RE-AIMing Predictive Text

Matthew Higgs, Claire McCallum, Selina Sutton, and Mark Warner
Northumbria Social Computing (NorSC) Group, Northumbria University, UK
{matthew.higgs, claire.mccallum, selina.sutton, mark.warner}
@northumbria.ac.uk

Abstract

Our increasing reliance on mobile applications means much of our communication is mediated with the support of predictive text systems. How do these systems impact interpersonal communication and broader society? In what ways are predictive text systems harmful, to whom, and why? In this paper, we focus on predictive text systems on mobile devices (Figure 1) and attempt to answer these questions. We introduce the concept of a 'text entry intervention' as a way to evaluate predictive text systems through an interventional lens, and consider the **R**each, **E**ffectiveness, **A**doption, **I**mplementation, and **M**aintenance (RE-AIM) of predictive text systems. We finish with a discussion of opportunities for NLP.

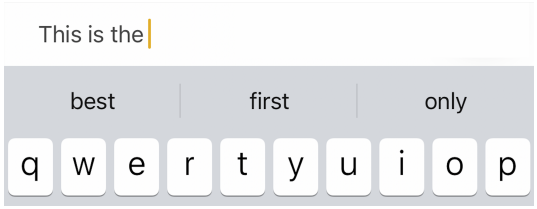

Figure 1: Screenshot of predictive text on an iOS device, with input text above and default keyboard below.

1 Introduction

Predictive text systems were born from an engineering paradigm (Church and Mercer, 1993), integrated into user interfaces (Garay-Vitoria and Abascal, 2006), and deployed at scale (Apple, 2014; Samsung, 2021) - with little consideration for their impact on interpersonal communication and broader society. Originally designed to increase the speed and ease of communicating with a computer (Darragh et al., 1990), predictive text systems have been optimised over time towards the goals of accuracy and efficiency on writing tasks (Bi et al., 2014; Quinn and Zhai, 2016), and creators of recent email-based predictive text systems highlight how their system saves users over "one billion characters of typing" (Chen et al., 2019). More recently though, Quinn and Zhai (2016), Arnold et al. (2018, 2020), and Hancock et al. (2020) respectively highlight concerns for the benefits of predictive text on mobile devices, the effects on the content we write, and the impact of predictive text as it mediates our communications.

We aim to begin a discussion on how research priorities and industry practice can change to better our understanding of the benefits and harms of predictive text systems, and to imagine how they might be redesigned. We introduce the concept of a 'text entry intervention' (an intervention delivered as a user enters text) where examples include: an app offering privacy and well-being advice alongside the messages children type (BBC, 2019); a writing assistant intended to influence a writer's grammar, spelling, style and tone (Grammarly, 2019); or an email assistant intended to reduce repetitive writing and the number of keys we type (Chen et al., 2019). We focus on the common 'text suggestion bar' on mobile devices (Figure 1) as an instance of a text entry intervention, and describe a path forward for evaluating its impact. Note that all of these examples of text entry interventions are real products, being used by real people, with little research into the effects they have on the content produced. Once we consider predictive text to be an intervention, we can draw on areas of research focused on the design and evaluation of interventions (Flay, 1986; Glasgow et al., 1999, 2019; Murray et al., 2016). In doing so, a new set of criteria with which to evaluate predictive text systems becomes available; enabling us to examine aspects of their design and potential impact, and encouraging us to consider the relationships between predictive text systems and their downstream effects (Figure 2).

