Evaluating transcription tools for Shetland dialect

Jack Irvine

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2023

Abstract

Automatic speech recognition (ASR) is used to reduce the effort associated with the transcription of natural spoken language. This approach has many advantages, with a significant reduction in the effort required to convert speech to text. Under-resourced languages stand to benefit significantly from new developments in this area as they often lack the resources to effectively transcribe and translate spoken content, develop language models, or create language-specific tools and applications. Unfortunately, these tools may also present biases impacting the production of linguistic resources which may reduce the value of human input. This project explores the impact of error highlighting in ASR generated transcripts when used as a starting point for human transcription of a language with a non-standard orthography. The Shetland dialect shares similarities with both Standard English and Scots, yet retains a rich vocabulary and varied orthography. Using the language as a case study, we consider how susceptible it may be to standardisation as a result of the use of ASR tools in assisted transcription tasks.

The main components of the project include the development and processing of a specialized corpus for the Shetland dialect, the training and evaluation of a language model based on the constructed corpus, the design and implementation of a cross-platform transcription interface, and a user study to evaluate the effects of error highlighting in the context of non-standard orthographies. We find that error highlighting has a significant effect on the spelling choice of human transcribers where the correct output is ambiguous. This project lays the groundwork for future research aiming to increase the value of human input in assisted transcription tasks for languages with non-standard orthographies, ultimately improving the overall accessibility and preservation of low resourced language resources.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 280672
Date when approval was obtained: 2022-11-03
The participants' information sheet and a consent form are included in the appendix.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Jack Irvine*)

Acknowledgements

I would like to acknowledge several individuals and organisations who have contributed significantly to the success of this project. First and foremost, I would like to thank my supervisor, Peter Bell, for his support and thoughtful feedback throughout the course of the project.

I'm grateful to my parents and family, who have assisted with vocabulary and helped to provide early feedback on the initial design of the study.

I'd like to thank Angus Johnson for sharing his work, having transcribed almost all of the recordings used in the project, and the Shetland Archives, who have collected and stored these resources.

Many thanks to my fellow committee members at Shetland ForWirds and other organisations that have promoted the study, helping it reach a wider audience and ensuring robust participation. Finally, I would like to thank all of the study participants, whose engagement and enthusiasm have contributed a great deal to the results of this project.

Table of Contents

Chapter 1

Introduction

Transcription is the process of converting the spoken word to text. This process has taken many forms throughout human history, from writing by hand, to typewriters, on-screen keyboards and eventually, advanced systems such as the dictation feature available on modern smartphones. These tools are constantly evolving to not only increase productivity for the general population, but to allow access to technology for those who struggle with mainstream interfaces. In a 2013 study investigating the effects of interface on the transcription speed of blind smartphone users, researchers found speech input to be five times faster than an on-screen keyboard with voice over enabled (Azenkot and Lee, 2013). The study highlighted the frustration of error correction, and recommended further improvement to the text review and error detection process.

Speech transcription presents several challenges to even experienced transcribers, such as accurately capturing the nuances of spoken language and understanding speakers with diverse accents and dialects. Transcribers often face difficulties in differentiating overlapping speech or conversations taking place in noisy environments and the presence of domain specific vocabulary can make transcription even more complex. To overcome these challenges and produce high quality transcriptions, experienced transcribers must possess a deep understanding of the language and be skilled at identifying and adapting to the various obstacles. As a consequence of these challenges, transcription can be a time consuming process, with even professional transcribers taking many times the duration of an audio file to convert it to plain text. For more challenging corpora, a minute of audio may take an average of 40 minutes to transcribe with unfamiliar speech requiring disproportionate effort (Foley et al., 2018).

1.1 The Shetland dialect

While transcription tools have made significant progress in recent years, the focus has largely been on widely spoken languages and there is still much work to be done regarding the preservation and documentation of endangered languages, such as the Shetland dialect. The Shetland dialect (also known as Modern Shetlandic Scots) is spoken throughout the Shetland Islands north of Great Britain. Though the exact num-

ber of speakers is unknown, the 2011 Scottish Census recorded 3500 Shetland residents who used Scots at home. A further 7500 of the islands' 22,000 population indicated that they could speak the language, but did not use it at home[1]. Unfortunately, these figures do not tell us much about the number of Shetland dialect speakers, as there are many Shetland residents who speak traditional Scots but not the local dialect. It is also possible that some respondents to the census instead recorded their native language as English, choosing not to identify with Scots as a separate entity.

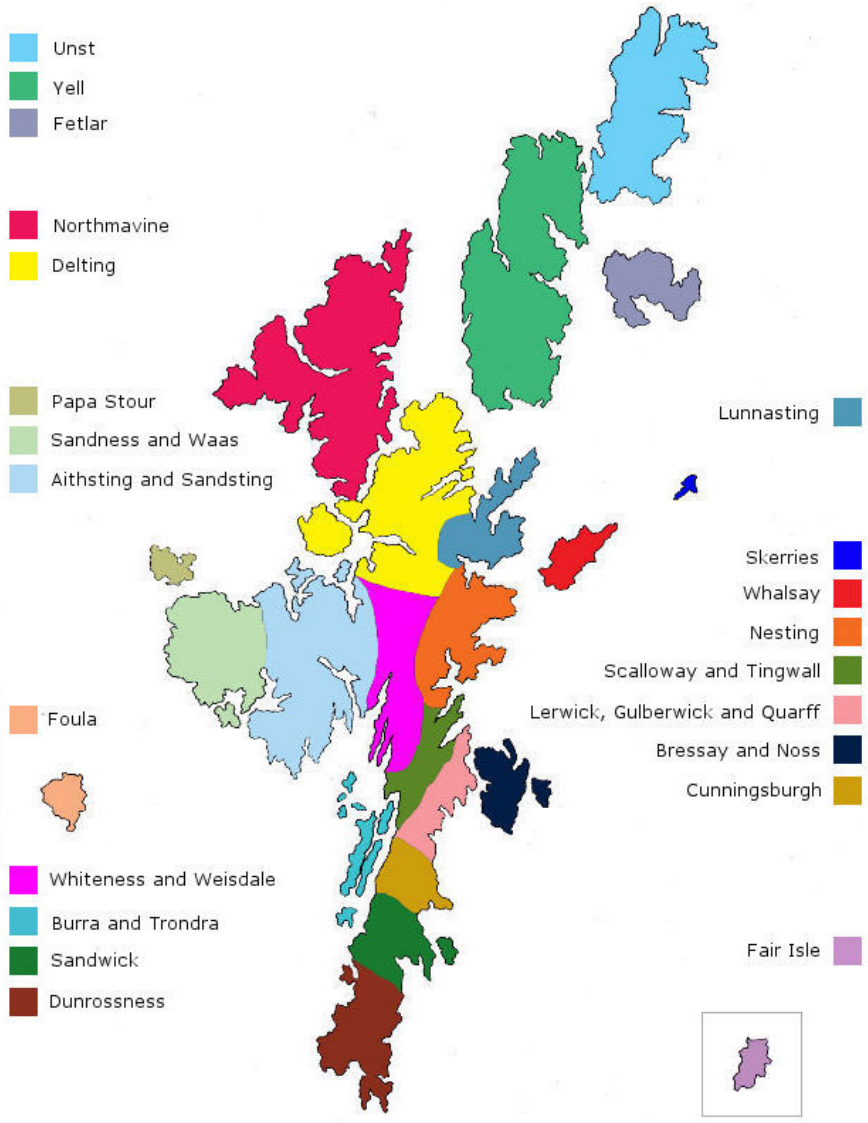

Figure 1.1: Dialect map of Shetland (Shetland ForWirds, 2022)

Though Shetland dialect shares much of its orthography with Standard Scots, it retains a rich vocabulary and varied writing system throughout the islands in which it is spoken (Graham, 1993). Local language charity Shetland ForWirds divides the isles into 22 parish regions on their website (see Figure 1.1), with examples of speech audio from

[1] https://www.scotlandscensus.gov.uk/census-results/at-a-glance/languages/

each of them (Shetland ForWirds, 2022). There are distinct differences in phrasing among the different highlighted regions, with similarities among the North Isles and strong contrasts among other regions.

Smith and Durham (2011) investigated the changing role of Shetland dialect in recent decades, considering the effects of the 1970s oil boom and subsequent changes to the local population. The study identified strong homogeneity among the use of dialect by older residents, with greater contrasts observed among the younger population. In a subsequent study, the same researchers built on their findings to identify a potential dialectal shift, in which the use of dialect was becoming more standardised over time and closer to that of Standard English (Smith and Durham, 2012). Native Shetland dialect speakers are also likely to exhibit code switching in day to day speech (Sundkvist, 2011), taking on a more standardised accent for English words which make up a significant proportion of day to day speech. This is referred to locally as "knappin" (De Luca, 2018) and tends to be more prevalent in conversations with non native speakers.

Shetland dialect speakers are proud of their local language and regard it as a strong source of cultural identity (Durham, 2014). Past initiatives to study and preserve local linguistic heritage are consistently met with strong support from the community and often yield fascinating results. As part of a recent drive to increase the number of languages available on their platform, Microsoft added Shetland dialect to the supported languages on SwiftKey, an assisted keyboard for touch screen devices (Shetland News, 2021) (Microsoft, 2017). The vocabulary of this system was based on a small corpus of spoken audio transcripts provided by the Shetland Museum and Archives.

As Sundkvist (2012) and many others have stressed, the loss of Shetland dialect would represent a significant loss to the local community, and a reduction in the diversity of language use in the British Isles. While language shift may be a natural process that represents changes in a culture over time, evolving communication technologies have the potential to further exacerbate this shift and lead to a premature extinction of endangered languages with limited resources. Designers of transcription tools for these communities should be mindful of the biases introduced by the interface employed.

1.2 Challenges

Despite robust community interest in the preservation and promotion of Shetland dialect, there continues to be a limited range of resources of Shetland dialect text and spoken audio, so most of the available content is largely dependent on a very small number of contributors. A good place to start investigating improvements to the existing resources is to design a transcription interface which makes it easy for Shetland dialect speakers to transcribe their own dialectal audio and add to the existing corpus. There are several challenges associated with this task.

A corpus which is representative of a language must be of suitable scale and express the linguistic diversity of the population. Wide participation in crowd sourced corpus building is dependent on accessible interfaces which do not require specialised knowledge to use and allow efficient submission of transcribed audio. In order to ensure

maximum benefit from the transcripts submitted by the local population, it is essential that the interface makes it as easy as possible to use accurate orthography which reflects both the user's style of writing and the speaker's use of the language.

In order to decrease effort associated with transcribing Shetland dialect, we may investigate the potential of ASR systems to do most of the work for us. After all, much of Shetland dialect is based on English words, and even more still on Scottish vocabulary. This task will involve building a language model and acoustic model to process an audio file and output Shetland dialect text. To make use of similarities with better resourced languages, we may start with existing models and fine tune them to better match the specific features of Shetland dialect.

Existing transcription interfaces are largely focused on well known languages, and few are designed to cater for languages with a non-standard orthography. As we know, ASR models are not perfect, and will need some correction, particularly if the transcript has the potential to vary by transcriber. We know that error highlighting can assist with this task, but also realise the potential for this feature to influence which words are chosen from the available suggestions. For dialectal transcription with a non-standard orthography, "error" words are much more difficult to define given the range of spellings which are acceptable in the language. Applied to the Shetland dialect, we might consider how geography affects the orthography that is used throughout the islands, in addition to speaking styles and contexts, and the speaker's overall relationship with the dialect. Previous work has highlighted a high level of dependence on an individual's linguistic heritage (Durham, 2014). All of these components influence the spelling and word choices that a Shetland dialect speaker uses to express themselves and should be considered carefully when designing systems intended for the collection of a representative corpus.

Before using an ASR system to increase the available corpus of transcribed audio, we must investigate how an interface which uses pre-generated transcripts to aid the task is likely to influence transcribers as they correct the text.

1.3 Key Contributions

We begin by collecting a corpus of Shetland dialect in Section 3.2 before using this data to construct a language model in Section 3.3. In Section 3.4.5, we design and implement a new cross-platform transcription interface which is then applied in Section 3.5 as we perform a user study evaluating the effects of error highlighting on the assisted transcription of Shetland dialect. We discuss the results of this study in Chapter 4 before concluding in Chapter 5 and recommending future work.

This project highlights the challenges associated with building a speech transcription interface for the Shetland dialect. We combine elements of natural language processing, sociolinguistics, interface design and data analysis to better understand the effects of interface on a non-standard orthography and suggest better strategies for maximising the value of transcriber input.

The main contributions of this project are

- a 275,778 token corpus of the Shetland dialect.
- a Shetland dialect language model.
- a cross-platform transcription interface which allows for maintained alignment, error highlighting, and inline text replacement.
- a study to evaluate the effects of error highlighting in the transcription of a non-standard orthography where ASR transcripts are used as a starting point.

Chapter 2

Background

In this chapter, we introduce automatic speech recognition as a tool for aiding in the task of transcribing dialectal audio. We then discuss corrective interfaces which can be used improve the usability of such tools before considering the potential influence of assistive technology in transcription tasks.

2.1 Automatic Speech Recognition

Modern automatic speech recognition (ASR) technology is capable of approaching, and in some cases exceeding human level performance on speech transcription tasks depending on the degree of supervision and domain familiarity (Radford et al., 2022). We evaluate performance on these tasks using the common metric of word error rate (WER), derived with the formula

$$WER = \frac{S+D+I}{N},$$

where S = substitutions, D = deletions, I = insertions and N = number of words in the reference text. This is equivalent to the rate at which a word from the transcript must be changed to match the original source. Character error rate (CER) is calculated in a similar fashion but finds error rate at the character level. We consider generalised human performance on conversational English speech to sit around 5-6% WER (Saon et al., 2017) (Stolcke and Droppo, 2017). Popular home assistants, such as Apple's Siri, Google Assistant or Microsoft's Cortana have been shown capable of matching this benchmark (Glasser, 2019).

