has provided various approaches to predicting phrases that users type. We introduce a simple extension to the familiar mobile keyboard suggestion interface that presents phrase suggestions that can be accepted by a repeated-tap gesture. In an extended composition task, we found that phrases were interpreted as suggestions that affected the content of what participants wrote more than conventional single-word suggestions, which were interpreted as predictions. We highlight a design challenge: how can a phrase suggestion system make valuable suggestions rather than just accurate predictions?

Author Keywords

phrase suggestions; mobile text composition

ACM Classification Keywords

H.5.2 Information Interfaces: User Interfaces; I.2.7 Natural Language Processing

INTRODUCTION

Most mobile keyboards include a *suggestion bar* that offers word completion, correction, and even prediction. While the suggestion bar plays a central role in commercial products used daily by billions of people, it has received limited attention from academic research [5, 28]. Today's suggestion bars generally offer single words. We consider the natural extension to multi-word phrases. Statistical language modeling has been shown to be capable of accurately predicting phrases [26], sentences [7] and even entire short messages [18], but it is not clear how to present multi-word suggestions for mobile text entry, or how people might interact with them.

Text *composition* involves both thinking and entering text [21]. So suggestions, especially of phrases, have the potential to affect the *content* of what people ultimately choose to write.

UIST 2016, October 16–19, 2016, Tokyo, Japan.

ACM ISBN 978-1-4503-4189-9/16/10 ...$15.00.
http://dx.doi.org/10.1145/2984511.2984584

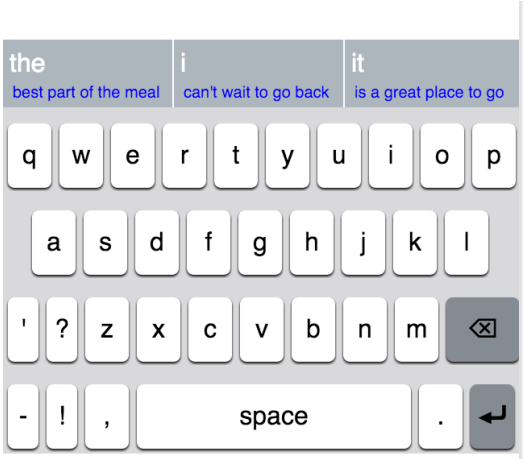

Figure 1. We augment the standard mobile keyboard suggestion bar (offering buttons containing individual words such as *The*, *I*, and *it*) by previewing the coming words in small blue text. A user inserts words by tapping repeatedly on the button, one tap per word. For each button press, the suggestion advances, e.g., if the rightmost button is quintuple-tapped, the text *it is a great place* is inserted, the word *to* is suggested, and the preview text advances to become *go for a romantic*.

This raises two challenges: first, a user interface challenge in suggesting phrases on a mobile device in a way that is at least as useful as single-word suggestions that have already proven valuable; and second, a statistical language model challenge of generating useful suggestions from context, which may require techniques other than standard prediction. This note addresses this first challenge.

We conducted a within-subjects study comparing a simple phrase suggestion interface (Figure 1) with a nearly identical system in which only a single word was suggested in each of the three suggestion boxes. Participants were given the task of writing reviews for restaurants of their choice. We found that participants accepted significantly more words with our phrase interface and, consequently, that the suggestions shaped what they were writing.

We found that phrases are interpreted as *suggestions of what to say and how to say it* more than the single word displays, which were viewed as *predictions of what will be typed*. Most participants commented that the system influenced the content of what they were writing. Some felt that the suggestions were helpful in writing better-worded and more comprehensive

reviews, while others felt that the suggestions were trite and generic.

The contributions of this paper are: (a) an extension of the mobile keyboard suggestion bar to offer phrases, (b) a restaurant reviewing task for studying long open-ended text composition, and (c) behavioral and subjective evidence that phrase suggestions are treated as suggestions and shape the resulting composition to a greater extent than single words, which are treated as predictions. We conclude with an opportunity and a challenge: phrase suggestions now enable interactive systems to help people compose text, but the conventional approach of generating the *most accurate prediction* does not always make the *most valuable suggestion*.

RELATED WORK

Word prediction has been shown to help people enter text more quickly and accurately, showing keystroke savings of up to 45% in both mobile keyboards [10] and accessibility domains [17, 34, 33]. Despite its utility and near-ubiquity on mobile devices, word prediction has received relatively little research attention in mobile text entry. With appropriate parameter choices, a system can offer both corrections for mistakes and completions for partially entered words [5], but offering suggestions can incur costs of perception and interaction that hinder entry speed [28].

