Bridging the Communication Rate Gap: Enhancing Text Input for Augmentative and Alternative Communication (AAC)

Hussein Yusufali, Stefan Goetze, and Roger K. Moore

Department of Computer Science
The University of Sheffield, Sheffield, UK
{hsayusufali1, s.goetze, r.k.moore} sheffield.ac.uk

Abstract. Over 70 million people worldwide face communication difficulties, with many using augmentative and alternative communication (AAC) technology. While AAC systems help improve interaction, the communication rate gap between individuals with and without speaking difficulties remains significant, and this has led to a low sustained use of AAC systems. The study reported here combines human computer interaction (HCI) and language modelling techniques to improve the ease of use, user satisfaction, and communication rates of AAC technology in open-domain interactions. A text input interface utilising word prediction based on BERT and RoBERTa language models has been investigated with a view to improving communication rates. Three interface layouts were implemented, and it was found that a radial configuration was the most efficient. RoBERTa models fine-tuned on conversational AAC corpora led to the highest communication rates of 25.75 words per minute (WPM), with alphabetical ordering preferred over probabilistic ordering. It was also found that training on conversational corpora such as TV and Reddit outperformed training based on generic corpora such as COCA or Wikipedia. Hence, it is concluded that the limited availability of large-scale conversational AAC corpora represent a key challenge for improving communication rates and robust AAC systems.

Index Terms: Text Input Prediction, Language Modelling, Augmentative and Alternative Communication (AAC), Speech Synthesis

1 Introduction

Speech impairments can have significant inhibiting effects on affected individuals. It is estimated that over 0.5% of the UK population and 7.5 million individuals in the United States have a specific type of vocal impairment [36,14]. The effects of speech impairments can vary significantly depending on the severity and extent of the individual's impairment. The degree of impairment determines whether an individual may benefit from an AAC device. The primary objective of the AAC strategy or device is to assist or facilitate a user's communication ability [41,42].

An inhibiting factor in the adoption of AAC is the significant gap in communication rates between standard spoken conversational interaction and that experienced by AAC device users [8]. AAC devices have been shown to reach communication rates of approximately 10–16 WPM whereas conversational rates can surpass 140–150 WPM [46,38]. Communication rates measure the rate at which individuals exchange information during an interaction; whereas conversational rates indicate the pace of a conversation between two or more individuals which incorporates turn-taking and the flow of a dialogue. The significant gap between these two rates restricts individuals with communication difficulties from effective participation in multiparty interactions with minimal delay [22]. Overall, current assistive technologies cannot enable AAC users to participate successfully in multiparty interactions due to large time delays and a lack of expressivity, naturalness and personalisation [46,17,48].

Prediction and predictive technologies are integral to the development of AAC devices, as these techniques have the potential to decrease the communication gap between participants by using text or character predictions (as in Dasher [52], a graphical user interface (GUI) based system presenting animated character predictions as an innovative text input interface at the time of its publication). However, for these techniques to be effective, the phrase, word or character prediction mechanisms must be accurate and have a minimal delay to ensure that the device is not detrimental towards the user or interaction [44,24]. If prediction mechanisms are not accurate, they can amplify the communication gap by increasing delays in the exchange of information. Nevertheless, AAC users widely use Pictureboards, Pragmatic Organisation Dynamic Display (PODD), speech generation AAC devices and predictive systems such as *Predictable*[1] [22,49], depending on an individual's unique communication abilities. The implementation of advanced text prediction techniques such as language modelling, topic modelling, using semantic or syntactic information and conversational context has significantly improved AAC devices [8,16,22,27].

In the design and evaluation of user interfaces, including assistive technologies like AAC devices, HCI plays a critical role. User modelling is essential to determine the efficacy of a system that is aimed at users with a diverse range of abilities and circumstances. Various models are utilised to understand how individuals interact with interfaces and systems, and these generate effective design and evaluation justifications for developing systems, specifically AAC systems. For instance, 'task analysis models' have been used to capture a user's ability to achieve a specific task outside of a computer system or interface [1]. In particular, AAC system designs are able to benefit from HCI modelling as they can inform the development of personalised and effective AAC strategies and devices. Furthermore, HCI modelling can identify potential barriers to device usage as well as be used to design interfaces that are intuitive and accessible to users with different requirements [1]. Indeed, it has been shown that by considering the needs and abilities of diverse users, AAC designers can create

[1] https://therapy-box.co.uk/predictable - Available on June 2023

more inclusive and effective assistive technologies [2]. However, despite these advances, there is still a need for improved prediction-based interfaces for AAC users.

This paper addresses these issues and is structured as follows: Section 2, discusses modelling techniques for user interface design. Section 3 describes the experimental method and interface designs used in modelling user behaviour for text prediction interfaces and the results are presented in Section 4. Section 5 concludes the paper. The main contributions of this paper are four-fold:

1. A general model for user predictive text-entry has been introduced based on established HCI laws.
2. Three word-prediction selection interfaces for text input have been modelled.
3. Choice vs. human response time (HRT) has been investigated for word selection, concluding in the number of optimal choices presented to the user in a text-input interface.
4. Fine-tuning large language models (LLM) for increased text-entry for AAC devices have been examined, with conversational data and testing text input user interfaces across four communication scenarios.

2 User Behaviour Modelling and Text Prediction for Increased Text Entry

Within the field of HCI there are specific models relating to how users interact with movements within an interface and how users interact with choices provided. Models such as 'Fitts' Law' [15] and the 'Hick Hymen Law' [21] provide designers with predictability information about user behaviour, which can be used to justify design decisions and enhance the user experience [45]. Additionally, predictive models allow for the calculation of metrics in an analytical manner [35], eliminating the need for time-consuming and resource-intensive experimentation.