2.2 Dialectal Transcription

ASR systems have made significant progress in recent years, but challenges remain in accurately recognizing and transcribing various dialects within a single language. Different dialects can have similar-sounding words, which can lead to increased ambiguity and decreased confidence in ASR outputs. Attempting to build a model which is

capable of recognising multiple dialects is likely to increase the model's WER when compared to a set of separate models. This was demonstrated by a 2012 study involving Arabic dialects in which a multi-dialectal model performed poorly compared to individually trained models (Biadsy et al., 2012).

The most successful approaches to this problem utilise a combination of multilingual deep neural networks and semi-supervised training with both manually annotated and automatically generated data (Yılmaz et al., 2018). For many small language communities however, this approach simply may not be feasible. Training a language model to consistently recognise speech requires vast amounts of data which may not be available to a low resource language community. A common approach is to develop a preliminary model using a small dataset, then use that to increase the efficiency of building a larger corpus. This method, known as bootstrapping, has been successfully employed in various studies on low-resource languages, allowing researchers to make better use of available data and improve ASR performance (Besacier et al., 2014).

Kaldi is an open-source framework for building ASR models using finite state transducers (Povey et al., 2011). The tool allows researchers to implement transfer learning by incorporating new vocabulary and speech features into an existing model. In 2022, researchers at the University of Edinburgh used Kaldi to produce an ASR system for Scottish Gaelic, a low resource language (Evans et al., 2022). By applying a cross lingual approach, in which they leveraged information from greater resourced languages, the researchers were able to significantly improve the performance of the ASR system.

While tools like Kaldi make it much easier for technically confident users to automatically convert large amounts of spoken audio to text, an accessible interface is required before wider participation can be expected. The Endangered Language Pipeline and Inference System (ELPIS) was designed in 2018 to allow non-technical users to build their own speech recognition models using Kaldi (Foley et al., 2018). The tool was developed to automate much of the ASR pipeline for users unfamiliar with the field and provide an interface for transcribing new audio with the resultant model. Furthermore, ESPnet, an end-to-end speech processing toolkit (Watanabe et al., 2018), has also emerged as a valuable resource in the ASR domain. The team behind ELPIS has since built on their original work by combining the software with ESPnet as an alternative to Kaldi (Adams et al., 2020). In their suggestions for future improvements, they highlight the value for an accessible front end user interface which would aid usability.

2.3 Transcript Correction

Tools like ELPIS make it much easier for proponents of under resourced languages to build and evaluate ASR systems. However, the transcripts produced are not perfect, and require correction before being used for further analysis. In 2001, a DARPA funded study (Suhm et al., 2001) evaluated a range of multi-modal text correction techniques including (in ascending order of correction accuracy): choosing from a list of alternatives, respeaking, handwriting, spelling and typing. The study found that users were most likely to prefer handwriting and following that, respeaking error regions. In the years following the study, both the ASR models and interfaces have changed

dramatically. Typing has largely remained the same, yet other methods have either been improved or replaced by far more intuitive interfaces. For example, multi-touch screens which are now ubiquitous and offer a range of new modes of human computer interaction.

In 2008, a new annotative interface was proposed by a team of Japanese researchers (Wang et al., 2008) that allowed transcripts to be corrected by marking error sections using a pen-like tool. For example, to substitute a word, the user would simply need to draw a circle around it and a list of alternatives would be suggested. The researchers highlight that this number must be low to avoid overcrowding the screen and putting too much load on the user, but they do not specify an appropriate limit. The research was further built upon in 2014 by two consecutive studies, the first using long context matching (LCM) and an n-gram language model (LM) to find substitutions for pen marked errors (Liang et al., 2014a) which was later revised to allow for insertions and deletions in the text by analysing the acoustic features of the error region (Liang et al., 2014b). Methods like these reduce the effort required from retyping error words by taking advantage of intuitive design patterns in touch screen interfaces.

In the context of assisted transcription, several studies have investigated the influence of an automatically generated starting point on the transcripts produced. Goddijn and Binnenpoorte (2003) compare human performance on both manual phonetic transcription tasks and tasks assisted by an automatically generated transcript. They highlight increases in efficiency from the assisted system but no significant disparity in transcriber agreement between the two tasks. This research indicates that the assisted system does not significantly hinder the quality of the transcription. However, the study focused only on phonetic transcription and may not be directly applicable to orthographic transcription.

Text correction interfaces have a substantial impact on the way we compose written content, as they play an active role in shaping our language and writing style. By offering suggested corrections or modifications, text correction interfaces can influence our choice of words and sentence structure, nudging us toward more standardized and widely accepted language conventions. A 2012 study into word alternative selection interfaces found that word suggestion systems would often encourage participants to select inaccurate substitutions when the correct text was not included in the list (Kolkhorst et al., 2012). Following this, a 2014 study found that these systems would also fail to produce correct replacements in up to two thirds of the errors encountered, further exacerbating potential inaccuracies (Harwath et al., 2014).

Following the results of these studies, Sperber et al. (2016) attempted to optimize retyping interfaces where the system learns from corrections made by the user to re-evaluate its confidence level. The researchers chose not to implement an alternative list selection interface, citing previous research. The team also suggested that the accuracy gains displayed by Wang et al. (2008) could be explained through the use of confidence highlighting, which they retain in their simplified interface without the more complex text selection features. In the context of non-standard orthography, these concerns become even more relevant, as the errors present in the text become much more difficult to define.

Gaur et al. (2016) also investigated the effects of varying WER of automatically generated transcripts to observe changes in correction efficiency, they found that transcripts with a WER of 30% or higher increase the amount of work required to correct it to the point that it would be better for the transcriber to start from scratch.

2.4 Influence of Assistive Text Entry

Despite their utility, assistive text entry tools present several drawbacks which should be considered alongside their obvious advantages. For example, further consideration is being applied to the influence they have on the way we think and communicate. In a 2018 Harvard study (Arnold et al., 2018), researchers investigated the effects of predictive text on the sentiment of restaurant reviews, demonstrating a significant positive bias when participants were offered more positive words and the opposite when offered negative words. The effects of predictive text on language are of particular relevance for our purposes as they demonstrate how potential phrasings offered to the user influence their style of communication.

Following their 2018 study, Arnold et al. (2020) compared a range of text suggestion systems in the application of image captioning. The study found that captions written with suggestions were not only shorter, but used more predictable and less descriptive words. In many ways, this uniformity reduces the value provided by human contributors working in tandem with an automatic system. Despite an increase in text entry speed for most users, faster typists conversely experienced a reduction in text entry speed. This means that the predictive text is not only reducing the value of input from these contributors, but slowing them down in the process.

In addition to the efficiency gains provided by assistive technology, we must consider the impacts it may have on the value provided by human contributors. In a low-resourced language community, the time provided by human participants is highly valuable and care should be taken to ensure that the maximum amount of information is being extracted from their contributions. It is crucial to strike a balance between leveraging the benefits of ASR in these contexts and maintaining the accuracy of manual transcription.

Chapter 3

Methodology

3.1 Outline and Scope

This project follows the development of a corpus of Shetland dialect, training a language model, designing a cross-platform transcription interface and a study investigating the effects of interface on the transcription of Shetland dialect where ASR generated text is used as a starting point.

The corpus used in this project was composed of work from several writers and contains no material originating from the the author (disregarding minor adjustments). Alignment and normalisation of the corpus were completed by Peter Bell, the supervisor for this project, using an existing system detailed in Section 3.2.2.

The development of an acoustic model for Shetland dialect is handled by another student completing a separate and complementary project focusing on ASR tools for Shetland dialect. The corpus described in Chapter 3.2 is shared between the two projects and the results from the incorporation of the language and acoustic models in Kaldi is used in part for the production of reference texts for the study detailed in Section 3.5.

The interface developed as part of this project represents a simple prototype which reduces the range of actions we may expect from the user. While we consider design choices which are likely to improve our user interface (see Section 3.4), the design used in our study is not intended to represent an interface we would expect to see in a final product.

The primary focus of this project is to facilitate the investigation into the effects of error highlighting on the transcription of non-standard orthographies. We use Shetland dialect as an example of one such language. Given the potential of auto-generative systems to alter the way in which we communicate, we intend to investigate how susceptible Shetland dialect might be to such bias.

While many of the results of this project advance the production of ASR tools for Shetland dialect, we do not propose any software capable of end-to-end transcription. We discuss the future work required for such systems in Chapter 5.

3.2 Corpus Development

3.2.1 Collection

As Bird (2021) notes in a 2021 article on sparse transcription, transcripts for low resourced languages are rare for a number of reasons. Mainstream resources such as TV, radio and newspapers are unlikely to employ specific dialects given the prioritization of understanding among a wider audience. The small number of dialectal spoken audio resources that are produced are also more likely to be translated to a standard form, as these are more practical for storage, searchability and distribution.

Given a lack of centralised public corpora for the Shetland dialect, a broad range of texts have been collected and normalised for the purposes of the study. There has been one past attempt to build a corpus of text from Shetland and Orkney, with resources identified for this purpose but unfortunately did not reach fruition (Ljosland, 2012).

The corpus built for the purposes of this project consists of two parts:

- Spoken audio files and accompanying text transcripts - A set of existing audio files recorded and stored over several decades by the Shetland Museum and Archives. These recordings have since been orthographically transcribed by Angus Johnson, a member of the Archives staff. Angus primarily used a typewriter to record these transcriptions.
- Text files - A larger corpus consisting of several publicly available resources in addition to a collection of works kindly provided by Christine De Luca, a local writer.

The spoken audio corpus has an original total duration of 22 hours; however many of the associated transcripts are missing or the speech recordings are not of sufficient quality for training a speech recognition model. With some processing, we arrive at a final total of approximately 5 hours of transcribed audio data. The total word count for this segment is 47,539 words.

The written corpus features a collection of texts sourced from the internet and individual requests to dialect speakers and writers. The word count of this segment (not including the transcriptions of the audio) is 228,239 words from 9 different writers. Therefore we arrive at a final word count of 275,778 words (see Figure 3.1). This is larger than any publicly referenced corpus of Shetland dialect text that is currently available, and is expected to grow over time as several writers have pledged further textual resources. Though it is of a scale similar to previous efforts relating to Nordic languages which have served productive research (Johannessen et al., 2009), the corpus is very small compared to what we might expect to need for building an ASR system.

Having ensured there is no repetition in the corpus (through programmatic utterance search), we are interested in the extent to which the contributors to our corpus share vocabulary. We generate a set of the 50 most common words in each writer's text and plot a confusion matrix of the length of the intersection between those sets (see Figure 3.3). As we might expect, larger contributions tend to share more vocabulary. Anecdotally, there is some correlation between geographical origin and vocabulary,

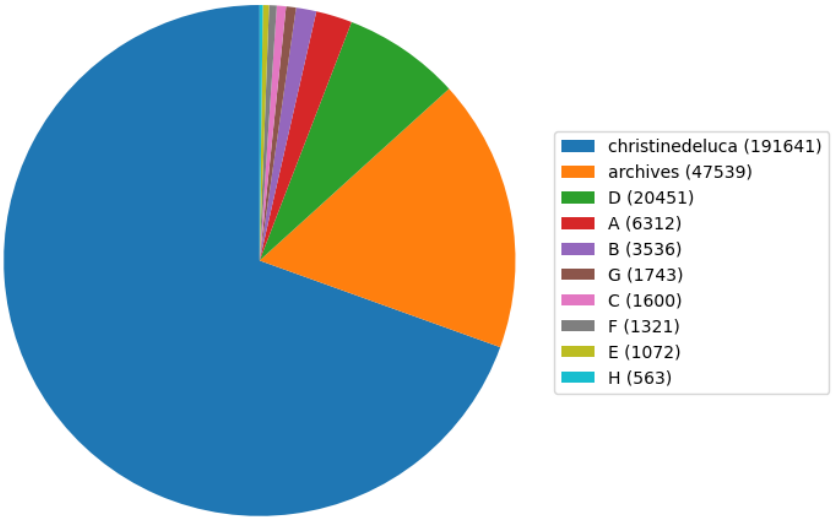

Figure 3.1: Distribution of writer corpus sizes

but given the lack of data for some contributors, further analysis was not possible at this stage.

Our corpus has a total vocabulary of 15,644 (including English words) across all sources. When we remove the English words from the corpus, we find a total of 6590 words which only appear in our corpus. Manual inspection of the removed words identifies 372 words which have a different meaning in Shetland dialect than in English (for example “raisin”), though there are likely to be some that were missed. The final total of unique Shetland dialect words is 6962. This is believed to be the largest collection of unique Shetland words in modern usage not including English.

3.2.2 Alignment and Normalisation

Alignment, tokenisation and lexicon development of the corpus were completed by the project supervisor, Peter Bell, using an existing system (Bell and Renals, 2015) designed to accommodate low quality audio of spoken English. The assumption here is that Shetland dialect shares enough similarities to the orthography of Standard English that the gaps can be filled in for the unique vocabulary of Shetland dialect. This process also divided the corpus into a collection of shorter utterances with timestamps to reflect the clip of audio where they could be found.

A pronunciation dictionary was assembled from the data with the phonetic structure for each word. Many of the words in the lexicon were produced purely from their written form in a process known as grapheme-to-phoneme (G2P) conversion. In this case, a joint multigram model (Cao et al., 2012) was applied to capture the relationship between graphemes (letter sequences) and their surrounding context.