Numerous innovations in text entry have been studied, including gesture keyboards [39, 4, 23, 29, 1], key-target resizing [30, 15], alternative layouts [9, 38, 6], and sensor-based adaptation [3, 13]. Speed and accuracy gains have been reported in systems where the user presents a complete utterance to the system, which the system can then process as a whole: speech recognition with editing [35], or typing with feedback only at the end of a sentence [37].

While text input methods (such as those cited above) have been conventionally evaluated using transcription tasks, composition tasks are becoming more favored as they are more representative of people's actual use of text entry systems [36, 21]. However, while previous composition tasks aimed at a representative breadth of scenarios, we instead focus on the opportunities posed by long phrase suggestions in a single scenario where we have sufficient data to generate contextually relevant suggestions.

Suggestions of phrases and sentences have been explored in other application areas on desktop interfaces. In language translation, system suggestions of phrases and even entire sentences have been shown to be helpful [14]. To copy text, an "AutoComPaste" interaction complements the traditional copy-paste interaction with an autocomplete interaction [40]. And when users delete and retype to fix mistakes, a system can suggest text that was previously written [2]. Few studies have investigated suggestions of novel content; one exception is a case-based reasoning system that suggests phrases to consider using in product reviews [8].

Word prediction systems generally employ statistical language models to make predictions. Typical mobile word-prediction systems use small bigram models (e.g., [10]), though one commercial predictive keyboard has implemented neural language models that are able to utilize earlier context information[1]. With the benefit of longer context, statistical language modeling has been shown to be capable of accurately predicting phrases [26, 7]. Continuing advances in contextual language modeling (e.g., [12]) and generation (e.g., [25, 22]) should lead to further improvements in generating appropriate phrases and sentences. We use an n-gram prediction model, as this paper focuses on the interface design question and not on developing improved prediction models.

Google has deployed a system called "Smart Reply" [18] that suggests short responses to email messages. Like our approach, it also extends the familiar 3-options interaction to support phrases. The design elegantly leverages interruption theory by providing suggestions at the task boundary between reading and replying. But a different design (such as the one we chose in this experiment) is required for making suggestions *during* the reply process. They discuss some ways that an effective suggestion is different from an accurate prediction, e.g. post-processing the predictions to encourage topic diversity; we may build on their insights in future work.

SYSTEM

Our system generates word predictions in a manner that closely matches existing predictive typing systems. Given the text written so far, our system predicts three likely next words and shows them in equal-sized suggestion buttons, with the most likely suggestion in the center. If the user starts typing a word, the suggestions are instead completions of that word, but the behavior is otherwise the same.[2] Tapping a word inserts it with automatic spacing.

The novel element of our approach is that we extend the next-word predictions into phrase suggestions. For each single-word suggestion, the system predicts the most likely 5-word continuation phrase that begins with that word[3] and shows as much of the continuation phrase as fits (typically all 5 words on our devices, for a total of 6 visible words per slot) below the suggestion in smaller blue text (Figure 1). Tapping repeatedly on that suggestion slot inserts that phrase, one word per tap. After each tap the system also shows suggestions in the two other prediction slots, also including a word plus a continuation phrase.

To increase the ability of the system to make useful long suggestions, we focus on a single domain—for this study, we choose restaurant reviews. This choice anticipates by perhaps only a few years the ability of mobile devices to run powerful language models (e.g., via model compression [27, 11]) such as contextual neural language models [12] that can leverage a user's complete activity context, such as what kind of artifact they are currently writing and to what audience, plus their location history, prior communications, and other information.

[1] `https://blog.swiftkey.com/neural-networks-a-meaningful-leap-for-mobile-typing/`

[2] We did not use autocorrect for this study.

[3] We use beam search, commonly used in machine translation to find high-quality translations [14, 31].

We use a large domain-specific language model on a server to generate contextually relevant word and phrase suggestions. We built a word-level 5-gram model from the 213,670 restaurant reviews in the Yelp Academic Dataset[4]. We used KenLM [16] for language model queries, which uses Kneser-Ney smoothing [19], and we pruned n-grams that occur less than 2 times in the corpus. We mark the start-of-sentence token with additional flags indicating if the sentence begins a paragraph or the entire review.

For simplicity of both the experiment interface and backend processing, we restricted the allowed set of characters to lowercase ASCII letters and a restricted set of punctuation (see Figure 1). This restriction allowed our experimental keyboard to only need a single layer. In our experiments we instructed participants to disregard capitalization.