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 109–115

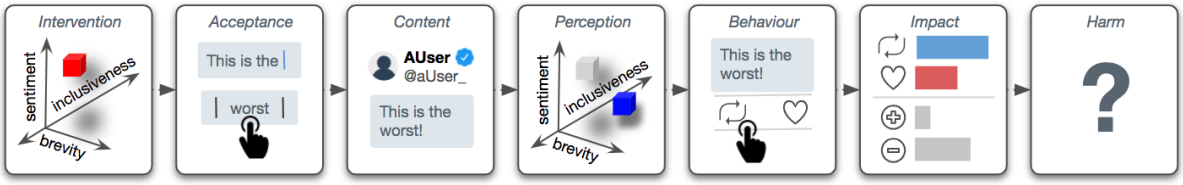

Figure 2: An example connecting a predictive text system to its impact and harms. The *intervention* is represented as a point in a design space, which (when combined with other factors) affects: the user's *acceptance* of the text suggestions and the *content* they post online, the reader's *perception* of the content and their *behaviour* (e.g. like, retweet, or nothing), and (over time) the *impact* and *harm* of the intervention.

We aim to: (i) begin a discussion about the benefits and harms of predictive text systems, (ii) conceptualise predictive text as an intervention, and (iii) highlight opportunities for NLP research.

2 Related Work

Predictive text systems have evolved from communication aids designed to benefit novice and reluctant typists (Darragh et al., 1990; Garay-Vitoria and Abascal, 2006) to default features on smartphones (Apple, 2014; Samsung, 2021) and popular email clients (Kannan et al., 2016). While there are some large scale studies of web-based predictive text system for structured transcription tasks (Palin et al., 2019), there is little research into their use in the wild (Buschek et al., 2018) and evidence that lab-based studies might not reflect real world use (Komninos et al., 2020). Recent research into the effects of predictive text on the content we write (Arnold et al., 2020) highlights the need for more studies in-the-wild, but this brings significant challenges (Buschek et al., 2018).

With regards to evaluating predictive text systems, a key challenge for any researcher-developed intervention is ensuring it changes real world practice. To help intervention developers assess real world impact, the RE-AIM framework (Reach, Effectiveness, Adoption, Implementation, Maintenance) was developed (Glasgow et al., 1999, 2019). It asks: does the intervention **R**each those it should? Is it **E**ffective in doing 'more good than harm' in real world contexts? (Flay, 1986)). Is it likely to be **A**dopted and **I**mplemented by those required to deliver it? Will any changes be **M**aintained over time? The RE-AIM framework leads to key questions we need to ask to decide how research priorities and industry practice should change (Murray et al., 2016).

With regards to improving predictive text systems, we propose re-purposing methods for controlling text generation (Ghazvininejad et al., 2017; Holtzman et al., 2018; Tambwekar et al., 2018; Keskar et al., 2019; Sahar et al., 2020) which are being used to control 'bias' in text generation (Sheng et al., 2020; Dinan et al., 2020). With regards to measuring downstream effects, the intersection of causal inference and language provide methods for estimating the causal effects of linguistic properties on downstream outcomes (Pryzant et al., 2020); such as the effect of wording, brevity, and tone on message propagation and impact on Twitter (Tan et al., 2014; Gligorić et al., 2019) and the effects of tone in online debates (Sridhar and Getoor, 2019).

3 Evaluating Text Entry Interventions

Conceptualising predictive text systems as a type of text entry intervention, we draw on the RE-AIM framework (Glasgow et al., 1999, 2019) to consider the reach and effects of predictive text systems and how new interventions might be adopted, implemented, and maintained by the industry actors who control their development and deployment. Specifically, application of the RE-AIM framework (and consideration for how research priorities and industry practice could change) forces us to identify significant gaps in our current knowledge regarding the impact of predictive text systems.

3.1 Reach

While predictive text systems are available on all mobile devices by default, there exists little evidence on their uptake and the extent to which text suggestions are actually used (Buschek et al., 2018). A recent study of a web-based predictive text system for transcription enrolled $37,000$ volunteers, suggesting significant reach, but the study was of a structured task set by researchers (Palin et al., 2019), as opposed to real world use. A significant barrier to assessing the reach of predictive text systems is limited access to usage data, which is often

held by industry gatekeepers. Nevertheless, reach must be considered early in the development and redesign of predictive text systems. What is the reach of current predictive text systems? What proportion of smartphone users have predictive text turned on? Which users accept and benefit from text suggestions? In which contexts (e.g. messaging, email, social media) are text suggestions used? How can we ensure new predictive text systems reach intended users? How can we be sure intended users participate and engage? Will reach be limited in some way through, for example, not accounting for various user characteristics or industry gatekeepers?