HCI is intrinsically linked to human-motor movement. Therefore, when designing interfaces and systems, models of movement can be utilised to inform the best approaches in the design of systems. An early example of such an HCI modelling is the keystroke level model (KLM) [5]. KLM predicts the approximate time it takes the user to perform a specific task using a system, assuming that there are no errors. However, users differ significantly, and some factors have more impact on performance than others [30]. For example, users have different ways to use their hands and motor skills. The topic of bimanual (the use of both hands simultaneously) and laterality (preference of the use of one hand) has been extensively studied with Guiard's model [18] being one of the most notable. The examination of Guiard's model is critical, due to enhancing user experience and developing efficient interfaces. This model proposes a framework for understanding preferred and non-preferred hand movements. Guiard's model proposes that individuals have a preferred hand (usually the dominant hand) that performs tasks that require precision and fine motor control, and a

non-preferred hand that performs tasks that require less precision and forceful movements. This understanding is important when analysing the efficacy of user interfaces, particularly for tasks involving fine motor movements.

Furthermore, other factors are also critical, such as the variety of goals that can be achieved with a system and the ability to recall a specific function after a period of not using a system [23]. There are also associated factors such as user fatigue, concentration, and satisfaction [30].

2.1 Interface Modelling

Fitts' Law: Fitts' law [15] states that the time taken for a cursor to move in a user interface is directly related to the area and width of the target on the interface. Fitts defines an Index of Difficulty (ID) as a function of the Euclidean distance D between two points on the screen $\mathbf{p}_1 = [x_1, y_1]^T$ and $\mathbf{p}_2 = [x_2, y_2]^T$ and width W of a target.

$$\mathrm{ID} = \log_2 \left(\frac{2D}{W} \right) \tag{1}$$

The ID is a measure of complexity or difficulty for a movement in an interface, such as clicking or selecting a widget via using a cursor or on a touchscreen. The ID is directly related to a prediction of a Movement Time (MT) as follows:

$$\mathrm{MT} = a + b \cdot \mathrm{ID} \equiv a + b \cdot \log_2 \left(\frac{2D}{W} \right) \tag{2}$$

The constants a and b reflect the efficiency of the interfaces and systems, in particular pointing and mouse cursor movements. They are constants that are determined empirically via regression analysis.

Fitts' Law (and its variants) have become ubiquitous in predicting the performance and difficulty of movement in interface designs; it is one of the most widely adopted models for human performance prediction and behaviour modelling [45,19].

Hick Hymens Law: A second model ubiquitous in HCI is the Hick-Hymens law [21,45], which states that the reaction time of an individual T will be logarithmically correlated to the number of choices n presented to the user. This model was investigated to determine the optimal number of words presented to a user in a text-input word prediction interface as follows:

$$T = b \cdot \log_2(n + 1) \tag{3}$$

The constant b is determined empirically via experimentation. There are several time latencies related to human motor and cognitive behaviour that must be examined to justify design interfaces and layouts.

2.2 Text Entry Speeds

A significant limiting factor for the sustained use of AAC devices is the difference between text input rates and spoken conversational rates. The QWERTY keyboard has become predominant in text-entry layout interfaces, both in physical text-entry and touchscreen interfaces [33]. Text-entry rates for a typical QWERTY keyboard can be predicted using a variation of Fitts' law (the Shannon formulation [32]) as follows:

$$\mathrm{MT}_{i,j} = a + b \cdot \log_2 \left(\frac{D_{i,j}}{W} + 1 \right) \tag{4}$$

In (4), i and j are the indices of the separate keys and $D_{i,j}$ is the distance between key i and key j. However, this model does not take into account if a key is repeated. This was investigated in [33] which predicted a text input rate of 30.1 WPM for a typical soft QWERTY keyboard. This model is in accordance with studies [26] representing entry speeds aligning to the predictive model established by [33].

Spoken conversational rates and text entry rates are considerably different [3]. There are also consequential differences between spoken and written language [6]. This was investigated by [29] establishing that the highest text entry rates that can be achieved are not only dependent on the text entry method, but also on the user transferring thoughts and information to written text. An 'inviscid' text-entry rate of 67 WPM was determined by [29], establishing a grand goal for text entry input methods. Table 1 summarises text entry rates reported in the state-of-the-art literature and the respective references.

Table 1: Comparison of text entry rates with varying text input methods.

Text Entry Method	**WPM Rate**	**Reference**
Inviscid Upper Bound	67	[29]
QWERTY Physical Keyboard	51.56 ± 20.2	[13]
QWERTY Touchscreen Keyboard	45	[26]
Gesture Keyboards (Swype)	45	[25]
Dasher	17.26 (Upper Bound)	[52]

2.3 Language Modelling

Statistical language modelling techniques are utilised in the predictive text input to increase communication rates [7]. These techniques are thus essential for text input into AAC devices. Prediction mechanisms can be character-based, word-based, or phrase-based (multiple words), demonstrated by the Dasher interface [52], where character predictions are presented to the user in a streaming interface with low latency. Prediction with gesture keyboards such

as Swype (virtual touch-based keyboard allowing users to join characters by a sliding gesture) [25] and phrase prediction techniques such as [27] have demonstrated the efficacy of language modelling techniques for AAC devices. LLMs have the ability to generate text based on user input with the incorporation of contextual history and semantic correctness. LLMs, such as transformer-based models, Bidirectional Encoder Representations from Transformers (BERT) [12,43] and Robustly Optimized BERT (RoBERTa) [55] can be fine-tuned (adapting the pre-trained model to a specific task) for specifically the use case in conversational open-domain AAC devices (cf. Section 3.3 for corpora used for fine-tuning). These language models are chosen because they have been shown to achieve high accuracy on language generation tasks and to score high on commonly used metrics such as General Language Understanding Evaluation (GLUE) and BiLingual Evaluation Understudy (BLEU) [39,51]. RoBERTa has proven to be effective in language generation due to dynamic masking in model training.