EXPERIMENT

We studied phrase suggestions in a free composition task of writing restaurant reviews. We chose this task because people often write reviews on mobile devices, and the task presents many opportunities for people to accept phrases that were not exactly what they might have written on their own but were still perfectly acceptable. We compared two conditions:

Phrase: phrase previews are shown in blue text beneath the single-word suggestions (as in Figure 1)

Word: identical behavior to Phrase except phrase previews are hidden

The only difference between the two conditions is whether or not the phrase preview is shown; identical one-word suggestions are shown in both conditions, and repeated taps on the same slot insert the same text that would have been inserted in the Phrase condition.

We used a within-subjects design: we asked participants to write four restaurant reviews, two for each condition (condition ordering was counter-balanced). To familiarize themselves with the keyboard and suggestion mechanism, participants first practiced with both conditions (order randomized). Then before writing reviews, participants wrote down names of four restaurants that they had visited recently. The system then guided them to complete each review in sequence (order randomized), alternating conditions between reviews. (This pre-commitment mechanism ensured that participants did not select restaurants based on, for example, the types of suggestions offered.) We instructed participants to write reviews that were at least 70 words in length, and displayed a word counter. We offered a reward for high-quality reviews.

Twenty students (undergraduate and graduate) participated in a 45-60 minute lab study for monetary compensation. We used 5th-generation iPod Touch devices, which have a 4-inch 1136x640 display. A remote server provides suggestions over WiFi; in practice, suggestions were shown in less than 100ms.

RESULTS

We report both behavioral data from system logs as well as subjective data from surveys done both after each review and at the conclusion of the session. All statistical analyses are mixed-effects models, with participant as a random effect and condition (Phrase or Word) as a fixed effect. Unless otherwise noted, we combine the logs of each participant's two trials for each condition. We exclude from analysis 19 reviews where more than 95% of the review was written using suggestions, leaving 61 reviews from 16 participants. We only report on whole-word suggestions, i.e., those suggestions offered when the participant had not yet begun typing a word.

Objective Measures

Participants accepted more whole-word predictions in the Phrase condition (F(1,15)=37.5, $p < .0001$): 45% of words[5] in Phrase condition compositions had been inserted by prediction, compared with 28% of words in Word condition reviews. This effect has two parts: (1) participants typed out a word when they could have used a suggestion more often in the Word condition (44% of times when a suggestion matching the final word chosen was offered) than in the Phrase condition (28%) (F(1,15)=19.1, $p < .001$), suggesting that participants paid more attention to suggestions in the Phrase condition, and (2) reviews written in the Phrase condition contain more words that had been offered as suggestions at the time they were entered: 63%, compared to 51% in the Word condition (F(1,15)=42.1, $p < .0001$). So showing phrases shaped the content that participants wrote more than showing the same suggestions one word at a time.

In both of our interfaces, repeated taps in the same suggestion slot insert successive words of a phrase. In the Phrase condition, where the participant saw a preview of upcoming words, participants accepted two suggestions in a row 1312 times, of which 85% were consecutively in the same slot, i.e., part of the same phrase. In contrast, in the Word condition, of the 776 times that participants accepted two suggestions in a row, 56% were consecutively in the same slot (F(1,15.3)=20.2, $p < 0.001$; one participant had no consecutive suggestion acceptances in either condition). As expected, the average length of phrases accepted (defined as consecutive whole-word suggestion acceptances in the same slot) was longer in the Phrase condition (mean 2.8 words) than the Word condition (1.5 words; F(1,15)=15.6, $p = .0013$); 14% of phrase acceptances were the full 6 words initially shown.

We compute the error rate by dividing the number of backspace taps (each deleting a single character) by the total number of taps. We did not observe a significant difference between conditions (25% in Phrase, 19% in Word, F(1,15)=3.2, n.s.). Our keyboard did not support any assistive gestures for correction, such as tap-and-hold to delete whole words, which we suspect would reduce the difference between conditions.

Overall typing rate was not significantly different between the two conditions (20.0 wpm[6] for Phrase, 20.9 for Word, F(1,15)=0.69, n.s.). On the one hand, participants were able to insert a phrase faster when they could see the preview (for same-slot transition times, Phrase mean = 0.8s, Word mean

[4] `https://www.yelp.com/dataset_challenge`

[5] Here, a "word" is a contiguous sequence of non-space characters.

[6] Here, a "word" is 5 consecutive characters, including spaces, the definition more common in text-entry studies.