To ensure consistency, we apply the same tokenisation and normalisation process to the pure text part of the corpus in addition to the transcripts. As part of the normalisation

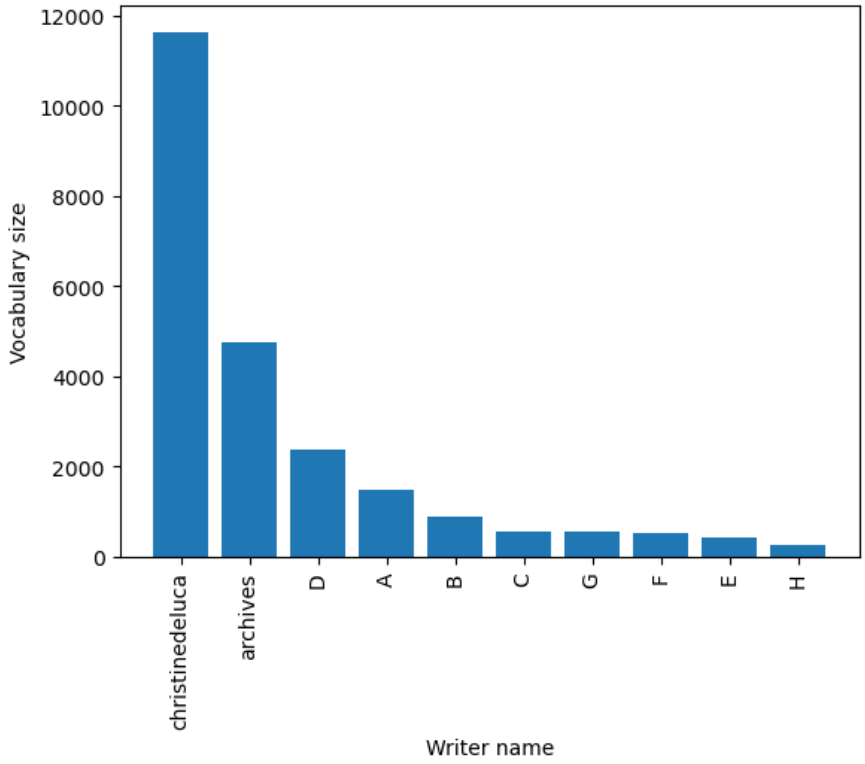

Figure 3.2: Vocabulary size for each writer's corpus

process, diacritics are removed from the text. Use of diacritics varies throughout our corpus. For example, De Luca uses the "ö" (umlaut o) symbol extensively in her work, however Johnson does not include it in his transcripts. Later conversations with the authors reveal that Angus may have used diacritics in his work if they had been available on the typewriter he used to construct the original transcripts. This is another example of how transcription tools affect the text produced.

While traditionally, a suitable replacement for "ö" might be "oe", this is not reflected in Shetland dialect. For example, Graham (1993) presents the Shetland word for "above" as "abön" due to it being a modified "o" sound. Another acceptable spelling would be "abun" , but subtleties in the text make it difficult to apply these changes consistently and without bias throughout the corpus. We chose to simply replace diacritics with the plain text version of the character, for example, the letter "ö" is replaced with "o".

Some further steps we apply are removing erroneous stops from the Shetland word "o" (English: of) which have been mistaken for initials and converting the entire corpus to lowercase. Apologetic apostrophes (e.g. endin') are also removed given their sparsity in the data.

3.3 Language Modelling

In this section, we train and evaluate a language model for the Shetland dialect. The experiments detailed in this section were in part performed for the purposes of provid-

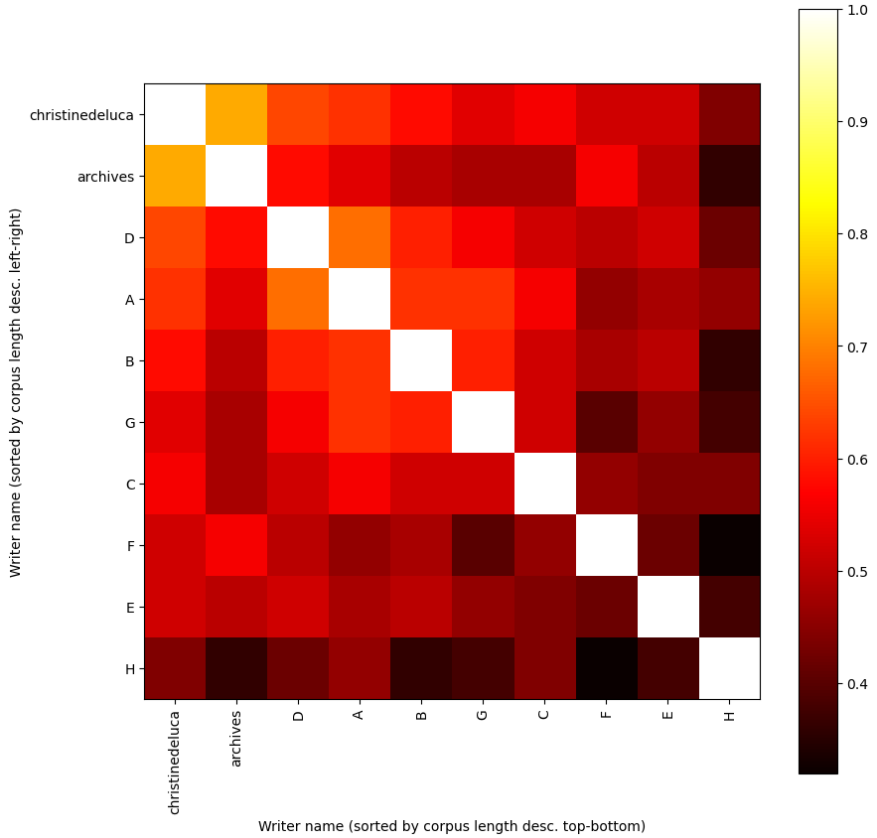

Figure 3.3: Overlap rate between 50 most common words in each writer's corpus

ing data to a complementary project (see Section 3.1) and are considered preparatory to the rest of the methodology. Transcripts provided by that project are then used as part of the study detailed in Section 3.5.

3.3.1 Approach

Given the small size of the corpus, careful consideration must be applied when choosing a language model to ensure optimal performance despite data limitations.

Our approach utilises both an existing language model trained on a large collection of English speech and a new model trained only on our Shetland dialect corpus. We choose to interpolate our model with a Librispeech trigram model, trained on a large collection of English text (Panayotov, 2014). This approach vastly improves the model's overall performance, particularly given the amount of vocabulary shared with Standard English. The Librispeech model is made up of a collection of 14,500 public domain books, giving it a broad vocabulary and range of contexts.

In a complimentary project, another student combines this Librispeech model with an American English acoustic model to obtain a WER of 71.55% when tested against the archives transcripts. This model contains no Shetland words, so the high WER is unsurprising. We may consider this result as our baseline.

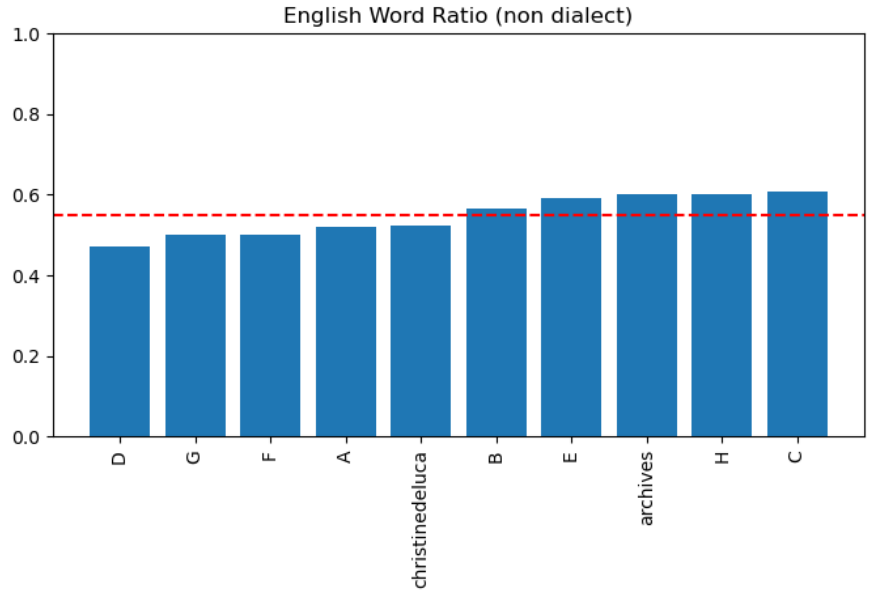

Figure 3.4: Ratio of English words to dialect words per line of the corpus

We use SRILM's N-gram tool[1] to train a trigram language model with Kneser-Ney (KN) smoothing on the Shetland dialect corpus (excl. transcriptions). KN smoothing reallocates probability mass from observed N-grams to unobserved ones by discounting the probability distribution of lower order N-grams (Kneser and Ney, 1995). This improves the model's performance when it encounters unseen word sequences, as it is particularly likely to with spelling variations and alternate phrasing in Shetland dialect. Unlike Additive or Laplace smoothing which simply add an additional constant to existing sequence counts, KN smoothing takes the context in which words appear into account. This should result in more accurate probability estimates. If we take the word "den" from the Shetland greeting "Noo den" (English: Hello/Now then), we are likely to find a high unigram probability for the common English definitions of the word (the home of certain wild animals, a type of room, a rough structure or a location for illicit activities) (Cambridge Dictionary, 2023). KN smoothing helps to re-attribute some of this probability mass to phrases which include the word "den" in Shetland dialect.

3.3.2 Evaluation

We train the Shetland dialect model only on the pure text part of the corpus and evaluate the model on the transcript data. This strategy is intended to reduce the effects of overfitting in the model. Unfortunately we sacrifice some generalisability here as the archives and wider corpus represent different writing styles and vocabulary. Given that the pure text corpus is largely derived from written materials largely from a single source, we can expect it to differ from the spoken archives corpus.

The interpolation scale (λ), decides the proportion of weight assigned to each of the models we are combining. By optimizing λ, we can strike a balance between the Shet-

[1] http://www.speech.sri.com/projects/srilm/manpages/ngram.1.html

land dialect and Librispeech models, ensuring that the interpolated model captures the characteristics of Shetland dialect while benefiting from the much larger Librispeech dataset. To identify a suitable weighting, we generate a range of models with varying λ weights and compare the perplexity against a test set. Perplexity is a measure of how well the model predicts the test data, with lower values indicating better performance. It is calculated using the following formula:

$$P(W) = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 p(w_i|w_{i-1})}$$

Jurafsky and Martin (2019) note that perplexity is very close to the concept of entropy, as it describes the average uncertainty or unpredictability of a language model in predicting the next word in a given sequence. When comparing the perplexity of two models, we must ensure that the vocabulary of both models is the same as the perplexity score is sensitive to the size of the vocabulary, and models with larger vocabularies may inherently have higher perplexity values due to the increased number of possible word choices.

Evaluating our interpolated models against the transcriptions, we identify the best weight to be $\lambda = 0.63$ (where 63% of the probability distribution is assigned to the Shetland model) with a perplexity of 489.6. When this model is incorporated into the Kaldi decoder (built as part of the complementary project mentioned in Section 3.3.1), we obtain a WER of 70.99%, representing a very small decrease in WER compared to the baseline model (0.56%). By scaling the acoustic model factor, WER is decreased to 70.85%.

Given that the Librispeech model contained no Shetland dialect vocabulary, we would expect a much more significant decrease in WER with the introduction of a representative lexicon. Besides this fact, the reduction in perplexity on the test set indicated a much more significant difference than was demonstrated in real world performance. The main conclusion we draw from these results is that a larger corpus is necessary in order to improve the performance of our model. The decoder is likely to have benefited from an acoustic model trained on Shetland dialect audio, however a limited audio corpus may have restricted the effectiveness of this strategy. These experiments are outwith the scope of this project.

We explored the effects of including some of the lower quality transcriptions in a trilinear interpolation of the existing dialect model, the Librispeech model and a further trigram trained on the transcripts (with the same paramaters as the pure text model). When we interpolate these models, we find that the ideal weighting is 20% Librispeech model, 20% pure text model and 60% speech transcript model, resulting in a perplexity of 429.5 when evaluated against the held out transcripts. This roughly follows our expectations in terms of the interpolation weights, as we see most of the information is taken from data with the same transcriber. However, the perplexity of the resultant model is only 12.3% lower than our linearly interpolated model. It is not known to what extent this increase in performance is the result of overfitting. Our primary conclusion from these experiments is that the model would benefit from being trained on a much larger collection of spoken text.

In another exploratory experiment, we attempt to further standardise the vocabulary presented in the corpus. Though many of the writers follow the spelling style prescribed by Graham (1993) (confirmed in conversations with both Angus Johnson and Christine De Luca regarding their own contributions), there remains a diverse range of spellings throughout the corpus. For example, we find at least six different spellings for the English word "because" throughout the corpus, three of which are represented in the archives transcripts. These alternate forms are dependent on both the transcriber's spelling choices and the phonetic variation of the speaker. We experiment with identifying clusters of words with low phonetic edit distance and making replacements based on context, but this does not yield any significant improvement to language model performance. The strategy also struggles to account for modified standard English words and other ambiguities.