To increase text entry rates and decrease the communication gap between AAC users and their conversational partners, predictive language modelling is utilised in this work (cf. Section 3.3), together with various input methods and interfaces (cf. Section 3.1).

The process of predictive text input: There are associated time latencies involved when utilising a system. To optimise a system, a functional component analysis can be conducted to optimise the controllable parameters [38]. A functional structure, i.e. decomposition of the components of the system into sub-functions to analyse the system either as a whole or at the component level, can be critical in analysing certain time latencies and shortcomings of a system. Figure 1 visualises a general overview of the functional flow and associated time components in a predictive text input system, which can be summed to equal the time per utterance (i.e. a sequence of written text):

- T_{Type}: time to enter the initial word and type characters for a word, if a word prediction is not selected by the user
- $T_{\text{Prediction}}$: time necessary for language model to predict the next word
- T_{Look}: time for the user to scan the search space of word predictions
- T_{React}: decision time of the user to decide if the word predictions are satisfactory
- T_{Select}: time for the user to select a word prediction or revert to typing.

This functional system flowchart, when used in conjunction with envelope analysis [28], i.e. analysing each sub-component with a user-centric approach, can be instrumental in improving the efficiency and effectiveness of system designs. Please note, that timings in Figure 1 do not correspond to the time T in (3).

3 Experimental Method

Informal experiments were conducted by the main author, being an AAC user for many years.

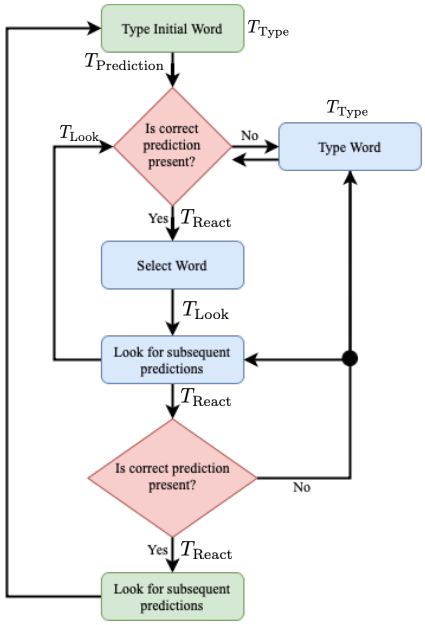


Fig. 1: Functional system flowchart for the developed interface to select predictive words, including time delays introduced by system and user.

3.1 Fitts' Law Modelling and Experimentation of User Interfaces

To model the impact of layout changes, Fitts' law [15] in (2) is utilised and tested to predict the time of movements, i.e. the time required to move to a target position on an interface, performed by users during use of the interface. Three interfaces are tested with differing button sizes, and distance lengths across the screen which influence the ID.

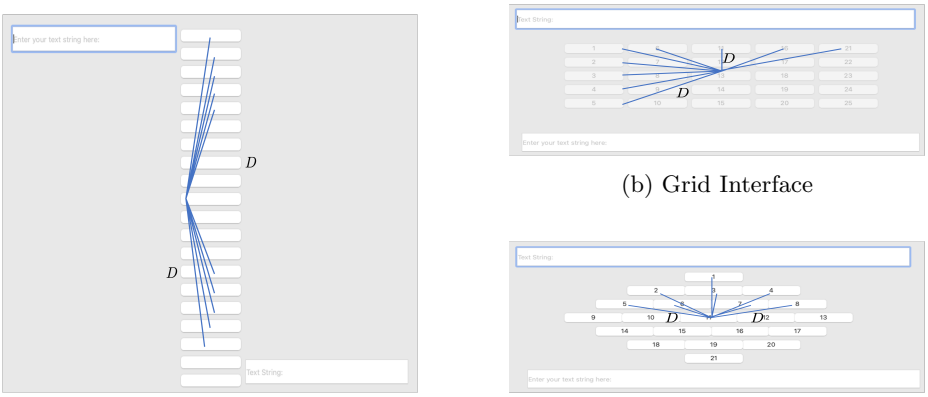
(a) Alphabetical Ordering

(b) Probabilistic Ordering

Fig. 2: Ordering strategies: Predicted words are presented to the user either in (a) alphabetical or (b) probabilistic order, i.e. sorted based on the confidence score of the LLM text prediction model. The word predictions are generated after the phrase 'The weather is' has been entered.

Further to this, two different ordering strategies of the predicted words are investigated as visualised in Figure 2: (i) alphabetical ordering (cf. Figure 2a)

and (ii) ordering via probabilities of the generated words (cf. Figure 2b). The probabilities are the confidence scores (likelihood of this word prediction following the utterance) which are generated together with the word predictions and which can be interpreted as confidence scores of the prediction.



(a) List Interface

(b) Grid Interface

(c) Radial Interface

Fig. 3: GUI Interface Designs.

Figure 3 shows the three layouts under test. The list layout interface in Figure 3a was chosen as this is similar to the list style of the Dasher Keyboard [52], a grid layout (cf. Figure 3b), since it is similar to the auto-predict selection above a QWERTY keyboard (QWERTY) on a smartphone, with a similar width to a smartphone auto-predict interface, and the radial interface (cf. Figure 3c), since it is similar to a hexagonal keyboard, such as e.g. the TYPEHEX or Hookes Keyboard [40,54]. Coefficients a and b in (2) are determined by regression analysis with x and y values empirically measured, where x and y are scalar coordinates on the Fitts' plots (Movement Time (MT) vs ID) which are empirically measured, correlated to the distances moved by the user in the interfaces.

Additionally for accurate measurements, the cursor origin must remain constant for each distance measurement when empirically evaluating an interface. For each interface layout in Figure 3, three cursor origin positions were tested; top of the interface, middle centre and bottom centre of the interfaces. The purpose of investigating the origin cursor positions was to determine which cursor origin reduces the ID in (2) for the user and, thereby, increase efficiency. Cursor movements can vary between different users due to dexterity, fine-motor movement and also the concentration of the user. However, Fitts' law provides a foundation for the predictability and efficacy of user interfaces, which is utilised to determine an efficient interface for predictive text input.