3.2 Effectiveness

To understand effectiveness, we need to consider both benefits and harms (Flay, 1986).

Originally designed to increase the speed and ease of communicating with a computer (Darragh et al., 1990), HCI research has focused on the benefits of predictive text in relation to accuracy and efficiency (Bi et al., 2014; Quinn and Zhai, 2016). However, there are doubts over the benefits of text suggestions on smartphones, and Quinn and Zhai (2016) provided evidence that the costs of attending to and using text suggestions impaired average time performance. So we must ask, who actually benefits from text suggestions on mobile devices, in which contexts, and how?

HCI research into the harms or unintended consequences of predictive text is more limited, though we can connect recent research into the effects of predictive text on content (Arnold et al., 2020) to research into the harms of more general NLP systems (Blodgett et al., 2020). For example, it has been suggested that predictive text may make our writing more predictable (Arnold et al., 2020), and when we consider the scale to which such systems are deployed and the effects of language on society (Blodgett et al., 2020), we must consider the possibility that predictive text on mobile devices is reinforcing social norms and leading to representational harms. Specifically, we can imagine differences in system performance (predictive text benefiting particular social groups more/less than others), stereotyping that propagates through suggestions, and minority languages slowly being rendered invisible through a lack of representation. Currently though, we do not fully understand the unintended consequences and potential harms of predictive text systems, and we need to ask: in what ways are predictive text systems harmful, to whom, and why?

The full extent and variety of downstream consequences and social outcomes can be difficult to anticipate, and capture within a trial (Oliver et al., 2019). As suggested by (Blodgett et al., 2020), qualitative and participatory approaches exploring the lived experiences of those likely to use and be affected by systems, are needed, and in advance of deployment. To understand the benefits and harms of predictive text systems, more evaluations should take place in-the-wild (Buschek et al., 2018). This will help us to understand not only *whether* a predictive text system is effective, but for who, why and in what context.

3.3 Adoption, Implementation & Maintenance

Whether new predictive text systems are being introduced, or existing systems are being re-designed, it is imperative that we explore the willingness of industry sectors to deliver them. Consulting and co-designing such interventions with those delivering them will help us understand and mitigate potential barriers to their adoption.

A key challenge to adoption and implementation will be understanding and aligning with the values and goals of the organisations who control delivery. Instances of industry further adapting predictive text interventions (e.g. to meet their own business goals), after they have been tested for their effectiveness, should be monitored. While adaptations and system updates are inevitable, these new versions should be re-tested for their positive and negative effects on users.

Implementation relies on real world user engagement patterns being similar to those found in effectiveness trials. The challenge, however, is assessing real world use outside of trials. Previous research has invested in developing logging systems that can measure typing behaviour (including use of predictive text) in-the-wild (Buschek et al., 2018). However, currently, industry professionals are under little obligation to share this with researchers. So we ask, will such organisations share (in a secure way) the data needed to properly evaluate ongoing implementation and maintenance (i.e. the long term impact) of text systems? And when provided with evidence for the effectiveness of interventions, will they adapt existing systems accordingly?

4 Theory to Practice

In this section, we consider how we can involve users and affected communities in the design of improved predictive text systems, and how we can engage with industry players.