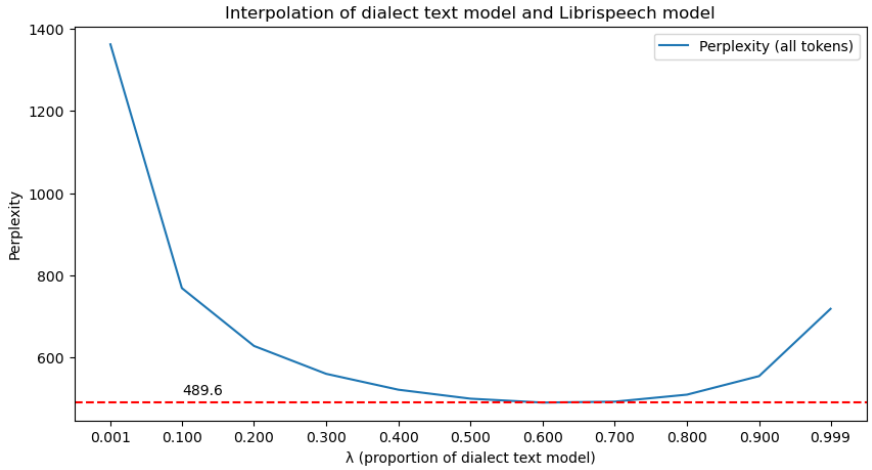

Figure 3.5: Evaluating perplexity of models trained with different λ weights (interpolation scale)

3.4 Interface Design and Implementation

3.4.1 Review of Existing Transcription Interfaces

Before building our own transcription interface, we consider several existing apps in the same domain, as listed on a Tech Radar (2023) compilation of the most popular transcription apps and the top listings on the Play Store and Apple's App Store. These include "Transcribe - Speech to Text" by Transcribe.com, "Otter: Transcribe Voice Notes" by Otter.ai, "Temi - Recorder & Transcriber" by Rev.com.

There are several common attributes of the app interfaces on this list:

- Error indication with text colour/highlighting - Several of the apps utilised colour to indicate words which the system considered potential errors.
- Audio playback speed - All of the apps reviewed allowed the user to vary the playback speed.

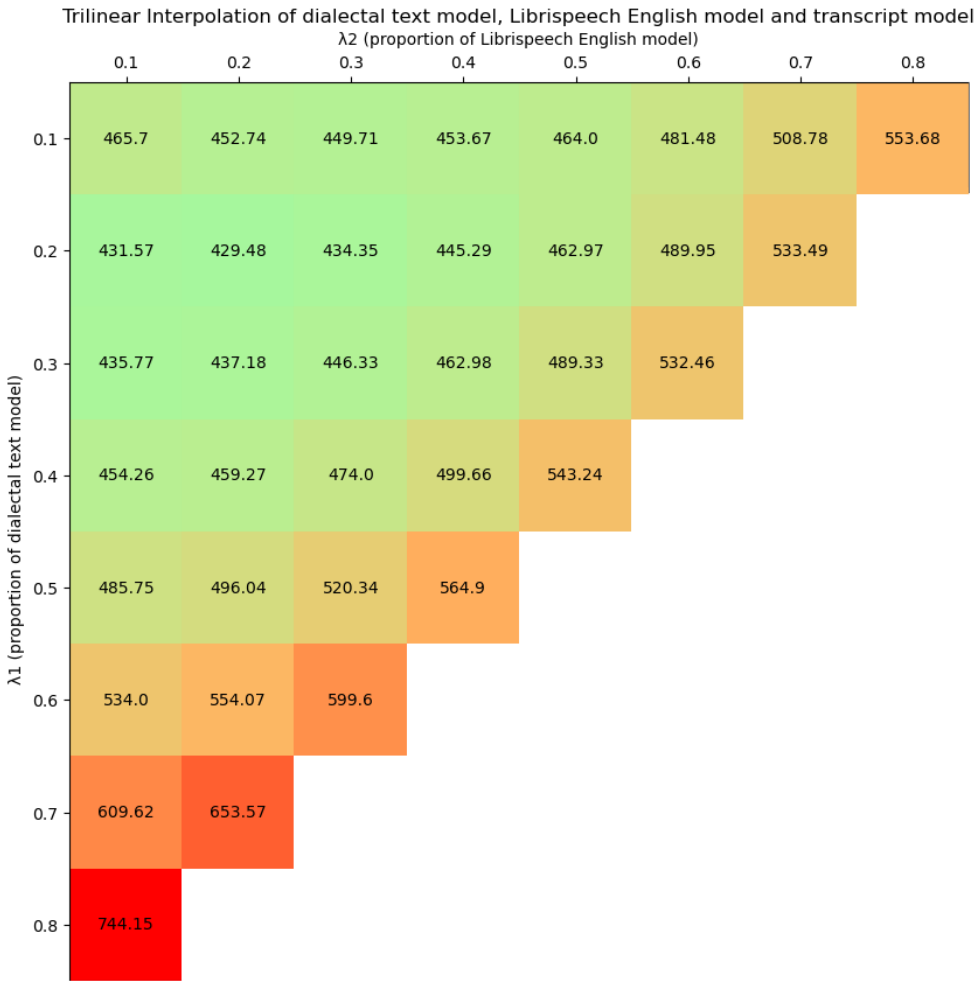

Figure 3.6: Evaluating perplexity of models which have been combined using trilinear interpolation (low values are better). λ1 is the mixing value of the dialect text trained model. λ2 is that of the Librispeech model. Not shown is the mixing value for the transcript trained model (1-λ1-λ2).

- Clip position slider - All of the apps allowed the user to choose the playback position on a slider.
- Clip jump (10-15s) button - Several of the apps allowed the user to skip back a set amount of time to a slightly earlier part of the audio.
- Segmentation and timestamps - Several apps divided audio input into segments which could be edited separately.

We can use these observations to evaluate the design decisions of our interface. Live transcription, accent specialisation and speaker differentiation are related to the app's ability to process new audio which is outwith the scope of this project so we disregard these.

Many of the apps used text colour to indicate potential errors in the transcripts produced. While this is a simple and eye catching method of drawing attention to parts

of the text, it is possible that people with colour blindness may struggle to distinguish these words. Text highlighting, bold text and underlining are other methods which differ visually beyond their colour.

We choose not to include a playback speed button to potentially allow more consistent review of how quickly users complete a transcription task.

Given the shortness of the clips of audio we intend to present potential study participants, it seems there is little need for a clip position slider or clip jump button. They are also likely to increase clutter on the screen which may cause confusion for some users and further challenges in the analysis of user actions. In a real world implementation, we would expect these features to be available so we should design the app with the expectation that these features be introduced in the future.

Audio segmentation is a feature which was present in all of the interfaces reviewed. The feature separates parts of the text based on speaker changes or pauses in the audio. Again, given the shortness of the audio clips intended for our study, this feature may not be necessary and could serve to increase on screen clutter. We would also like to ensure users have as much context as possible at any given time to ensure consistency throughout their submission. We choose to implement a more concise segmentation through text highlighting. This is a slight deviation from the norms of transcription interfaces so the instructions will need to explicitly detail this feature.

3.4.2 Requirement Specification

We outline the functional requirements for the transcription interface as follows:

1. The interface must accept a "sausage" like structure represented as a 3D array, where subsections of the text are represented as lists of lists of alternate word suggestions in addition to the associated audio clip start and end times. Also included in this case, is the original index of the segment (used only for logging purposes). For example, if we consider a possible output from a system for the (deliberately ambiguous) input "Whar's du fae den, Mark?" (English: Where are you from, Mark?):

   ```
   [[[["ores ", "whars "], [0, 1000 (clip end)], [0]],
   [["due "], [0, 1000], [1]],
   [["fae ", "fame ", "flame "], [1000, 2000], [2]],
   [["denmark"], [1000, 2000], [3]]]]
   ```

   Each row here contains the word present in the text (with the alternative suggestions following that), the audio clip start and end times and a singleton array with the index of the segment. When flattened to plain text, this example would evaluate to "ores due fae denmark".

2. Words with more than one option, in this case "ores" and "fae", should appear as dark blue underlined text of the first word in the list. Words with only one option should be represented as plain text.

3. When plain black text is edited, the system must retain these changes.
4. When a dark blue word is long pressed (i.e. the user places their finger on the word and continues to hold for a few moments), a popup menu should appear near the word with the other options listed. When one of these options is tapped, the text should be replaced in the passage and become plain text.
5. When the user deletes the space at the end of a blue highlighted word, the word should change to plain text.
6. All words with the same clip times should be highlighted with a semi-transparent light blue colour when one of these sections has been selected. When the app first loads, the first section should be highlighted, in this case "ores due".
7. A play button should play back the clip of audio for the section currently highlighted.
8. A help button should allow access to the instructions at any point during the task.
9. All aspects of the interface must function consistently on both iOS and Android devices.

There are also several qualitative requirements which should be taken into account, mainly related to accessibility:

1. All touch targets should be larger than 40 pixels.
2. All text should be easily readable on all screen sizes.
3. All aspects of the interface must function consistently on both small screens (smartphones) and large screens (tablet/laptop).
4. Colours with low contrast should not be used next to one another (or red and green).
5. The interface must be usable in both portrait and landscape mode.

3.4.3 Low Fidelity Prototype

We incorporated the design specifications into an early design (see Figure 3.7) using Figma, a common interface design tool. This allowed us to get early feedback on an initial concept and identify potential issues. The design was intentionally made simple to ensure focus was placed on core functionality and to encourage open and honest feedback (Krug, 2014).

Our first concept for the design involved the inclusion of the play button as part of the suggestion highlighting display, merging playback with error correction. Following further consideration, we moved the play button to the top of the screen so that it is always visible throughout the transcription task. The initial design would have constrained playback to very short sections of text, potentially inhibiting transcribers from hearing the full context of a given phrase.

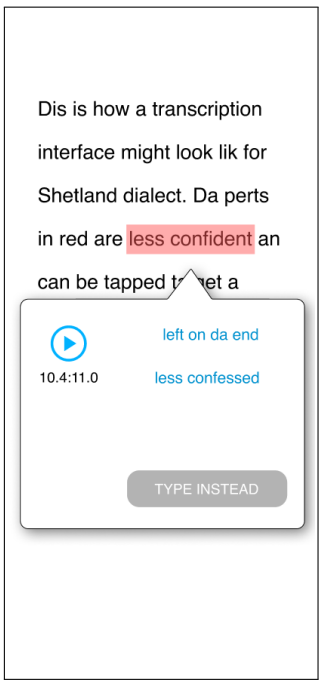

Figure 3.7: The initial mockup of the interface, designed in Figma.

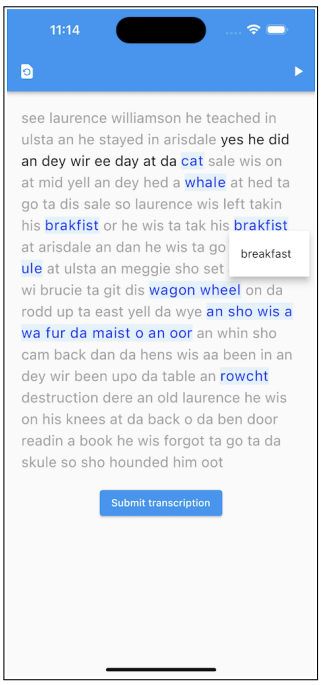

Figure 3.8: The first interface design implemented in Flutter.

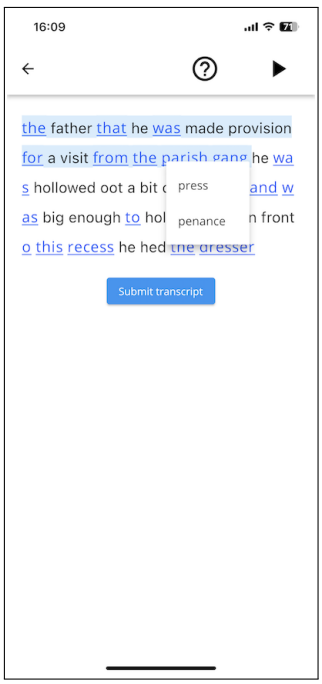

Figure 3.9: The final interface design used in production.

We also removed the "type instead" button from the suggestion panel. Users can be expected to manually delete text if none of the options are suitable as this is a standard design pattern for toolbar based text editing (Cooper, 2014).

The final change from the initial design is the colour choice used for highlighting. Though red is typically associated with errors, it draws a significant amount of attention which may cause unintended influence on the transcriber. We also needed a method of indicating errors in the text, and large amounts of red text may have been overwhelming. We chose dark blue as a much calmer colour with less contrast to its surroundings.

The interface design used in production includes a help button at the top of the screen and increased touch target sizes. We also made interactive words dark blue while playable sections highlighted with a semi transparent light blue.

The implementation of the text editing widget followed a test driven development (TDD) approach. Several example texts were taken from the audio corpus and converted to the 3D array format expected by the interface. We designed several tests for both the deserialisation of JSON transcripts and modification of text within the widget itself. These tests ensure that the original indexing and audio clip data is maintained throughout the editing process and no unexpected changes are made to the text itself given the possible interactions with the interface. The app was initially implemented as a standalone interface for the purposes of designing the underlying structures and mechanisms. When all tests were passed and the requirements detailed in Section 3.4.2 had been met, we ported the app to a production environment for a pilot test

with a small number of users who were unfamiliar with the project. This helped us to confirm the assumptions made in the original design and identify any remaining issues. One such finding was that users encountered some confusion with the utterance highlighting feature (we discuss this further in Section 3.5.3).

3.4.4 Use of the Shetland Dictionary app

In 2022, Shetland ForWirds released the Shetland Dictionary app (Shetland News, 2022) with the goal of further preserving and promoting Shetland dialect. Given an existing user base of predominantly Shetland dialect speakers (further discussed in Section 3.5.2), the app makes an excellent candidate for supporting the transcription interface.