3.2 Number of Choice of Words vs Human Response Time

To optimise the time latencies for text entry, controllable parameters in the system can be refined. An initial experiment is conducted to investigate the relationship between the number of word choices and the HRT. Users were provided with lists containing varying numbers of words, denoted as n in (3), and their reaction times to select a specific word from the list were measured. To ensure accurate time measurements and to minimise human errors, an automatic timer was used.

Hick-Hymen's law according to (3) is commonly used to investigate the choice paradigm. However, this experiment aimed to investigate word choices given a list, which is not covered by the choice paradigm. Therefore, the experiment focused on this aspect. The word *"the"* was chosen as the target word for selection from the list. *"The"* was chosen as it is a common English word that is easily comprehensible to all users. Furthermore, the experiment aimed to validate and predict the optimal number of words to present to the user for predictive text input. Human response times were measured with different numbers of words presented to the user. Random word lists were created, lists consisting of common English words and users were instructed to select the word *"the"* from each list. Each trial, corresponding to each list size, was repeated 3 times to ensure fairness and accuracy. The lists increased in length by 5 word increments.

To accurately represent human response times for word selection, factors and environmental conditions such as screen size, brightness, location, mouse or cursor speeds, font and button sizes and text colour were kept constant. Additionally, users were given practice instructions and a warm-up period of 3 minutes enabling them to familiarise themselves with the task.

3.3 Language Model Fine-tuning

A limiting factor in AAC device usage is currently the inability of devices to support open-domain multiparty interactions. Some devices are only helpful in very limited domains, e.g. for communication with friends or family.

To enable a greater variety of domains, LLM are trained that have demonstrated to be successful on various tasks. LLM can be fine-tuned on specialist corpora for downstream tasks, such as conversational word prediction. The corpora below are utilised to fine-tune both BERT and RoBERTa models [12,55].

Corpora Used: The data used for model fine-tuning is outlined below:

- The **TV Corpus** [10] contains over 325 million words and has collected data from over $75,000$ TV episodes and shows. The TV corpus contains informal language and dialogues from a collection of TV shows and is considered the largest corpora of informal language available.
- The **Switchboard Corpus** [4] consists of spontaneous telephone conversations between American-English speakers of over 300 hours of

recorded and transcribed multi-gender and multi-topic speech. There is a total of 2300 conversations covering 70 topics.
- **The Corpus of Contemporary American English (COCA)** [9] is a widely used American English corpus and contains over one billion words, spanning eight genres: fiction, popular magazines, newspapers, academic texts, TV, Movies and blogs.
- The **Wikipedia Corpus** [11] is one of the largest corpora based upon a large Wikipedia collection from 2014. The corpus contains over 2 billion words, covering a variety of genres and topics, together with over 4.4 million pages of data.
- The **AAC Corpus** [50] is one of a limited amount of specialist AAC corpora available for AAC applications, especially to train large-scale models on. The corpus consists of imagined sentences and responses from users who imagine themselves using AAC devices. The AAC corpus is small in size, with approx. $6,000$ messages from 289 unique workers on glsAMT. However, it is considered to be beneficial when training language models for specific AAC applications.
- The **Reddit Corpus** [34] is a collection of cross-domain text, scraped from Reddit and containing over 256M conversational threads across a variety of unspecified domains, ranging from $2015 - 2018$.
- The **Daily Dialog** [31] corpus is a smaller, fully annotated corpus utilised for multi-turn dialogues, with 13k dialogues and the average speaker turns per dialogue of 7.9 turns. It is an open-domain corpus with utterances being more formal than other widely used corpora, such as Twitter or Reddit corpora.

The training times for fine-tuning times vary depending on the size of the training data, between 4 hours to 7 days on 2 NVIDIA V100 GPU's (each with 32GB RAM). Table 2 summarises parameters for the model adaptation. The learning rate, weight decay and Adam weights were utilised as the same as fine-tuning BERT or RoBERTa models in the original tuning parameters.

Table 2: Hyperparameter values used during fine-tuning of LLMs.

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Learning Rate	$5 \cdot 10^{-5}$	Adam Beta_1	0.9
Batch Size	16	Adam Beta_2	0.99
Number of Training Epochs	50	Adam Epsilon	$1 \cdot 10^{-8}$
Weight Decay	0.01		

3.4 Interface Testing for AAC Scenarios

An informal study was conducted to evaluate the interfaces together with the fine-tuned language models by the first author of this work. To mimic

close-to-realistic situations, four communication scenarios and tasks were defined: a scripted dialogue for ordering coffee, a half-scripted dialogue where topics were provided but the responses were not scripted an open-ended question, and a picture description.

Task 1 - Scripted dialogue

- **AU**: *A coffee with cream and sugar, please.*
- **CP**: *Which size?*
- **AU**: *A small, please. Thanks.*
- **CP**: *Here you are.*
- **AU**: *I think I have some change somewhere; let me check.*
- **CP**: *Thanks.*
- **AU**: *Thanks, have a nice day.*

AU is the AAC user utilising the interface and responding to the utterances given by CP, the conversational participant.

Task 2 - Half prompted dialogue

Task 2 aims to simulate a workplace scenario where the participants are asked to engage in small talk. They are prompted to start with a general greeting, sustain the conversation, transition to discussing an after-work event and conclude by talking about their workload. This dialogue intends to total 9 utterances in the conversation, between both the AAC user and conversation participant and mimic a realistic situation.