1.1s, F(1,15)=21.7, $p < 0.001$ for log-transformed times); overall, 24% of all suggestion-to-suggestion sequences in the Phrase condition took less than 300ms, compared with 0.3% in the Word condition. But on the other hand, participants spent more time before starting to accept a suggestion (Phrase mean 1.2s, Word mean 0.9s, F(1,15)=19.7, $p < 0.001$) and after finishing accepting a suggestion (Phrase mean 1.3s, Word mean 1.3s, F(1,15)=16.4, $p = 0.001$).

Analyzing the two trials for each condition separately, we do not find any main effect of trial number on rate of suggestion usage (F(1,41.9)=3.87, n.s.) or error rate (F(1,43.1)=2.01, n.s.). Interaction of condition and trial number was also not significant for either analysis (F(1,41.75)=.002 for usage, F(1,42.7)=.002 for error rate, n.s.).

Subjective Measures

Participants reported that suggestions helped them think of *how to say what they wanted to say* more in the Phrase condition (1=*strongly disagree*, 5=*strongly agree*, mean 2.8) than the Word condition (mean 2.1; F(1,15)=6.4, $p = .02$). Participants also rated whether suggestions helped them think of *what to say*; ratings were marginally higher in the Phrase condition (mean 3.0, vs mean 2.3 for Word; F(1,15)=3.8, $p = .07$). In a cumulative survey, they more often reported that Phrase suggestions gave them ideas. (Phrase mean 3, Word mean 2.2, t(19)=2.3, $p = .03$.)

Overall preference was split nearly evenly: 11 participants preferred the Word keyboard and 9 preferred the Phrase keyboard. Participants liked that the phrase keyboard gave them ideas of both what to say and how to say it, sometimes in ways that were better than what they had in mind or better matched the style of the genre (in this case, restaurant reviews). But they disliked that the phrases suggested were often generic or trite, and felt that the phrases forced them into certain less creative ways of thinking. In contrast, the "Word" suggestions helped people write in "my own words" and be more "thoughtful." They also liked that text entry felt faster and easier in the Phrase condition, but some commented about spending a lot of time reading phrase suggestions (though there was no significant difference in ratings on "I felt like I spent a lot of time reading the suggestions": Phrase mean 3.0, Word mean 2.6, F(1,15)=1.5, n.s.). Participants commented that both Word and Phrase suggestions were often distracting, confirming the findings of a prior study [28].

We did not find a statistically significant difference in writing quality between the conditions: the mean of two independent ratings of overall quality by MTurk workers showed no difference (F(1,15)=.09, ns).

DISCUSSION

Phrase suggestions affected both the *process* and *product* of composition. The short delays between successive suggestion insertions indicate that participants successfully inserted phrases as units, rather than re-evaluating the suggestions for each successive word. (Consistent with a previous study on word suggestions, the additional cost to evaluate suggestions counteracted the speed benefit of inserting a suggestion, so the overall speed did not improve [28].) The phrase suggestions also shaped the final review: when shown phrases, participants accepted a greater number of suggestions, and those suggestions were more often repeated taps in the same suggestion slot. Since the single word shown in a suggestion slot was identical in the Word condition (i.e., it was also part of a phrase), this evidence indicates that people made different choices about what to say when predictions were presented as phrases.

Sometimes our system's phrase suggestions were helpful in showing examples of common phrases in restaurant reviews. But the phrases shown were the *most likely*, thus not novel or creative—and poor examples have been shown to hinder creativity [20, 24, 32].

CONCLUSION

In this paper, we propose a way to offer people phrase suggestions while writing in a way that leverages the familiar word-prediction interface. Our study found that many people appreciate the suggestions for content, style, and speed.

Ecological validity is a prominent limitation of our study: although our open-ended composition task is arguably much more representative of mobile typing in the wild than the transcription task used in almost all text-entry studies, our task was still a lab study with corresponding limitations: memories may have faded, and participants are not as invested in the resulting review as if it were to be actually published. We are exploring refinements to the task in follow-up studies.

A prominent limitation of multi-word suggestion systems is that high-quality suggestions are hard to generate unless something is known about what the user intends to write about. In this present study we partially avoided this limitation by focusing on the domain of restaurant reviews, where a large corpus of domain-specific text is available. Our approach is directly applicable to other task- or domain-specific entry tasks, such as other kinds of reviews (products, movies, etc.), customer relations management, or product support. In communications such as email, the recipients, subject, message thread, and prior sent messages can all serve as context with which to generate suggestions. In a general context across applications (where the active application is of course context), our interface could easily adapt by only showing phrase suggestions when there is sufficient context to merit suggestions. Gathering appropriate context across users, in an efficient and private manner, is left as an interesting open challenge. Since much of this data resides within specific applications, the keyboard software would need to interact heavily with applications so as to offer better suggestions.