4.1 Community involvement

Regarding qualitative and participatory approaches involving users and affected communities, we focus on collaborative research through 'co-design' where users and communities would play a large role in knowledge development, idea generation and concept development (Zamenopoulos and Alexiou, 2018). We propose four outputs from such research: 'stakeholder maps' to identify and characterise the communities affected and industry players involved; 'evaluative constructs' (e.g. measures of impact) designed to increase benefits and decrease harms (Metcalf et al., 2021); 'system designs', such as custom keyboards and other text entry interventions (Arnold et al., 2018, 2020); and 'study designs' to evaluate alternative systems with regards to the measures of impact. Additionally, to better understand (and measure) the benefits and harms of predictive text, we need to better understand people's communication goals and use of predictive text in different domains. Focusing on communication through mobile devices, we propose to characterise and compare the benefits and harms of using predictive text when used alongside different types of apps; such as messenger/communication, email, and social.

4.2 Industry involvement

Cooperation from industry players is essential for answering many key questions, but industry objectives might be not be in line with the objectives of this work. If this is the case, and the research meets resistance from industry players, we propose a path forward consisting of three (escalating) levels of engagement with industry players.

The first level ('co-design') would involve industry players in the design process, and would attempt to understand the suitability of current predictive text systems in meeting their business needs. Specifically, why are predictive text systems part of their offering? Why it is cost-effective to maintain predictive text? Which objectives are considered when evaluating predictive text? Research would aim to incorporate such objectives into the design process, and work with industry players and affected communities to design new solutions balancing stakeholder needs. The second level ('negotiation') would occur if industry actors refuse to be involved in the design process, and would aim to support the identification and communication of the benefits and harms of predictive text. Specifically, research would aim to develop tools for surfacing and communicating the benefits and harms of predictive text, and building a body of evidence which can be used to negotiate changes in current predictive text systems. Along with this, we might consider establishing a forum to act on the evidence and mandate changes in the implementation of predictive text systems (Metcalf et al., 2021). The third level ('resistance') would occur if efforts relating to the first two levels fail, and would aim to provide affected/concerned communities with the tools to take action themselves. The simplest example of this would be an 'opt out' movement where users (beyond affected communities) are encouraged to turn off predictive text, though it is not clear the impact this would have. An alternative would be to design 'adversarial' text entry interventions to counter the effects of existing predictive text systems. These could take the form of custom keyboards which users install on their devices, though developing and maintaining such systems at scale would come with significant cost and (potentially) need to undergo industry review.

5 Opportunities for NLP

In this section, we draw on the literature on controllable text generation (Ghazvininejad et al., 2017; Holtzman et al., 2018; Tambwekar et al., 2018; Keskar et al., 2019; Sahar et al., 2020; Sheng et al., 2020; Dinan et al., 2020) and the intersection of causal inference and language (Tan et al., 2014; Gligorić et al., 2019; Sridhar and Getoor, 2019; Pryzant et al., 2020) to discuss the role NLP can play in the development of improved predictive text systems. Specifically, we focus on: (1) implementing new systems, and (2) measuring downstream effects.

5.1 Implementing new systems

To develop improved predictive text systems, we need to be able to align future systems with desired outcomes and communication goals, and we believe existing methods for controlling text generation can be re-purposed for this task.