Task 3 - Open-ended scenario

Task 3 aims to test the accuracy and capability of the language model by instructing the participant to utilise the interface, the participant is instructed to answer the following question with the interface: *Please describe a weekend when you have had fun memories*

Task 4 - Picture description

Task 4 is primarily aimed at testing the open-domain capabilities of the fine-tuned language models, as well as assessing their ability to construct long sentences effectively. To accomplish this, an image depicting a cat trapped on a tree [20,37] was used. This was chosen due since it allows for the description of various topics and is commonly used in testing aphasia patients.

During the study of the interfaces, two metrics, i.e.

$$\text{WPM} = \frac{\text{number of words typed}}{\text{time taken in seconds}} \cdot 60, \tag{5}$$

and

$$\text{Accuracy} = \frac{\text{number of words selected in utterance}}{\text{total number of words in utterance}} \cdot 100 \tag{6}$$

were calculated for each utterance. Subsequently, the average WPM and accuracy rates were computed across all utterances in the given communication scenario.

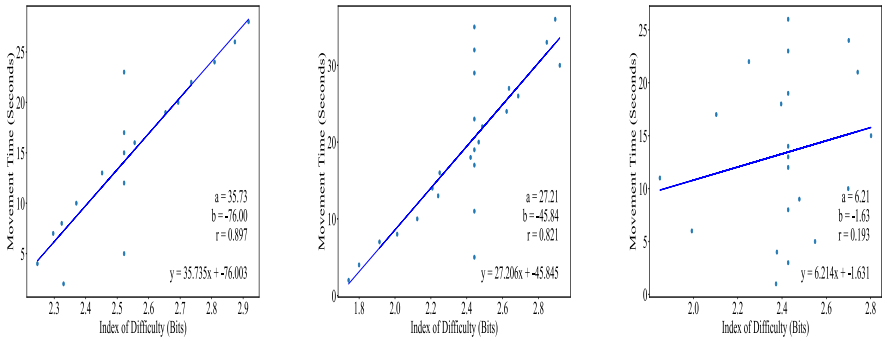
To minimise and mitigate any unnecessary delays during the use of the AAC system, the predictions generated from the language models are processed

through a stemmer [47] to remove any individual punctuation predictions. The decision to remove individual punctuation predictions is in line with previous work [27]. By this, the AAC system offers a more efficient and streamlined user experience.

4 Results

The results of the initial experiments conducted by the main author to investigate the speed of text input and the various factors involved in users' text entry are presented in the following.

4.1 Fitts' Law Experimentation for User Interfaces



(a) List Interface (b) Grid Interface (c) Radial Interface

Fig. 4: Comparison of movement time in seconds over ID for the three user interfaces shown in Figure 3.

Figure 4 shows movement time in seconds over ID for the interface layouts shown in Figure 3, determined by the methodology described in Section 3.1. The interface buttons have a constant size (width W). Points in Figure 4 represent ID for each button on the interfaces, and the time MT necessary for the user to move towards the target selection area of the button according to (2). The coefficients of Fitts' law, a and b, were determined by regression analysis and are summarised in Table 3.

The results indicate that the list interface Figure 3a is the slowest and most difficult for the user. The results align with Fitts' law; the user has to move further in the list interface and cover longer distances to reach the target selection area, consequently resulting in higher ID and MT. The grid interface (cf. Figure 3b) is less efficient for text entry. Fitts' law is indicative of preferring

Table 3: Fitts' law's parameters for the user interface layouts.

Interface Layout	Cursor Position	Equation of Regression Line
List	Middle	MT = 35.73 − 76.00 ID
Grid	Middle	MT = 27.21 − 45.84 ID
Radial	**Middle**	**MT = 6.21 - 1.63 ID**

radial layouts (as in Figure 3c) over list layouts, which is reflected in Figure 4. Radial layouts offer shorter and closer movement distances for the users, resulting in lower ID.

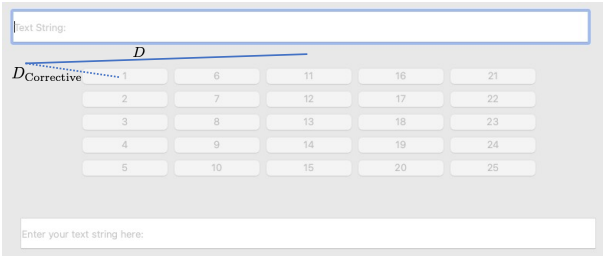


Fig. 5: Representing a corrective distance by a user if the user has missed the target selection area and has to revert back.

The experiments also show that radial layouts are not always preferred by users. This is because the options in radial layouts are in close proximity to each other, therefore increasing the risk of error selection. The experiments also assume that the distances and timings measured follow one single smooth movement by the user and no corrective movements are considered. Corrective movements occur when the user moves towards a target selection area and misses the target; therefore introducing a corrective distance as visualised in Figure 5, which can affect the MT. Faster movements do usually result in increased errors, particularly with small target widths [53].

Fitts' law indicates a strong correlation between the ID and MT. For the three interfaces tested, only the list and grid interfaces depicted in Figure 3a and Figure 3b show a strong correlation between the target areas (buttons) and ID in Figure 4. The list layout interface shows the strongest correlation. The radial layout shown in Figure 4c shows a weaker correlation, however, the Fitts' law coefficients a and b were also low, indicating a radial layout is effective, if the target selection areas (i.e. the buttons for selecting the word predictions) are closely aligned in the interface. The distances between the buttons are in close proximity in the interface, so the user's MT is reduced, therefore decreasing the ID.