The app was built using Flutter, an open source framework developed by Google for cross-platform app development. Flutter allows for the simultaneous development of iOS, Android and Web apps with a single code base[2]. The Shetland dictionary is currently supported on iOS and Android. While it would also be easy to publish the Shetland dictionary as a browser accessible website for the purposes of this study, the closed environment of a smartphone allows for greater control over external factors such as clipboard access and web browser versioning.

Flutter's development ecosystem has fostered robust support for accessible interfaces and it is easy to import plugins to expedite accessibility guideline conformance on all supported platforms. The `Semantic` widget can be used to better accommodate screen readers for example [3]. By implementing the transcription interface in Flutter, future work can benefit from increased options for improving accessibility.

3.4.5 Implementation

Our first step was to verify that the components available in the Flutter framework were capable of supporting the requirements detailed in Section 3.4.2. We found that there were a limited range of plugins that allowed interactivity with editable text, however these all relied on regular expressions (regex) to define interactive sections of text. These plugins are intended for chat applications where users may want to share web links or make user handles interactive. The regex that defines these elements is very simple and is therefore prone to some unexpected behaviour when certain characters are deleted or the opening/closing tags are used unintentionally.

Regex also makes it difficult to keep track of different sections of the text. If one highlighted section is deleted, it may affect another section or change the expected ordering of the highlighted words. These issues are less relevant in the applications where they are expected to be used, for example spelling correction. However, in our case, we needed to keep track of separate context dependent text replacements throughout the text which cannot be confused with one another. Parsing techniques whereby highlighted text was surrounded by tags e.g. `Lorem <s i="5">ipsum </s>dolor`

[2] https://flutter.dev

[3] https://api.flutter.dev/flutter/widgets/Semantics-class.html

proved unsuccessful as false equivalencies in Flutter's keybindings resulted in unexpected caret movement. These issues are tracked in an open issue on Flutter's repository[4] but are not expected to be resolved until after this project's completion. With existing options deemed unsuitable, we defined a new algorithm which could handle our requirements.

Flutter `TextField` widgets are stateful and keep track of both the displayed version of the text (with fonts, colours and stylings) and a simpler string only variable. The widget is rebuilt every time a user interacts with the text field. We propose an algorithm which feeds back changes within the displayed text to the underlying data structure so that the state could be maintained (see Algorithm 1). We rely on the assumption that only one part of the text is edited at any given moment. To ensure this, we implement a new extension to the `EditableText` class which the `TextField` is based on. This allowed for restrictions to be placed on the toolbar which would prevent unexpected changes resultant from partial selection of the text and subsequent use of the clipboard.

These design decisions allowed for rapid implementation of a Flutter widget capable of highlighting various sections of the text based on the data structure. Unfortunately, problems arose from attempting to interact with these highlighted sections given that they could be simultaneously edited as text. A solution where highlighted text is represented as tappable buttons presented similar issues to the parsing techniques previously explored. Finally, we identified a novel solution in which individual characters are expressed as a series of interactive `InlineSpan` widgets within an array of `TextSpan` widgets. Not only does this resolve conflicts between the display and data structure, but also enables line wrapping to function as expected. This is the first Flutter based interface allowing consistent interaction with editable text without relying on simple Regex formulae.

3.5 Study Design

3.5.1 Approach

Having developed a suitable transcription interface, we are interested in the extent to which the corrections made by dialect speakers on an existing transcript are influenced by the suggestion highlighting.

We hypothesize that participants will be more likely to replace an English word with the dialectal equivalent if the word is highlighted in the text. To test this hypothesis, we design an experiment which investigates the influence of error highlighting on the spelling choice of Shetland dialect speakers when correcting an existing transcript.

We generate a set of transcripts using the ASR system constructed as part of the complementary project (see 3.3). Given that the error rate of these transcripts is very high, we replace many of the errors with the correct version of the word to reduce the WER of the passage to approximately 30%. This rate was chosen as it allows us to assess the effectiveness of the transcription interface without overwhelming participants with

[4] https://github.com/flutter/flutter/issues/34688

Algorithm 1 Get Modified Text Array

Require: String *newText*
Require: $oldTextArray = [[[[textalt_1, \ldots], [start, end], [originalIndex]], \ldots], \ldots]$
Ensure: Modified text array
oldText $\leftarrow$ join elements in oldTextArray
if newText $==$ oldText **then**
    return oldTextArray
end if
if oldTextArray.isEmpty **then**
    Log the change
    return singleton text array with newText
end if
Find changeStartIndex and changeEndIndex
if changeStartIndex $== -1$ **then**
    Handle text added to the end
    Log the change
    return newTextArray
end if
if changeEndIndex $== -1$ **then**
    Handle text added to the start
    Log the change
    return newTextArray
end if
Find unchanged sections before and after modified section
sameStart $\leftarrow$ unchanged sections before change
sameEnd $\leftarrow$ unchanged sections after change
Build newTextArray with changes and sameStart and sameEnd
Remove empty sections from newTextArray
Log the change
return newTextArray

an excessive number of errors (Gaur et al., 2016). This means the average time to complete the task will be much lower and we can expect a much higher rate of participation. We select transcripts with a high number of English words which should be dialect words (for example, "another" when the word spoken is "anidder"). Despite these changes, we are in effect evaluating the reaction to errors demonstrated in a real ASR system. Though the surrounding context of each word has been artificially corrected to more closely match the ground truth, we might consider this to be a transcript with unfinished corrections requiring a native speaker to finish it.

Two variants are produced for each of these transcripts, one with highlighting on the words we intend to check, and the other with plain text. Both transcripts contain other errors (highlighted and not highlighted) alongside those which are being monitored. By comparing the two sets of completed transcripts, we can evaluate the extent to which this highlighting affects a participant's word choice.

The English words (and suggested dialect words) which are being compared are: *and (an), another (anidder), from (fae), like (laek), that (dat), the (da), there (dere), they (dey), this (dis), to (ta), was (wis)*. Of particular interest here is the word "from" which is phonetically very different from the Shetland equivalent present in the audio "fae". We might expect that Shetland speakers will translate the word in their head on inspection of the plain text version, and the highlighted version will help draw attention to the error. The rest of the words were chosen as they are phonetically similar to their English counterparts and are very common throughout the text.

We also include a transcript where the word "of" is replaced with an "o" in the text. This will serve as a control variable to see how often an erroneous dialect word is replaced with the English version heard in the audio.

The null hypothesis is that there is no significant difference ($p >= 0.05$) between the use of these words if they are left as plain text or highlighted. Given the expected small sample sizes, we will use Fisher's Exact test to validate the significance of our findings (Fisher, 1922).

3.5.2 Participants

The participants of the study are native speakers of the Shetland dialect. Given that the study is contained within the Shetland Dictionary app, we can make several assumptions about participants:

- They are more likely to be Shetland dialect speakers - According to feedback on social media, the app is mainly used by Shetland dialect speakers to look up vocabulary. Since its launch, the app was promoted largely by organisations in Shetland and local news. We can also confirm the speaker's linguistic identity in the survey following the submission of their corrected transcript.
- Interest in Shetland dialect - At the time of publishing the study, the app had around 1000 users. This represents around 4.5% of the population of Shetland[5]. We might assume that this demographic is composed primarily of those most

[5]https://www.scotlandscensus.gov.uk/census-results/at-a-glance/languages/

interested in the use and preservation of local dialect, and therefore more likely to participate in the study.

- Confidence with touch screen technology - Before downloading and using the Shetland dictionary app on a compatible device, we may assume that the participant is accustomed to touch screen interfaces and is likely to understand the instructions for the task.

The study was included in the side menu of the app. Participants were not prompted to take part in the study through any direct notifications. Instead, the study was promoted on social media by the author and a local dialect charity with detailed instructions on how to access the interface. This decision was made partly to reduce burden on the app's users by ensuring voluntary participation, as well as encouraging only a specific group of users to take part. Despite a locally concentrated user base, the app is also used by an international audience. Care was taken to word any promotional text consistently and avoid sharing images of the interface beforehand so as not to influence how participants used it.

Comments on social media indicated a general enthusiasm for the further incorporation of Shetland dialect into technology and the study was shared widely by many locals with whom the author had no prior connection (Appendix A.1). Though not instructed to submit multiple times, many participants submitted several transcripts and indicated continued interest in the project.

3.5.3 Experimental Procedure

Upon accessing the study page, each participant is provided with a short summary of the participant information sheet along with a link to the full document (Appendix C) and is asked to complete the consent form upon reading this. Having agreed to the conditions of participation, the instructions are then made visible. When the participant has confirmed that they have read the instructions, a randomly assigned transcript is revealed with the first section of the text highlighted.

The instructions ask that the participant find a quiet place to complete the transcription task, recommending headphones be used. They are then asked to correct the transcription on the following page to match the associated audio "as closely as possible". Note that this does not specify that the participant correct the transcript to match the speaker's pronunciation nor any specific set of spellings. The same instructions are available via a help button on the transcription page.

The interface is explained in detail, but does not provide a tutorial or an example video of the interface being used. This was intended to reduce the amount of time required to participate in the study and allow basic assessments to be made on the intuitiveness of the interface.

By retaining explicit details of the experimental procedure, we minimize the risk of participants modifying their behaviour (consciously or subconsciously) to align with the goals of the study. This ensures that the information collected accurately reflects the natural behaviour and decision making throughout the transcription task. As the

user edits the text, we log character deletions, insertions, suggestion views, phrase substitutions, audio plays, and help button presses. We do not log caret movement as we do not expect this to vary significantly between transcripts.

Chapter 4

Results and Discussion

4.1 Exploratory Analysis and Processing

The study received a final total of 119 submissions from 78 unique participants over the course of a week. Of the 22 available dialectal regions, we find 15 of these represented in the study (see Figure 4.1). We see that a large number of participants (23%) identify with the regional dialect of the island of Yell. This primarily stems from the fact that the author is a native of that island and as a result, the study was primarily shared within that region.

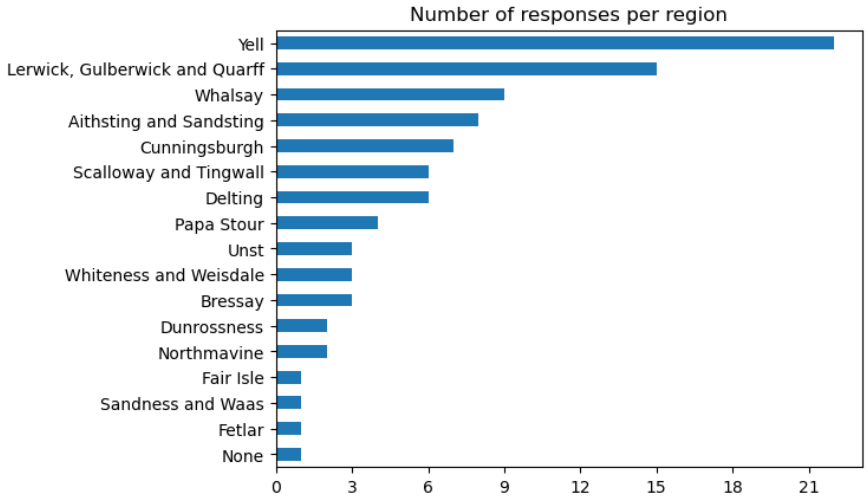

Figure 4.1: Number of responses per dialectal region - regions with no responses are not shown (Foula, Burra and Trondra, Sandwick, Lunnasting, Skerries and Nesting). This data is also available as a table in Appendix B.2.

Given that the transcript variants are distributed randomly to each participant, the submission count for each variant follows a normal distribution (see Figure B.1). While it may have been advantageous to have a flatter distribution of responses, by randomising transcript allocation we reduce some potential for bias and system malfunction throughout the study.

The median time taken to correct a transcript was 3 minutes and 49 seconds, with the median number of editing actions being 61. Each participant submitted an average of 1.5 responses. Some participants submitted multiple responses for the same transcript (i.e., they submitted a transcript where errors are not highlighted and another of the same text where the errors are highlighted). This was not intended as multiple transcripts by the same user would have the potential to bias results. We choose to include multiple submissions where the reference texts differ but not where the reference texts are the same. Eight such instances were identified and removed before further analysis.

We might have assumed that users who submitted multiple times would increase their word per minute (WPM) rate, as they became accustomed to the interface. However, we only find a very small difference between the median WPM for first and subsequent submissions at 11.5 WPM and 12.6 WPM respectively (see Appendix B.3). We may attribute this consistency to the shortness of the source transcripts meaning participants had too little time to develop familiarity with the task, or a potential bias, given that participants who have submitted multiple times may have had more time to contribute to the study in general and therefore spent slightly more time on both their initial response and subsequent ones.

We find that for some participants, the text highlighting feature of the interface was not immediately intuitive. Despite details of this aspect of the interface being explained in the instructions (see Appendix A.2), 21% of participants (25 responses) only played audio for the first section of the transcript, with most electing to either delete or ignore the rest of the passage. Interestingly, several of these participants corrected other sections of the passage without hearing the audio for them, taking context from the surrounding text alone. Unfortunately, these responses could not be used in the analysis of the transcripts in Section 4.2 as we can expect correction of words in one section to depend on the participant's choices in others.

4.2 Influence of Error Highlighting

Note: A coding error resulted in some differences between the two variants of transcript presented to participants. The differences in monitored spelling choices were therefore manually verified. Though we do not expect this to detract from the validity of the findings analysed here, we discuss the implications further in Chapter 5.