We believe methods for controlling text gener-

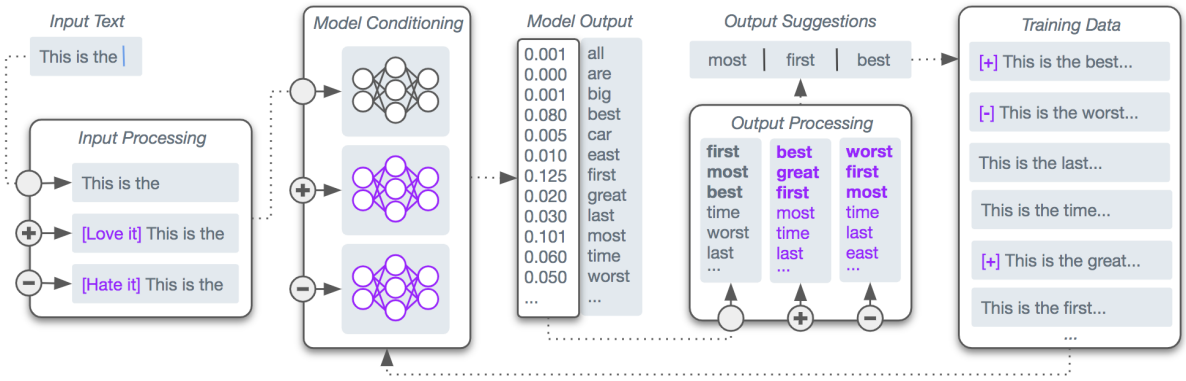

Figure 3: An example text suggestion pipeline showing the three main *control points* for biasing text suggestions: *Input Processing* where the input text is processed before being passed to a language model, *Model Conditioning* where the language model is conditioned during training, and *Output Processing* where the output of the language model is processed before the selection of text suggestions. Each control point can be used to provide different conditions to different groups of participants, and toy examples are shown for generating control ( ), positive (+), and negative (−) sentiment biases at each control point.

ation fall into three main groups, depending on the point in the text processing pipeline at which they exert control (Figure 3). Each have their own advantages and disadvantages with regards to the effort/cost/skills required to implement a new system to achieve a desired outcome, and in how likely the system is to be accepted by users. We highlight: input processing methods such as 'smart prompts' (Sheng et al., 2020), model conditioning methods such as in Tambwekar et al. (2018); Keskar et al. (2019); Dinan et al. (2020), and output processing methods such as 'guided decoding' (Ghazvininejad et al., 2017; Holtzman et al., 2018). Note that input/output processing methods do not require (potentially expensive) data collection/augmentation and/or language model training/fine-tuning.

5.2 Measuring downstream effects

To evaluate predictive text, we need to be able to measure its downstream effects (Figure 2).

With regards to measuring the effects of a predictive text system (intervention) on what we write (content), there is evidence that predictive text encourages predictable writing (Arnold et al., 2020) and biased text suggestions leads to biased content (Arnold et al., 2018) for writing tasks, but it's not clear if these results hold in-the-wild. We believe user acceptance is key to successfully delivering a (predictive text) intervention, and studies will need to consider how acceptance will vary for different contexts of use. Indeed, Buschek et al. (2018) observed variation in the use of text suggestions across messenger/communication, email, and social apps. Further downstream, there are opportunities for measuring the effects of content (produced with the support of predictive text) on reader behaviours and perceptions (Pryzant et al., 2020), message propagation (Tan et al., 2014; Gligorić et al., 2019), and online dialogue (Sridhar and Getoor, 2019). With these measurement methods, we hope to co-design and monitor suitable proxies for harms arising from predictive text.

6 Conclusion

Through this work, we hope to begin a discussion about the benefits and harms of predictive text systems, and believe the first step is to conceptualise predictive text systems as interventions. We focused on predictive text systems on mobile devices (the 'text suggestion bar'), but believe research opportunities exists for other NLP-driven text entry interventions. We believe research in this area can benefit from evaluation criteria taken from the field of behaviour change, which focuses on taking interventions from research to practice. We believe there are many opportunities for NLP and HCI researchers to improve existing, and design new, predictive text systems. In particular, HCI can center the work around the lived experiences of those affected, and NLP can help implement new systems and measure downstream effects. This will be challenging, and involve conducting studies in-the-wild and negotiating with the organisations who control the delivery of predictive text systems and the collection of user data.

References

Apple. 2014. ios 8 - quicktype - apple.

Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2018. Sentiment bias in predictive text recommendations results in biased writing. In *Proceedings of the 44th Graphics Interface Conference*, GI '18, page 42–49, Waterloo, CAN. Canadian Human-Computer Communications Society.

Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 128–138, New York, NY, USA. Association for Computing Machinery.

BBC. 2019. Bbc launches 'digital wellbeing' own it app for children.

Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2014. Both complete and correct? multi-objective optimization of touchscreen keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2297–2306, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. *ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild*, page 1–14. Association for Computing Machinery, New York, NY, USA.

Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. *Gmail Smart Compose: Real-Time Assisted Writing*, page 2287–2295. Association for Computing Machinery, New York, NY, USA.

Kenneth W. Church and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

J. J. Darragh, I. H. Witten, and M. L. James. 1990. The reactive keyboard: a predictive typing aid. *Computer*, 23(11):41–49.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Brian R. Flay. 1986. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine*, 15(5):451–474.

Nestor Garay-Vitoria and Julio Abascal. 2006. Text prediction systems: a survey. *Universal Access in the Information Society*, 4(3):188–203.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada. Association for Computational Linguistics.

R E Glasgow, T M Vogt, and S M Boles. 1999. Evaluating the public health impact of health promotion interventions: the re-aim framework. *American Journal of Public Health*, 89(9):1322–1327. PMID: 10474547.

Russell E. Glasgow, Samantha M. Harden, Bridget Gaglio, Borsika Rabin, Matthew Lee Smith, Gwenndolyn C. Porter, Marcia G. Ory, and Paul A. Estabrooks. 2019. Re-aim planning and evaluation framework: Adapting to new science and practice with a 20-year review. *Frontiers in Public Health*, 7:64.

Kristina Gligorić, Ashton Anderson, and Robert West. 2019. Causal effects of brevity on style and success in social media. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Grammarly. 2019. Great writing, simplified.

J. Hancock, M. Naaman, and Karen Levy. 2020. Ai-mediated communication: Definition, research agenda, and ethical considerations. *J. Comput. Mediat. Commun.*, 25:89–100.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

Anjuli Kannan, Peter Young, Vivek Ramavajjala, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, G.s Corrado, Laszlo Lukacs, and Marina Ganea. 2016. Smart reply: Automated response suggestion for email. pages 955–964.

Nitish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Andreas Komninos, Kyriakos Katsaris, Emma Nicol, Mark Dunlop, and John Garofalakis. 2020. Mobile text entry behaviour in lab and in-the-wild studies: Is it different?

Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. *ACM Conference on Fairness, Accountability,and Transparency (FAccT '21)*.

Elizabeth Murray, Eric B Hekler, Gerhard Andersson, Linda M Collins, Aiden Doherty, Chris Hollis, Daniel E Rivera, Robert West, and Jeremy C Wyatt. 2016. Evaluating digital health interventions: Key questions and approaches. *American journal of preventive medicine*, 51(5):843–851.

Kathryn Oliver, Theo Lorenc, Jane Tinkler, and Chris Bonell. 2019. Understanding the unintended consequences of public health policies: the views of policymakers and evaluators. *BMC Public Health*, 19(1):1057.

Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How do people type on mobile devices? observations from a study with 37,000 volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA. Association for Computing Machinery.

Reid Pryzant, D. Card, Dan Jurafsky, V. Veitch, and D. Sridhar. 2020. Causal effects of linguistic properties. *ArXiv*, abs/2010.12919.

Philip Quinn and Shumin Zhai. 2016. *A Cost-Benefit Study of Text Entry Suggestion Interaction*, page 83–88. Association for Computing Machinery, New York, NY, USA.

Hady El Sahar, Maximin Coavoux, Matthias Gallé, and Jos Rozen. 2020. Self-supervised and controlled multi-document opinion summarization. *ArXiv*, abs/2004.14754.

Samsung. 2021. How can i personalise and turn predictive text on and off on my samsung galaxy device?

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1872–1878. International Joint Conferences on Artificial Intelligence Organization.

Pradyumna Tambwekar, Murtaza Dhuliawala, Animesh Mehta, Lara Martin, Brent Harrison, and Mark Riedl. 2018. Controllable neural story generation via reinforcement learning.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland. Association for Computational Linguistics.

Theodore Zamenopoulos and Katerina Alexiou. 2018. *Co-design As Collaborative Research*. Connected Communities Foundation Series. Bristol University/AHRC Connected Communities Programme, Bristol.