4.2 Number of Choice of Words vs Human Response Time

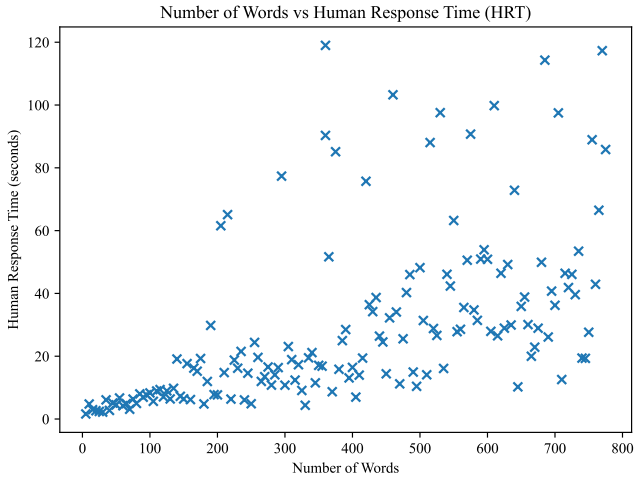


Fig. 6: Relationship between the number of words presented to a user in a list, and the response time taken to select a specific word from the list.

The graph depicted in Figure 6 analyses results generated from the methodology described in Section 3.2. The results indicate that the HRT increases as the number of words presented to the user increases, which is consistent with Hick-Hymen's law as expressed in (3). However, it is noteworthy that the observed relationship is not logarithmic, deviating from Hick-Hymen's law in (3). While the correlation between the number of words presented and the HRT does appear to be linear, there are noticeable outliers that become prominent when the number of words exceeds a search space of 200.

Figure 6 show a strong correlation between HRT and the number of words presented for a small search space, below 100 words. However, as the search space becomes larger, there is an increased number of observed outliers. Thus, for efficient text entry, the search space n must be limited to mitigate and minimise latency.

4.3 Language Model Fine-tuning and Interface Testing for AAC devices

Results for fine-tuning of the LLM as described in Section 3.3 are presented in Figure 7, highlighting the advantages of utilising conversational corpora for fine-tuning language models, particularly for AAC systems. Results show that the baseline models BERT and RoBERTa show lower performance and

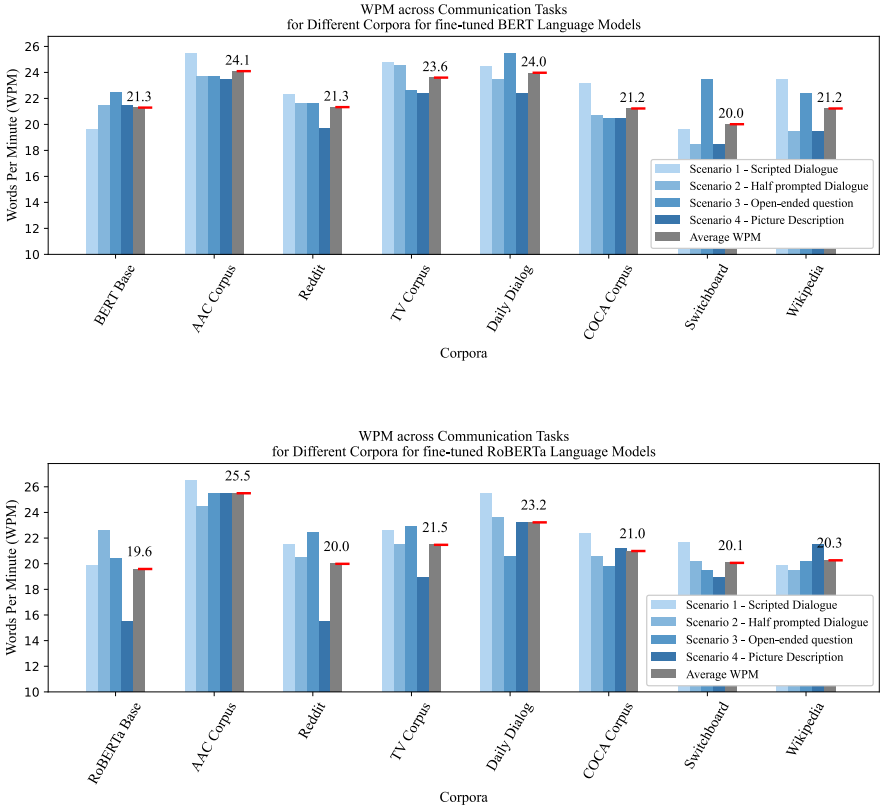


Fig. 7: Comparison of WPM across four communication scenarios utilising a radial interface with BERT and RoBERTa fine-tuned language models.

are therefore less suitable for predictive text entry, both, in terms of accuracy and WPM rates. The results for the baseline models can be attributed to the model's inability to provide accurate or satisfactory predictions across the four communication scenarios. In comparison, fine-tuning with conversational corpora produces higher accuracy and WPM rates compared to using more generic corpora such as the Wikipedia or COCA Corpus. However, results also show that specific corpora, such as the Switchboard and Reddit corpora, were not especially useful for open-domain applications (across the four communication scenarios when testing the interfaces) because of their limited domain applications.

To determine the best ordering of the word predictions, i.e. alphabetical or probabilistic (cf. Figure 2), subsequently, interface testing was conducted across the four communication scenarios. Initial experiments and use of the interfaces indicated that alphabetical ordering via the first character of the word

predictions was preferred over probabilistic order (ordering via the confidence scores of the generated predictions). The alphabetical ordering preference did result in higher communication rates. However, further experimentation is necessary to investigate the impact of changing the layout or style of how the word predictions appear to the user on text input rates.

The results presented in Figure 7 indicate that RoBERTa models outperformed BERT models, with higher resulting communication rates, primarily due to RoBERTa models having higher prediction quality. Consequently, the RoBERTa models outperformed certain fine-tuned BERT models. However, the difference in certain fine-tuned models is marginal. This study emphasises the importance of selecting appropriate corpora for specific applications. While conversational corpora are useful for open-domain applications, domain-specific corpora may be more suitable for limited certain tasks, specifically helpful for long-term AAC usage. The limited robustness of AAC systems and the lack of adaptability of AAC systems in various domains pose challenges. However, the lack of available and open-source AAC corpora makes training such open-domain models challenging. The results also showed that both fine-tuned BERT and RoBERTa models that were fine-tuned on the smaller specialist AAC [50] corpora achieved higher communication rates, compared to other generic corpora fine-tuned models. These fine-tuned models consistently outperformed the other models across the four communication scenarios.