When measuring WER using the ground truth transcripts as a reference, we find that the submitted transcripts in which the selected English words were highlighted exhibit a 6% lower WER (and a 3.5% decrease in CER) than those where the errors are left as plain text (see Figure 4.2). This is to be expected, given that the "correct" word in the original transcript is available in the alternative suggestions for all of these errors. CER tends to be more consistent across the transcripts with increased error suggestion presence, with a slight positive correlation with transcript length (see Figure 4.2). This consistency can be explained by the reduced effect of word count on CER, with the word count of the plain transcripts differing from their references by 0.7 word and the highlighted transcripts only differing by 0.3.

Closer inspection reveals pronounced differences in vocabulary between the two tran-

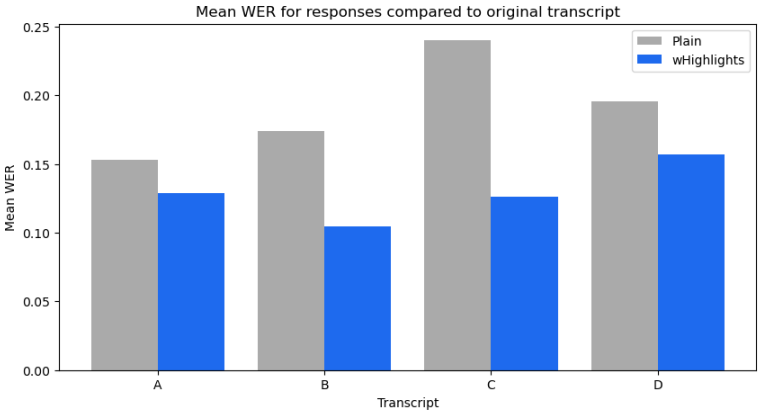

Figure 4.2: WER for both versions of each transcript, where errors are highlighted and left plain.

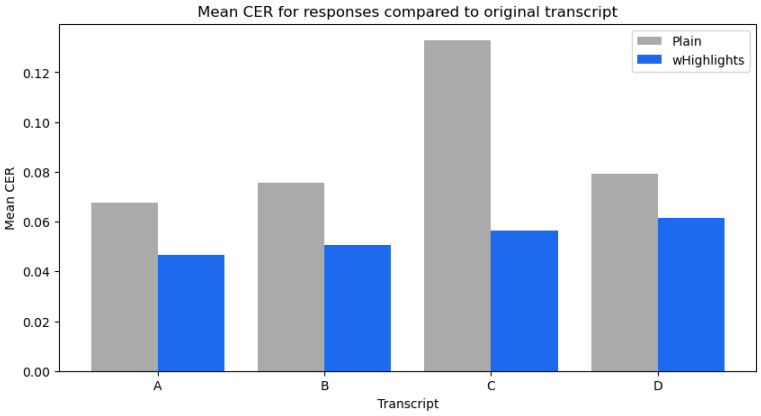

Figure 4.3: CER for both versions of each transcript, where errors are highlighted and left plain.

script variants. Our first observation is that there is no significant difference between the highlighted and plain text word “of” when left as the Shetland equivalent “o” in the text. Both are changed every time with only one exception (see Table 4.2). This shows that participants had no trouble identifying when the Shetland equivalent of a word should be written as the English version. Though there was a slight apparent tendency to change the word “from” to “fae” in the highlighted version (27.8% more likely if highlighted p=0.31), we do not have enough data to demonstrate a significant statistical difference between the two variants (see Table 4.3).

The first significant change seen in the transcripts is when the word “anidder” is left as “another” in the text. When the word is left plain, 55% (6/11) do not correct it with only 36% (4/11) choosing a dialectal spelling of the word (see Table 4.1). When the word is highlighted, all 7 opt to change it, with 4 participants changing the word to the suggested spelling, and 3 using alternate spellings. This means that the participants are 120% more likely to change the word if it is highlighted (p=0.038).

This trend is continued throughout the text. Across the “plain” transcripts, we see that participants changed the erroneous English spellings 39.7% of the time. Across

Transcript 1			
	Spelling	Plain (11)	Highlighted (7)
Text	and	7	1
Expected	an	4	6
Text	another	6	0
Expected	anidder	4	4
	annider	1	2
	annidder	0	1
Text	that	6	0
Expected	at	3	4
	dat	2	2
	it	0	1

Table 4.1: Differences in word form occurrence counts between the two variants of Transcript 1.

Transcript 2			
	Spelling	Plain (4)	Highlighted (10)
Text	o	0	1
Expected	of	4	9

Table 4.2: Differences in word form occurrence counts between the two variants of Transcript 2.

Transcript 3			
	Spelling	Plain (10)	Highlighted (19)
Text	the	7+6+6=19	4+5+2=11
Expected	da	3+4+4=11	15+14+17=46
Text	was	7+8+10=25	4+4+7=15
Expected	wis	3+2+0=5	15+15+12=42
Text	from	3	2
Expected	fae	7	17
Text	to	9	4
Expected	ta	1	9
	tae	0	6
Text	and	8	8
Expected	an	1	11
	(none)	1	0
Text	this	6	3
Expected	dis	4	16

Table 4.3: Differences in word form occurrence counts between the two variants of Transcript 3.

Transcript 4			
	Spelling	Plain (11)	Highlighted (11)
Text	they	6+5=11	1+1=2
Expected	dey	4+5=9	10+10=20
	day	1+1=2	0+0=0
Text	there	4	1
Expected	dere	4	9
	dare	3	0
	der	0	1
Text	the	6+5+6=17	0+2+2=4
Expected	da	5+6+5=16	11+9+9=29
Text	like	5	1
Expected	laek	1	6
	lik	5	4

Table 4.4: Differences in word form occurrence counts between the two variants of Transcript 4.

the variants in which the same errors were highlighted, participants elected to change the spelling to a dialectal one 80.5% of the time, meaning that participants were almost twice as likely to change an English word to a dialect equivalent if the word was highlighted (p=4e-18).

The most significant change was demonstrated with the word “to” which was 7 times more likely to be changed to the dialect equivalent (”ta” or “tae”) if highlighted (p=0.001) (see Figure 4.4).

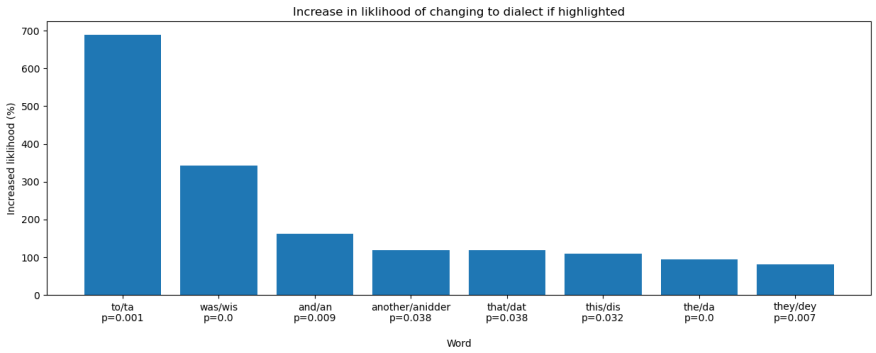

Figure 4.4: Increase in likelihood of changing an erroneous English spelling to dialect if the word was highlighted as opposed to plain text (only displaying statistically significant differences).

When asked to rate the helpfulness of word suggestions present in the text, we see that participants who completed the transcript with additional highlights were 4.25 times as likely to rate the suggestions as “very helpful” than the group who corrected the plain transcript (see Figure 4.5). It is worth noting here that 50% of participants who rated the suggestions as unhelpful did not successfully use them in the intended way (by long pressing a highlighted word) while only 20% of those who rated the suggestions

as helpful did not use them. While it is clear that some participants did not realise that the highlighted text could be interacted with, it would appear that many participants found some value in having the error words pointed out in the text. We find that there is a moderate positive correlation (0.5) between the helpfulness rating and number of times the highlighting feature was used to substitute a word in the text. We might consider this to be an indication that users are comfortable with a large number of error highlights in the text, although it is also important to note that all monitored highlights had the correct word in their list of suggestions.

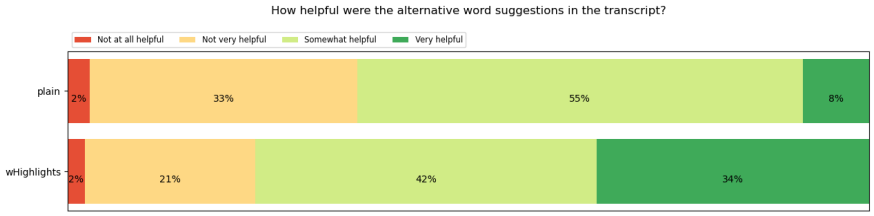

Figure 4.5: Distribution for responses when participants were asked to rate the helpfulness of word suggestions.

We find that there is a large amount of variation in average WER between different dialectal areas in Shetland. Responses from the west end of Shetland (Papa Stour, Whiteness and Weisdale, Aithsting and Sandsting) generally record much higher WERs than other regions. We also see that the North Isles (Yell, Unst and Fetlar) have a smaller WER on average than most areas (see Figure 4.6). It is important to note two biases here however. There are very few responses for some regions, for example Fair Isle which only submitted one highlighted response which we know tend to have a lower WER on average. This means that we can only make broad comparisons based on larger regions or a combination of smaller ones. We also cannot make direct comparisons between the plain and highlighted variants of transcripts between different regions as there were only a small number of responses recorded for each variant in each region. Both of these comparisons would result in a very high dependence on individual responses.

Our final finding is a moderate difference in the WER of participants with different linguistic inheritance from their parents. Users with two Shetland dialect speaking parents exhibited a 5% lower WER than those with only one (see Figure 4.7). We also find that users with two dialect speaking parents replaced more highlighted words on average than those with one. These findings indicate that participants with two dialect speaking parents were generally better at identifying erroneous English spellings throughout the transcript. The finding also highlights the role of linguistic inheritance in the performance of ASR systems. The observed difference in vocabulary choice between participants with varying levels of dialect exposure suggests that ASR models trained on data predominantly from speakers with strong dialectal backgrounds may struggle to generalize well to speakers with mixed linguistic inheritance.

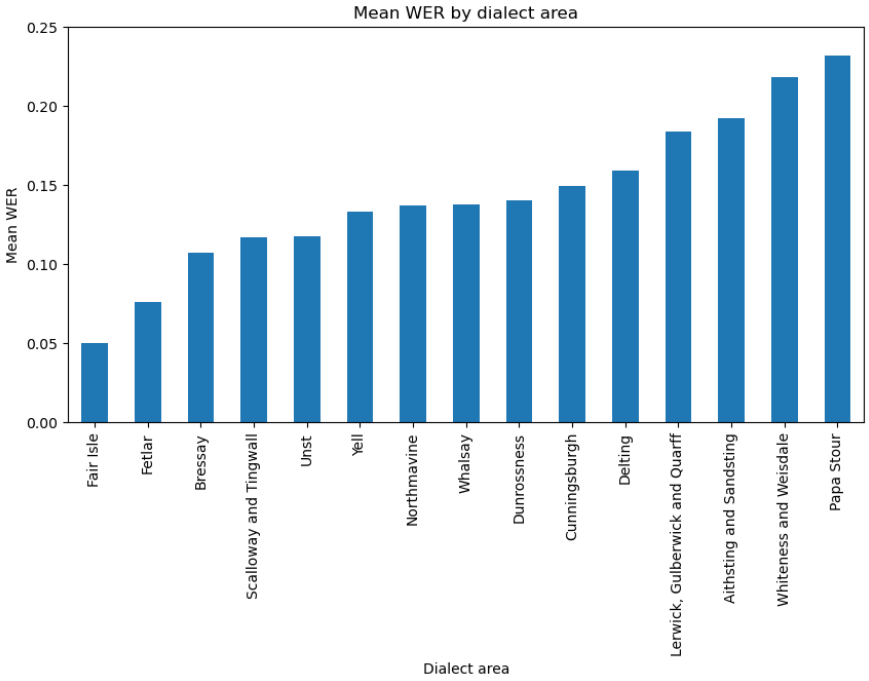

Figure 4.6: WER for different areas in Shetland. Note: this graph does not control for the distribution of transcript type presented to the participant.

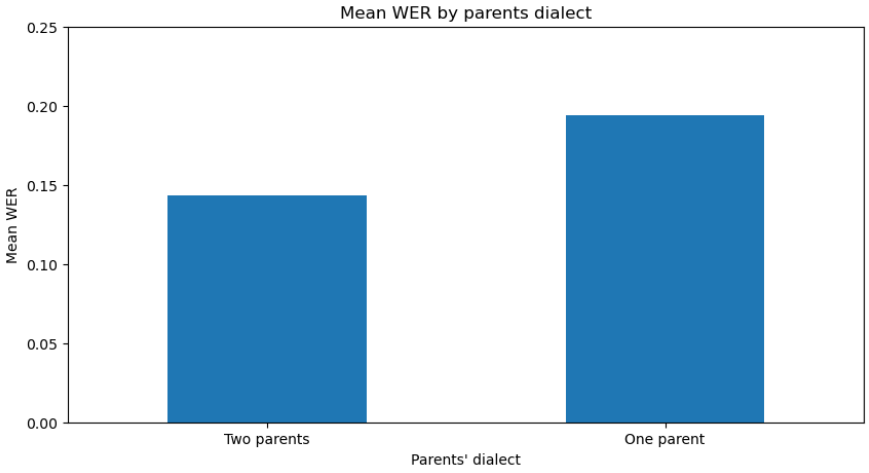

Figure 4.7: WER by parents' dialect.

Chapter 5

Conclusions and Future Work

The main elements of this project were

- The development and processing of a corpus for the Shetland dialect.
- The training and evaluation of a language model based on that corpus.
- The design and implementation of a cross-platform transcription interface.
- A study to evaluate the effects of error highlighting in a language with a non-standard orthography.