We pose a research challenge in natural language processing: how to generate inspiring and valuable suggestions. Approaches may include 'offline' improvements in predictive modeling (e.g., modeling individual writing style, tracking topics have already been discussed, or focusing training on high-quality examples), or 'online' modeling of how authors respond to suggestions (e.g., are certain kinds of words or grammatical structures particularly inspiring?). *Novelty* is an important parameter for future work in phrase suggestion systems: the system could aim to generate suggestions that are common and likely to accepted but thus unoriginal, or sugges-

tions that are more novel and likely to be inspiring but less likely to be accepted verbatim. If a future system were able to offer suitable suggestions that were also *inspiring*, *eloquent*, or *clever*, or topics ripe for inclusion, our results suggest that people may embrace and appropriate them.

We studied one interaction design in a potentially vast design space. Since we saw such different behavior with two extremes of suggestion length (1 word in our Word condition and 6 words in our Phrase condition), suggestions of intermediate length are ripe for future study. Other future study could investigate showing suggestions only in certain conditions, varying the number of suggestions, or offering alternative interactions with suggestions.

Ethical deployment of a phrase suggestion system will require careful consideration of questions of authorship, freedom of expression, and originality. For example, if the system tends to offer suggestions with positive valence, will that bias the user towards writing a more positive review?

This work demonstrates that phrase completions can be offered in a way that they are accepted by the user and interpreted as suggestions rather than predictions. It also opens the door to future work on generating valuable phrase suggestions.

Acknowledgments

We thank Anna Huang for help conducting the study, Ofra Amir for manuscript feedback, and Kai-Wei Chang for many helpful discussions.

REFERENCES

1. Alsharif, O., Ouyang, T., Beaufays, F., Zhai, S., Breuel, T., and Schalkwyk, J. Long short term memory neural network for keyboard gesture decoding. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, IEEE (2015), 2076–2080.
2. Arif, A. S., Kim, S., Stuerzlinger, W., Lee, G., and Mazalek, A. Evaluation of a smart-restorable backspace technique to facilitate text entry error correction. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems*, CHI '16 (2016).
3. Azenkot, S., and Zhai, S. Touch behavior with different postures on soft smartphone keyboards. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, ACM (2012), 251–260.
4. Bi, X., Chelba, C., Ouyang, T., Partridge, K., and Zhai, S. Bimanual gesture keyboard. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, ACM (2012), 137–146.
5. Bi, X., Ouyang, T., and Zhai, S. Both complete and correct?: Multi-objective optimization of touchscreen keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, ACM (New York, NY, USA, 2014), 2297–2306.
6. Bi, X., Smith, B. A., and Zhai, S. Quasi-qwerty soft keyboard optimization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2010), 283–286.
7. Bickel, S., Haider, P., and Scheffer, T. Learning to complete sentences. In *Machine Learning: ECML 2005*. Springer, 2005, 497–504.
8. Bridge, D., and Healy, P. GhostWriter-2.0: Product reviews with case-based support. In *Proceedings of AI-2010, The Thirtieth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, M. Bramer, M. Petridis, and A. Hopgood, Eds., Springer London (London, 2011), 467–480.
9. Dunlop, M., and Levine, J. Multidimensional pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 2669–2678.
10. Fowler, A., Partridge, K., Chelba, C., Bi, X., Ouyang, T., and Zhai, S. Effects of language modeling and its personalization on touchscreen typing performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, ACM (New York, NY, USA, 2015), 649–658.
11. Geras, K. J., Mohamed, A., Caruana, R., Urban, G., Wang, S., Aslan, O., Philipose, M., Richardson, M., and Sutton, C. Blending LSTMs into CNNs. In *4th International Conference on Learning Representations (ICLR), workshop track* (2016).
12. Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., and Heck, L. Contextual LSTM (CLSTM) models for large scale NLP tasks. In *KDD Workshop on Large-scale Deep Learning for Data Mining (DL-KDD)* (2016).
13. Goel, M., Jansen, A., Mandel, T., Patel, S. N., and Wobbrock, J. O. ContextType: using hand posture information to improve mobile touch screen text entry. In *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM (2013), 2795–2798.
14. Green, S., Chuang, J., Heer, J., and Manning, C. D. Predictive translation memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, ACM (New York, NY, USA, 2014), 177–187.
15. Gunawardana, A., Paek, T., and Meek, C. Usability guided key-target resizing for soft keyboards. In *Proceedings of the 15th international conference on Intelligent user interfaces*, ACM (2010), 111–118.
16. Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Sofia, Bulgaria, August 2013), 690–696.
17. Higginbotham, D. J. Evaluation of keystroke savings across five assistive communication technologies. *Augmentative and Alternative Communication 8*, 4 (1992), 258–272.

18. Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., Corrado, G., Lukács, L., Ganea, M., Young, P., et al. Smart reply: Automated response suggestion for email. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, vol. 36 (2016), 495–503.
19. Kneser, R., and Ney, H. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, IEEE (1995), 181–184.
20. Kohn, N. W., and Smith, S. M. Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology 25*, 3 (2011), 359–371.
21. Kristensson, P. O., and Vertanen, K. The inviscid text entry rate and its application as a grand goal for mobile text entry. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '14, ACM (New York, NY, USA, 2014), 335–338.
22. Lipton, Z. C., Vikram, S., and McAuley, J. Capturing meaning in product reviews with character-level generative text models. *arXiv preprint arXiv:1511.03683* (2015).
23. Markussen, A., Jakobsen, M. R., and Hornbæk, K. Vulture: a mid-air word-gesture keyboard. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, ACM (2014), 1073–1082.
24. Marsh, R. L., Landau, J. D., and Hicks, J. L. How examples may (and may not) constrain creativity. *Memory & cognition 24*, 5 (1996), 669–680.
25. Mei, H., Bansal, M., and Walter, M. R. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment. *Proceedings of NAACL* (2016).
26. Nandi, A., and Jagadish, H. Effective phrase prediction. In *Proceedings of the 33rd international conference on Very large data bases*, VLDB Endowment (2007), 219–230.
27. Prabhavalkar, R., Alsharif, O., Bruguier, A., and McGraw, L. On the compression of recurrent neural networks with an application to LVCSR acoustic modeling for embedded speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (March 2016), 5970–5974.
28. Quinn, P., and Zhai, S. A cost-benefit study of text entry suggestion interaction. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 83–88.
29. Reyal, S., Zhai, S., and Kristensson, P. O. Performance and user experience of touchscreen and gesture keyboards in a lab setting and in the wild. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM (2015), 679–688.
30. Rudchenko, D., Paek, T., and Badger, E. Text text revolution: a game that improves text entry on mobile touchscreen keyboards. In *Pervasive Computing*. Springer, 2011, 206–213.
31. Rush, A. M., Chang, Y.-W., and Collins, M. Optimal beam search for machine translation. In *EMNLP* (2013), 210–221.
32. Siangliulue, P., Arnold, K. C., Gajos, K. Z., and Dow, S. P. Toward collaborative ideation at scale: Leveraging ideas from others to generate more creative and diverse ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, ACM (New York, NY, USA, 2015), 937–945.
33. Trnka, K., McCaw, J., Yarrington, D., McCoy, K. F., and Pennington, C. User interaction with word prediction: The effects of prediction quality. *ACM Transactions on Accessible Computing (TACCESS) 1*, 3 (2009), 17.
34. Trnka, K., and McCoy, K. F. Evaluating word prediction: framing keystroke savings. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, Association for Computational Linguistics (2008), 261–264.
35. Vertanen, K., and Kristensson, P. O. Intelligently aiding human-guided correction of speech recognition. In *AAAI* (2010).
36. Vertanen, K., and Kristensson, P. O. Complementing text entry evaluations with a composition task. *ACM Transactions on Computer-Human Interaction (TOCHI) 21*, 2 (2014), 8.
37. Vertanen, K., Memmi, H., Emge, J., Reyal, S., and Kristensson, P. O. 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*, ACM (2015), 659–668.
38. Zhai, S., and Kristensson, P. O. Interlaced QWERTY: accommodating ease of visual search and input flexibility in shape writing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2008), 593–596.
39. Zhai, S., and Kristensson, P. O. The word-gesture keyboard: reimagining keyboard interaction. *Communications of the ACM 55*, 9 (2012), 91–101.
40. Zhao, S., Chevalier, F., Ooi, W. T., Lee, C. Y., and Agarwal, A. AutoComPaste: Auto-completing text as an alternative to copy-paste. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, ACM (New York, NY, USA, 2012), 365–372.