Experiments also show that the fine-tuned models achieve higher accuracy rates in both scripted and prompted communication tasks, as described in Section 3.4. This higher accuracy can be attributed to the specific communication scenarios being scripted, allowing the models to generate more accurate predictions. However, when tested on open-domain communication scenarios, the fine-tuned models did not perform as well. Resulting in consistently lower communication rates. This emphasises a limitation in AAC devices, where the systems are more effective in short scripted or prompted dialogues and not open-domain tasks.

The four communication scenarios described in Section 3.4 aimed to simulate realistic situations and varied topics. However, these experiments are still not indicative of the long-term usage of AAC systems. The study could benefit from a long-term longitudinal study, to gain a comprehensive understanding of the capabilities of the system and to test the robustness of the interface. Overall, this study highlights the benefits of leveraging conversational data and fine-tuning language models for improving the performance and usability of AAC systems.

5 Conclusion

The limited communication output rates of individuals using AAC systems significantly hinder their ability to engage in multiparty open-domain interactions. This paper proposed a solution by combining techniques from HCI and language modelling to bridge the communication rate gap between AAC

users and typical speakers. The preliminary results demonstrate that a radial text input interface integrated with a RoBERTa language model fine-tuned on conversational corpora, specifically fine-tuned on AAC, outperforms other text input interfaces, such as a grid or list interface. This approach achieves a communication rate output of 25.75 WPM across four simulated communication scenarios. Although the text input methods do not match the typical typing rate of a QWERTY keyboard, the communication rates surpass those of other AAC devices. Future research has to focus on refining user interfaces by minimising redundant time latencies and improving language prediction capabilities in conjunction with HCI modelling.

References

1. Benyon, D., Murray, D.: Applying user modeling to human-computer interaction design. Artificial Intelligence Review **7**(3-4), 199–225 (8 1993)
2. Biswas, P., Robinson, P.: Automatic evaluation of assistive interfaces. In: Proceedings of the 13th international conference on Intelligent user interfaces. pp. 247–256 (2008)
3. Cai, S., Venugopalan, S., Tomanek, K., Kane, S., Morris, M.R., Cave, R., Macdonald, R., Campbell, J., Casey, B., Kornman, E., et al.: Speakfaster observer: Long-term instrumentation of eye-gaze typing for measuring aac communication. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–8 (2023)
4. Calhoun, S., Carletta, J., Brenier, J.M., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D.: The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language resources and evaluation **44**(4), 387–419 (2010)
5. Card, S.K., Moran, T.P., Newell, A.: The keystroke-level model for user performance time with interactive systems. Communications of the ACM **23**(7), 396–410 (1980)
6. Chafe, W., Tannen, D.: The relation between written and spoken language. Annual review of anthropology **16**(1), 383–407 (1987)
7. Copestake, A.: Augmented and alternative NLP techniques for augmentative and alternative communication (AAC). In: Natural Language Processing for Communication Aids (1997)
8. Curtis, H., Neate, T., Vazquez Gonzalez, C.: State of the art in AAC: A systematic review and taxonomy. In: Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility. pp. 1–22 (2022)
9. Davies, M.: The Corpus of Contemporary American English as the first reliable monitor corpus of English. Literary and linguistic computing **25**(4), 447–464 (2010)
10. Davies, M.: The TV and Movies corpora: Design, construction, and use. International Journal of Corpus Linguistics **26**(1), 10–37 (2021)
11. Denoyer, L., Gallinari, P.: The wikipedia XML corpus. In: International Workshop of the Initiative for the Evaluation of XML Retrieval. pp. 12–19. Springer (2006)
12. Devlin, J., Chang, M., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, p. 2 (2019)

13. Dhakal, V., Feit, A.M., Kristensson, P.O., Oulasvirta, A.: Observations on typing from 136 million keystrokes. In: Conference on Human Factors in Computing Systems - Proceedings. vol. 2018-April. Association for Computing Machinery (4 2018)
14. Dupré, D., Karjalainen, A.: Employment of disabled people in Europe in 2002. Statistics in focus pp. 3–26 (2003)
15. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology **47**(6), 381–391 (6 1954)
16. Garay-Vitoria, N., Abascal, J.: Text prediction systems: a survey. Universal Access in the Information Society **4**(3), 188–203 (2006)
17. Goetze, S., Moritz, N., Appell, J.E., Meis, M., Bartsch, C., Bitzer, J.: Acoustic User Interfaces for Ambient Assisted Living Technologies. Informatics for Health and Social Care, SI Ageing & Technology **35**(4), 161–179 (Dec 2010)
18. Guiard, Y.: Asymmetric division of labor in human skilled bimanual action: the kinematic chain as a model. Journal of Motor Behavior **19**(4), 486–517 (1987)
19. Guiard, Y., Olafsdottir, H.B., Perrault, S.T.: Fitt's law as an explicit time/error trade-off. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1619–1628 (2011)
20. Hameister, I., Nickels, L.: The cat in the tree–using picture descriptions to inform our understanding of conceptualisation in aphasia. Language, Cognition and Neuroscience **33**(10), 1296–1314 (2018)
21. Hick, W.E.: On the rate of gain of information. Quarterly Journal of Experimental Psychology **4**(1), 11–26 (3 2008)
22. Higginbotham, D.J., Shane, H., Russell, S., Caves, K.: Access to aac: Present, past, and future. Augmentative and alternative communication **23**(3), 243–257 (2007)
23. John, B.E., Kieras, D.E.: The goms family of user interface analysis techniques: Comparison and contrast. ACM Transactions on Computer-Human Interaction (CHI) **3**(4), 320–351 (1996)
24. Krause, J., Taliaferro, A.: Supporting students with autism spectrum disorders in physical education: There's an app for that. Palaestra **29**(2), 45 (2015)
25. Kristensson, P.: Discrete and continuous shape writing for text entry and control. Ph.D. thesis, Linköping University (2007)
26. Kristensson, P.O., Brewster, S., Clawson, J., Dunlop, M., Findlater, L., Isokoski, P., Martin, B., Oulasvirta, A., Vertanen, K., Waller, A.: Grand challenges in text entry. In: CHI'13 Extended Abstracts on Human Factors in Computing Systems, pp. 3315–3318. Association for Computing Machinery, ACM (2013)
27. Kristensson, P.O., Lilley, J., Black, R., Waller, A.: A Design Engineering Approach for Quantitatively Exploring Context-Aware Sentence Retrieval for Nonspeaking Individuals with Motor Disabilities. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. vol. 20, pp. 1–11. Association for Computing Machinery (ACM) (4 2020)
28. Kristensson, P.O., Müllners, T.: Design and analysis of intelligent text entry systems with function structure models and envelope analysis. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2021)
29. Kristensson, P.O., 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 & services. pp. 335–338 (2014)