The production of a language model for the Shetland dialect presented several challenges throughout the training process. One significant challenge was the limited availability of existing resources and datasets for the dialect, which necessitated the creation of a new corpus from scratch. This involved the time-consuming process of gathering, organizing, and annotating authentic Shetland dialect data from various sources such as local publications, audio recordings, and community-contributed content.

The collection of Shetland dialect resources produced by this project is the most comprehensive known centralised corpus of the Shetland dialect. It is not however without its drawbacks. With only 5 hours of fully transcribed audio, there is little phonetic analysis possible with this data set. The small number of contributors also increases the speaker dependence of any models produced. It is hoped that future additions to this corpus will make it better reflect common usage of the Shetland dialect.

The study performed for this project uncovered interesting findings regarding the influence of error highlighting on the transcription of the Shetland dialect. By implementing error highlighting techniques within the cross-platform transcription interface, we were able to observe the impact on user behaviour during the task. We find that English words in the text, which should have been in the Shetland dialect according to the audio, were replaced by the dialect version almost twice as often when highlighted. This finding indicates a high level of sensitivity to the parameters of the correction interface when editing text based on a non-standard orthography. Unfortunately, a coding error led to several differences being created between the two variants of a transcript altering some of the context that surrounded the words we intended to check between the two

scripts. This meant we were unable to verify several metrics which may have informed a more descriptive analysis of the resulting transcripts or evaluation of the accessibility of the interface. If a similar experiment were run again, transcripts which were intended for comparison would be better controlled. Further research into this area may explore the benefit of a user learning model to identify patterns of orthography in a user's input and make suggestions based on this.

The interface created for the purposes of this study may be easily applied to other purposes to allow for easier editing and realignment of audio resources using the Flutter framework, which can be applied to a range of existing tools and interfaces. For example, as an addition to the ELPIS framework. However, as noted in Section 4.1, many users had trouble with the audio clip highlighting feature of the interface, despite instructions regarding its intended functionality. We consider this strong justification for adherence to the interface design that is commonplace in the set of popular transcription apps reviewed in Section 3.4. In a future design, utterances would be separated into paragraphs such that separate play buttons operated the audio clip position. Further additions to the interface may include an optional set of special characters for allowing easy access to diacritics in the transcription of a non-standard orthography like Shetland dialect.

Writing in Shetland dialect is a highly personal activity. We have demonstrated this both in the collection of our corpus and by observing differences in spelling choices between participants with varying linguistic inheritance. With even a small number of participants, we have collected a large number of spelling variants over a small vocabulary. Augmenting existing corpora with feedback from crowd sourced transcription is likely to yield much greater benefit to training ASR systems than a static corpus and allow for more robust and accurate systems in the future.

Bibliography

O. Adams, B. Galliot, G. Wisniewski, N. Lambourne, B. Foley, R. Sanders-Dwyer, J. Wiles, A. Michaud, S. Guillaume, L. Besacier, et al. User-friendly automatic transcription of low-resource languages: Plugging espnet into elpis. *arXiv preprint arXiv:2101.03027*, 2020.

K. C. Arnold, K. Chauncey, and K. Z. Gajos. Sentiment bias in predictive text recommendations results in biased writing. In *Graphics Interface*, pages 42–49, 2018.

K. C. Arnold, K. Chauncey, and K. Z. Gajos. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 128–138, 2020.

S. Azenkot and N. B. Lee. Exploring the use of speech input by blind people on mobile devices. In *Proceedings of the 15th international ACM SIGACCESS conference on computers and accessibility*, pages 1–8, 2013.

P. Bell and S. Renals. A system for automatic alignment of broadcast media captions using weighted finite-state transducers. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 675–680. IEEE, 2015.

L. Besacier, E. Barnard, A. Karpov, and T. Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100, 2014.

F. Biadsy, P. J. Moreno, and M. Jansche. Google's cross-dialect arabic voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4441–4444. IEEE, 2012.

S. Bird. Sparse transcription. *Computational Linguistics*, 46(4):713–744, 2021.

Cambridge Dictionary. Definition of "den" - english dictionary, 2023. URL https://dictionary.cambridge.org/dictionary/english/den.

M. Cao, S. Renals, P. Bell, A. Li, and Q. Fang. Grapheme-to-phoneme conversion methods for minority language conditions. In *2012 International Conference on Speech Database and Assessments*, pages 151–156. IEEE, 2012.

A. Cooper. *About Face: The Essentials of Interaction Design*. 4 edition, 2014.

C. De Luca. Mother tongue as a universal human right. *International journal of speech-language pathology*, 20(1):161–165, 2018.

M. Durham. Thirty years later: Real-time change and stability in attitudes towards the dialect in shetland. *Sociolinguistics in Scotland*, pages 296–318, 2014.

L. Evans, W. Lamb, M. Sinclair, and B. Alex. Developing automatic speech recognition for scottish gaelic. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 110–120, 2022.

R. A. Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the royal statistical society*, 85(1):87–94, 1922.

B. Foley, J. T. Arnold, R. Coto-Solano, G. Durantin, T. M. Ellison, D. van Esch, S. Heath, F. Kratochvil, Z. Maxwell-Smith, D. Nash, et al. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *SLTU*, pages 205–209, 2018.

Y. Gaur, W. S. Lasecki, F. Metze, and J. P. Bigham. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*, pages 1–8, 2016.

A. Glasser. Automatic speech recognition services: Deaf and hard-of-hearing usability. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pages 1–6, 2019.

S. Goddijn and D. Binnenpoorte. Assessing manually corrected broad phonetic transcriptions in the spoken dutch corpus. In *Proceedings of ICPhS*, pages 1361–1364, 2003.

J. J. Graham. *The Shetland Dictionary*. The Shetland Times Ltd, 1993.

D. Harwath, A. Gruenstein, and I. McGraw. Choosing useful word alternates for automatic speech recognition correction interfaces. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

J. B. Johannessen, J. Priestley, K. Hagen, T. A. Åfarli, and Ø. A. Vangsnes. The nordic dialect corpus–an advanced research tool. In *Proceedings of the 17th nordic conference of computational linguistics (nodalida 2009)*, pages 73–80, 2009.

D. Jurafsky and J. H. Martin. *Speech and Language Processing*. 3 edition, 2019. Draft of September 23, 2019. Retrieved from https://web.stanford.edu/~jurafsky/slp3/.

R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE, 1995.

H. Kolkhorst, K. Kilgour, S. Stüker, and A. Waibel. Evaluation of interactive user corrections for lecture transcription. In *Proceedings of the 9th International Workshop on Spoken Language Translation: Papers*, 2012.

S. Krug. *Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability*. 2014.

Y. Liang, K. Iwano, and K. Shinoda. Simple gesture-based error correction interface

for smartphone speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014a.

Y. Liang, K. Iwano, and K. Shinoda. An efficient error correction interface for speech recognition on mobile touchscreen devices. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 454–459. IEEE, 2014b.

R. Ljosland, 2012. URL https://connected-communities.org/index.php/project/the-orkney-and-shetland-dialect-corpus-project-scoping-study/.

Microsoft. A swiftkey employee has made it his mission to upload obscure languages. 2017. URL https://news.microsoft.com/en-gb/2017/02/21/a-swiftkey-employee-has-made-it-his-misson-to-upload-obscure-languages/.

V. Panayotov. Librispeech language models, vocabulary and g2p models, 2014. URL http://www.openslr.org/11/.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, et al. English conversational telephone speech recognition by humans and machines. *arXiv preprint arXiv:1703.02136*, 2017.

Shetland ForWirds. Dialect map of shetland, 2022. URL https://www.shetlanddialect.org.uk/dialect-map-of-shetland.

Shetland News. Local dialect predictive text comes to mobile keyboard. 2021. URL https://www.shetnews.co.uk/2021/11/24/local-dialect-predictive-text-comes-to-mobile-keyboard/.

Shetland News. Dialect group releases dictionary app for mobile devices. 2022. URL https://www.shetnews.co.uk/2022/07/27/dialect-group-releases-dictionary-app-for-mobile-devices/.

J. Smith and M. Durham. A tipping point in dialect obsolescence? change across the generations in lerwick, shetland 1. *Journal of sociolinguistics*, 15(2):197–225, 2011.

J. Smith and M. Durham. Bidialectalism or dialect death? explaining generational change in the shetland islands, scotland. *American Speech*, 87(1):57–88, 2012.

M. Sperber, G. Neubig, S. Nakamura, and A. Waibel. Optimizing computer-assisted transcription quality with iterative user interfaces. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1986–1992, 2016.

A. Stolcke and J. Droppo. Comparing human and machine errors in conversational speech transcription. *arXiv preprint arXiv:1708.08615*, 2017.

B. Suhm, B. Myers, and A. Waibel. Multimodal error correction for speech user interfaces. *ACM transactions on computer-human interaction (TOCHI)*, 8(1):60–98, 2001.

P. Sundkvist. 'standard english'as spoken in shetland's capital. *World Englishes*, 30 (2):166–181, 2011.

P. Sundkvist. 'insular isles, insular speech'? language change in the shetland islands. *Moderna språk*, 106(2):150–158, 2012.

Tech Radar. Best speech to text apps: Mobile speech to text apps to consider, 2023. URL https://www.techradar.com/news/best-speech-to-text-app#mobile-speech-to-text-apps-to-consider.

L. Wang, T. Hu, P. Liu, and F. K. Soong. Efficient handwriting correction of speech recognition errors with template constrained posterior (tcp). In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, et al. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*, 2018.

E. Yılmaz, M. McLaren, H. van den Heuvel, and D. A. van Leeuwen. Semi-supervised acoustic model training for speech with code-switching. *Speech Communication*, 105:12–22, 2018.