30. Kurosu, M.: Human-Computer Interaction: Human-Centred Design Approaches, Methods, Tools and Environments: 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part I, vol. 8004. Springer (2013)
31. Ma, K., Jurczyk, T., Choi, J.D.: Challenging Reading Comprehension on Daily Conversation: Passage Completion on Multiparty Dialog. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **1**, 2039–2048 (2018)
32. MacKenzie, I.S.: Fitts' Law as a Research and Design Tool in Human-Computer Interaction. Human–Computer Interaction **7**(1), 91–139 (3 1992)
33. Mackenzie, I.S., Zhang, S.X., Soukoreff, R.W.: Text entry using soft keyboards. Behaviour & Information Technology **18**(4), 235–244 (1999)
34. Medvedev, A.N., Lambiotte, R., Delvenne, J.C.: The anatomy of reddit: An overview of academic research. Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches 10 pp. 183–204 (2019)
35. Moore, R.K.: Modeling data entry rates for ASR and alternative input methods. In: Interspeech (2004)
36. Morris, M.A., Meier, S.K., Griffin, J.M., Branda, M.E., Phelan, S.M.: Prevalence and etiologies of adult communication disabilities in the united states: Results from the 2012 national health interview survey. Disability and health journal **9**(1), 140–144 (2016)
37. Nicholas, L.E., Brookshire, R.H.: A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. Journal of Speech, Language, and Hearing Research **36**(2), 338–350 (1993)
38. Ola Kristensson, P., Müllners, T.: Design and Analysis of Intelligent Text Entry Systems with Function Structure Models and Envelope Analysis. Analysis **12** (2021)
39. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
40. Pritom, A.I., Mahmud, H., Ahmed, S., Hasan, M.K., Khan, M.M.: Typehex keyboard: A virtual keyboard for faster typing in smartphone. In: 2015 18th International Conference on Computer and Information Technology (ICCIT). pp. 522–526. IEEE (2015)
41. Rackensperger, T., Krezman, C., Mcnaughton, D., Williams, M.B., D'silva, K.: "When I first got it, I wanted to throw it off a cliff": The challenges and benefits of learning AAC technologies as described by adults who use AAC. Augmentative and alternative communication **21**(3), 165–186 (2005)
42. Rennies, J., Goetze, S., Appell, J.E.: Personalized Acoustic Interfaces for Human-Computer Interaction. In: Ziefle, M., C.Röcker (eds.) Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications, chap. 8, pp. 180–207. IGI Global (2011)
43. Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics **8**, 842–866 (2021)
44. Schepis, M.M., Reid, D.H., Behrman, M.M.: Acquisition and Functional Use of Voice Output Communication by Persons with Profound Multiple Disabilities. Behavior Modification **20**(4), 451–468 (1996)
45. Seow, S.C.: Information theoretic models of HCI: A comparison of the Hick-Hyman Law and Fitt's Law. Human-Computer Interaction **20**(3), 315–352 (2005)

46. Shane, H.C., Blackstone, S., Vanderheiden, G., Williams, M., DeRuyter, F.: Using AAC technology to access the world. Assistive technology **24**(1), 3–13 (2012)
47. Sharma, D., Cse, M.: Stemming algorithms: a comparative study and their analysis. International Journal of Applied Information Systems **4**(3), 7–12 (2012)
48. Shire, S.Y., Jones, N.: Communication partners supporting children with complex communication needs who use aac: A systematic review. Communication Disorders Quarterly **37**(1), 3–15 (2015)
49. Todman, J., Alm, N., Higginbotham, J., File, P.: Whole Utterance Approaches in AAC. Augmentative and Alternative Communication **24**(3), 235–254 (2008)
50. Vertanen, K., Kristensson, P.O.: The imagination of crowds: conversational AAC language modeling using crowdsourcing and large data sources. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 700–711 (2011)
51. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 353–355. Association for Computational Linguistics, Brussels, Belgium (Nov 2018)
52. Ward, D.J., Blackwell, A.F., MacKay, D.J.C.: Dasher—a data entry interface using continuous gestures and language models. In: Proceedings of the 13th annual ACM symposium on User interface software and technology. pp. 129–137 (2000)
53. Wobbrock, J.O., Cutrell, E., Harada, S., MacKenzie, I.S.: An error model for pointing based on fitts' law. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 1613–1622 (2008)
54. Zhai, S., Hunter, M., Smith, B.A.: Performance optimization of virtual keyboards. Human–Computer Interaction **17**(2-3), 229–269 (2002)
55. Zhuang, L., Wayne, L., Ya, S., Jun, Z.: A robustly optimized BERT pre-training approach with post-training. In: Proceedings of the 20th Chinese National Conference on Computational Linguistics. pp. 1218–1227. Chinese Information Processing Society of China (Aug 2021)