Appendix A

Transcription Interface

A.1 Screenshots

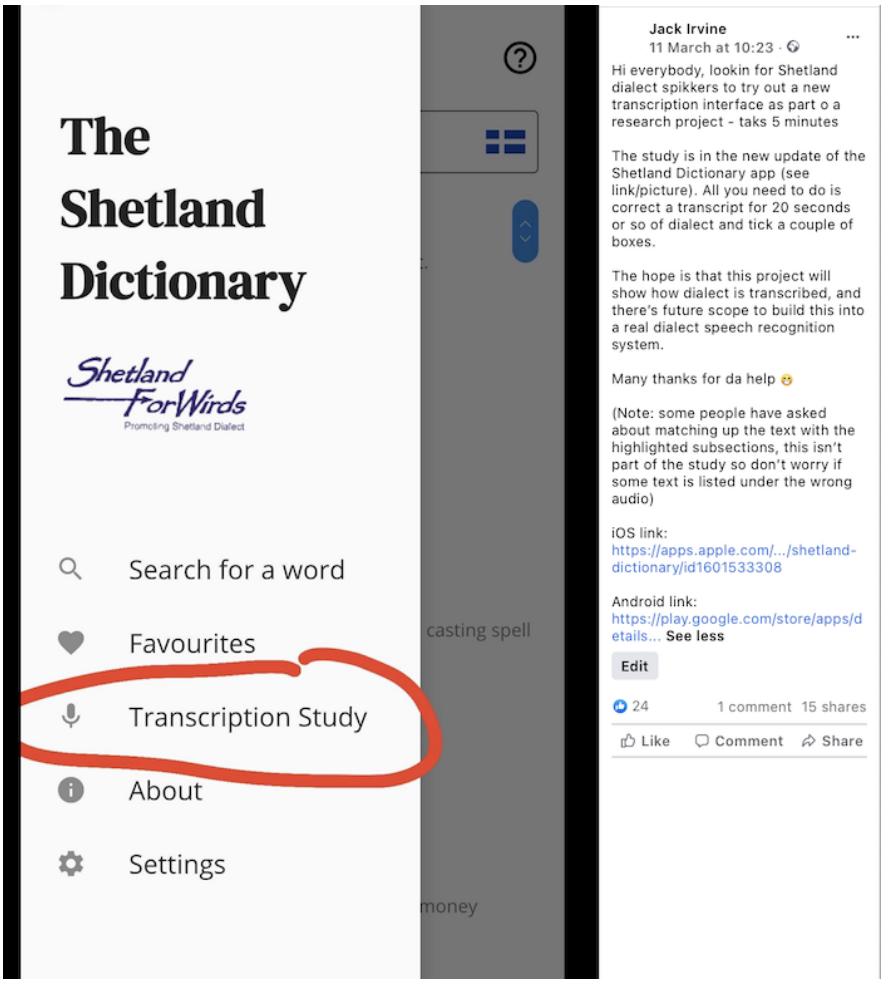

Figure A.1: Screenshot of the social media post used to promote the study

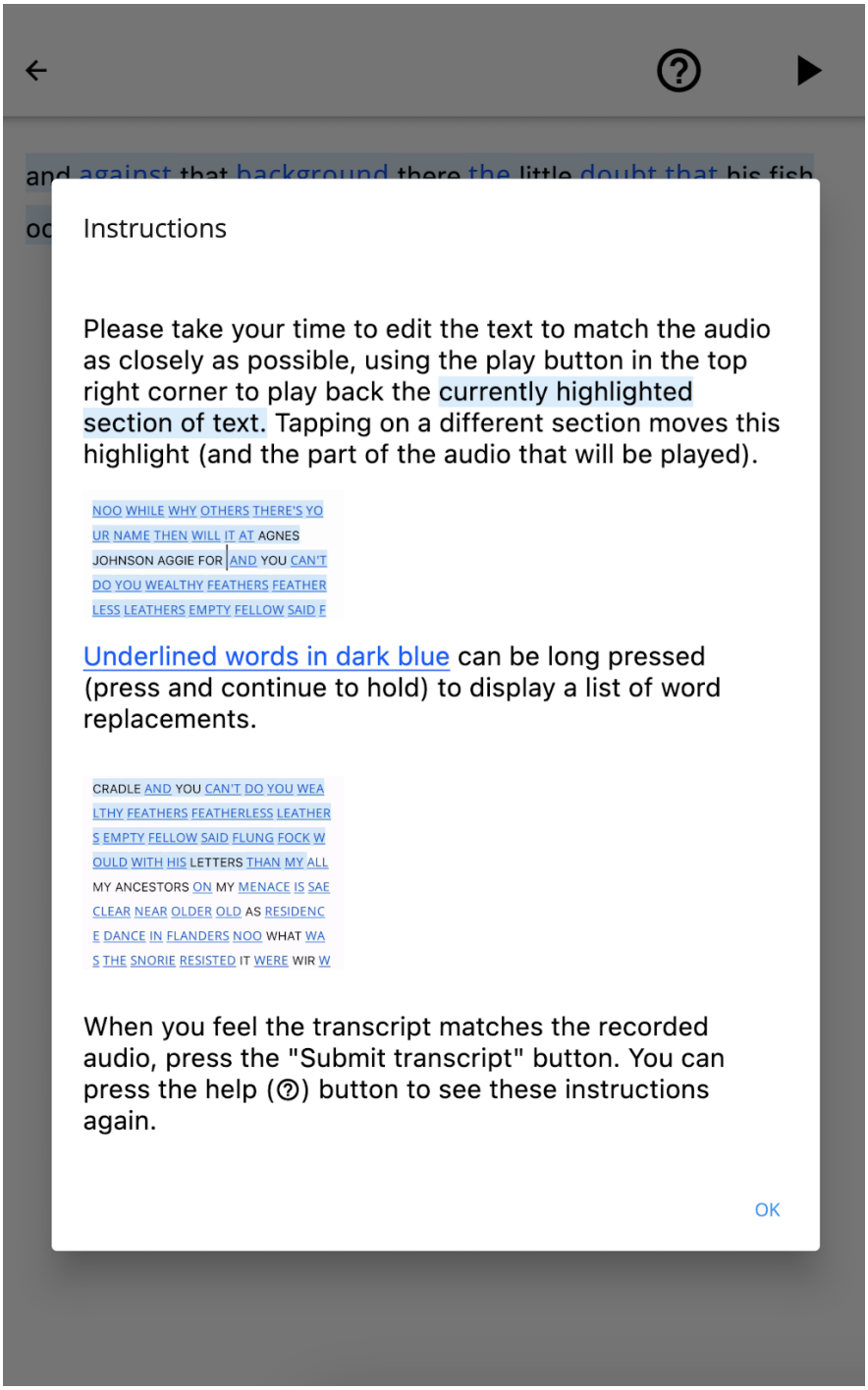

Figure A.2: Screenshot of the transcription task instructions. The blocks of smaller text between the instructions are animated GIFs which display the explained action taking place.

Appendix B

Additional Figures

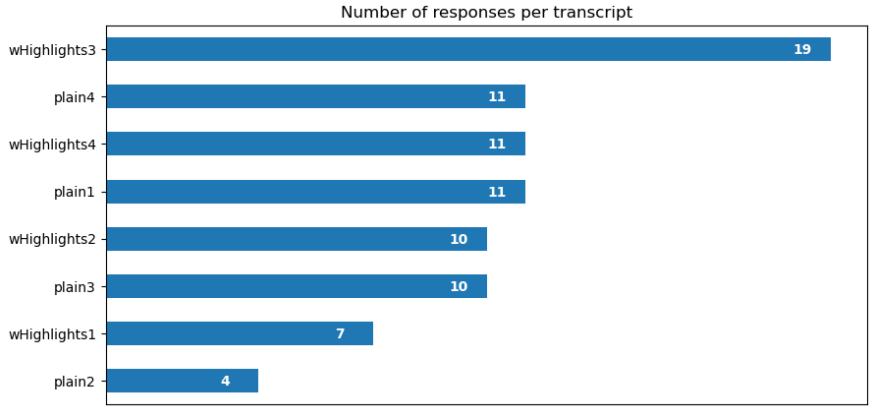

Figure B.1: Number of responses per transcript variant. Variant counts grouped by region are available in a table in Appendix B.2.

	plain1	plain2	plain3	plain4	wHighlights1	wHighlights2	wHighlights3	wHighlights4
Aithsting and Sandsting	-	-	1	3	-	-	3	1
Bressay	-	-	-	-	-	1	1	-
Cunningsburgh	-	-	-	1	1	2	1	-
Delting	2	-	1	-	-	-	2	1
Dunrossness	-	-	-	-	-	1	1	-
Fair Isle	-	-	-	-	-	-	1	-
Fetlar	-	-	-	-	-	-	-	1
Lerwick, Gulberwick and Quarff	2	-	2	3	2	-	3	3
Northmavine	-	-	-	-	1	-	-	1
Papa Stour	1	-	1	-	-	-	-	1
Scalloway and Tingwall	1	1	-	-	-	1	2	1
Unst	1	-	-	-	-	1	-	1
Whalsay	2	-	1	1	-	1	1	1
Whiteness and Weisdale	1	1	-	-	-	-	1	-
Yell	1	2	4	3	3	3	3	-

Figure B.2: Responses grouped by area

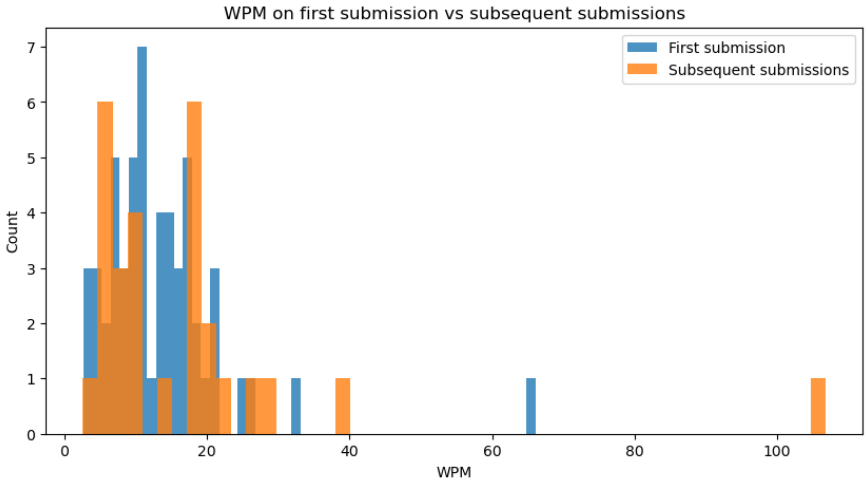

Figure B.3: Distribution of WPM for first submission and subsequent submissions.

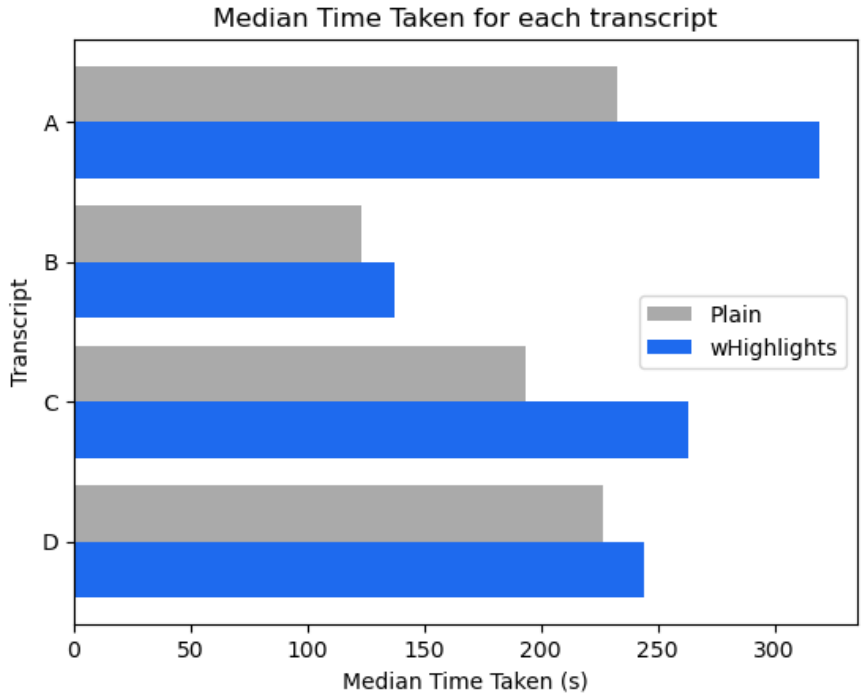

Figure B.4: Median time taken to complete the task, grouped by transcript

Appendix C

Participants' information sheet

Participant Information Sheet

Project title:	Designing accessible speech transcription interfaces
Principal investigator:	Peter Bell
Researcher collecting data:	Jack Irvine
Funder (if applicable):	

This study was certified according to the Informatics Research Ethics Process, RT number 280672. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

Jack Irvine – Artificial Intelligence and Computer Science student designing accessible interfaces for speech transcription as part of a final year honours project.

Casey Gong – Computer Science student focussing on a complementary final year honours project developing speech recognition for a low resource language.

Peter Bell – Project supervisor

What is the purpose of the study?

The purpose of this study is to involve native Shetland dialect speakers in the testing and evaluation of a new transcription interface. It is hoped that this interface will reduce the challenges associated with speech transcription for those less familiar with the task while maximising value to training automatic speech recognition systems. This study will focus on producing an effective and accessible speech transcription interface for Shetland dialect.

Why have I been asked to take part?

You have been asked to take part in this study as you are a native speaker of the Shetland dialect.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time without giving a reason. If you wish to withdraw, contact Jack Irvine

(j.l.irvine@sms.ed.ac.uk). We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

You will be asked to complete a small number of tasks using a transcription interface and submit a short questionnaire reflecting on this experience. These questions may ask about the participant's relationship with Shetland dialect, where they learned it, their parents' dialect and how they use dialect in day-to-day life. Some data will be recorded about how the transcription interface is used. In some cases, participants may be asked to participate in a brief interview or focus group (online or in-person) to evaluate various aspects of the user interface. These sessions will take no longer than an hour and you do not have to participate in them if you do not wish to do so.

Are there any risks associated with taking part?

There are no significant risks associated with participation.

Are there any benefits associated with taking part?

There are no direct benefits of taking part in this study. However, your participation in the study will support the development of speech recognition technology for Shetland dialect, making it easier for some people to interact with technology using their natural voice. This may in turn pave the way for further development and integration with existing platforms.

What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the

research team listed above and a small number of supporting researchers, to whom your data will be non-identifiable.

All electronic data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint) and all paper records will be stored in a locked filing cabinet in the PI's office. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the lead researcher, Jack Irvine, by email at j.l.irvine@sms.ed.ac.uk.

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.

If the research project changes in any way, an updated Participant Information Sheet will be made available on http://web.inf.ed.ac.uk/infweb/research/study-updates.

Alternative formats.

To request this document in an alternative format, such as large print or on coloured paper, please use the contact details listed above.

General information.

For general information about how we use your data, go to: edin.ac/privacy-research

Appendix D

Participants' consent form

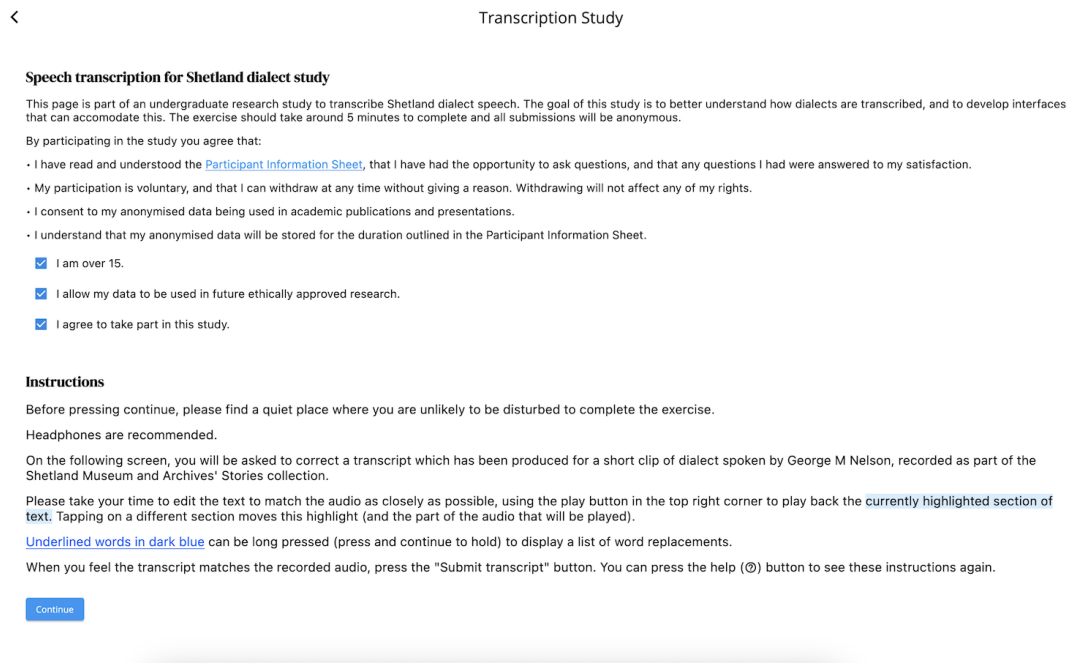

Figure D.1: Page within the Shetland dictionary app summarising the participant information sheet, gathering consent and explaining the instructions of the task