Assessment 2 - CS5063

Group work AI system evaluation

08th November 2020

Jasmin Uhlhorn - **52095681**

Jess McGowan - **52093344**

Vladimir Yesipov - **52093351**

Yassin Dinana - **52094480**

Table of Contents

1) Reflection on evaluation

a. Introduction to TTSReader

TTSReader is a tool that is free to use and requires no registration. It allows the user to listen to a text which is read aloud by the system.

The application allows the reading of typed or pasted text, the reading of a pdf file by extracting the content, and when using Google Chrome a free extension can be used to let the system read out websites.

Depending on the browser and operating system used, different voices and languages are available. Where some languages have a higher number of representative voices, including male, female, and local accents. Whereas, other languages are not featured at all and others which might have a single voice.

Another feature of the TTSReader allows the user to adapt the speed setting of the reading towards their preferences. The speed setting allows eight different settings ranging from very slow to three times the normal speed.

On multiple browsers, an automatic saving function is supported which allows the user to exit the session and continue the reading wherever it left off, another time.

The interface of the TTSReader is intuitive and understandable but additional documentation describing the system was not found. The documentation available on the web site does not describe the system functionalities detailed enough. Barely information available about what languages system supports and how the language options depend on the browser and the platform.

b. Features of TTSReader for evaluation

In the evaluation of TTSReader with the users, four different languages are mainly being evaluated depending on the user's native language, these languages are English, German, Russian, and Arabic.

Depending on the user's device and operating system that is used during this evaluation, different versions of the languages are available. Each user pastes a sample text into the TTSReader text block window to test it, where the text is the same but just translated into the four different languages, therefore, the results are more accurate.

TTSReader has different features, such as reading full PDF documents, reading eBooks, and narrating movies and presentations. The main feature used in the evaluation is simply testing the Text-To-Speech technique with a sample paragraph given to users in different languages.

c. Evaluation measures

We measure the performance of the system by asking the user to rate their experience. We focused on two main aspects. First, the quality of the reading. Was the reading understandable? Was the reading fluent? Second, the accuracy. Was the pronunciation accurate? (https://ehudreiter.com/2019/10/08/accuracy-fluency-and-utility/)
We will compare different languages with each other. Thus, we use the users' evaluation of the English language as our baseline and compare it with the other languages.

The languages will be compared with English from the 3 main groups (German, Russian and Arabic) and the group for other languages was added for other language participants. The effectiveness will be measured according to the participants response (answers) on questions from the questionnaire provided. The first parameter we are going to compare is the understandability of the text read by the system, which is going to be evaluated by the user. The second parameter is the fluency of the reading. The third parameter is the accuracy of the reading. The last parameter is the experience of using the TTSReader by the participant. During our preparation for the TTSReader research we noticed that different functionalities of the system available for different systems and different browsers. That is why we included the section for participants with questions about their system. These questions possibly can help us to explain the results we are going to get from participants.

The participants are provided with a five point Likert scale to evaluate their experience of using the TTSReader system. The Likert scale means that each question provided with the series of answers ranged from 1 (poor quality or experience) to 5 (excellent). Compared to binary questions, where we could only have two answers to choose from, Likert-type questions can give us a more detailed feedback. Using the Likert scale in our survey will help us to uncover the degrees of opinion that can give us a deeper understanding of the feedback we are receiving. It can also help identify the areas in which the TTSReader system could be improved, such as speech understandability, fluency, accuracy and front end - online representation of the system and its content which are responsible for the user-system interaction experience. The main advantage of using a human rating evaluation method is that people are able to evaluate the semantics and the usefulness of the speech.

2) Evaluation design

a. Research Questions and Hypotheses

The research question for this project is “Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language?”. Our hypothesis is that native English speakers will have a better experience in using TTSReader as English is one of the predominant languages in natural language processing. Following from this, the null hypothesis is “There is no difference in quality of experience between native English speakers and non-native English speakers in using TTSReader in their native language”, and our alternate hypothesis is “There is a significant difference in quality of experience between native English speakers and non-native English speakers in using TTSReader in their native language”.

b. Experiment Design

As the research question is “Do native english speakers have a better experience with TTSReader compared to non-native english speakers using TTSReader in their native language?”, the evaluation for all languages is done using the same paragraph translated in different languages in order to get more accurate results, apart from English, three other languages are used for the evaluation such as Russian, Arabic, and German where the results of the evaluation of the English section will later be compared to the results of other languages.

This is done using the *between subjects approach*, where different participants will be testing each condition depending on their native language in order to later analyze the results between the english speakers and non-native english speakers which will help achieve an adequate answer to the research question.

Through asking participants to rate TTSReader using a likert scale, a statistical analysis can be carried out to gain insights into how participants viewed the quality of their experience in using the web application. By comparing the average ratings provided, alongside other statistical measures including standard deviation, variance, and tailed hypothesis tests to measure whether participants who were non-native English speakers had a significantly different experience to native English speakers when using TTSReader in their native languages.

c. Experiment Procedure

Participants will be recruited through convenience sampling methods. This is due to the short timescale of this project, as well as having access to people who have non-English native languages. This will be carried out through requesting responses through online communities, including Facebook, Teams, and by hosting an online survey on Google Forms, which will include the participant information sheet, as well as a consent form.

Unfortunately, due to Covid-19 and social distancing restrictions, this experiment has had to be carried out remotely online. Ideally, in-person sessions would have been used in order to

carry out this survey, using fixed TTSReader options for all participants of the same native language. However, while creating the survey, it was discovered that TTSReader has different options for voices which are dependent on the participant's web browser and operating system. This has made creating a remote survey more challenging, as there are no voices that are universal for these variables. As a result, the survey asks users to self report which browser they are using, and on which operating system they are completing the survey on, as well as reporting which voice option that they selected.

Users are also asked to self report on their native language for this study, as this survey is carried out anonymously online there is no method to verify their claims. There is little for users to gain by lying about this, and there is no other personal information that is collected that could possibly lead a user to lie about this, but we are aware that this could affect results.

For the evaluation of TTSReader, we ask participants to select one of the voice options available on the website for their native language. Ideally we would have preselected a particular voice to reduce that as a variable, but as discussed earlier, due to participants having to complete this task on their own computers, there was no voice option that was available for all users.

The participants are then asked to paste a preselected paragraph into TTSReader to test the functionality, these are provided in English, German, Russian and Arabic. This preselected text is an extract from the Red Cross website in the various languages. Preselected text was chosen in order to reduce a variable that some people may select more complex text when given a free choice, and to have text with the same meaning between languages so that the potential for participants to have biased results based on the text has been reduced. Additionally, the chosen text offers a good coverage. It includes commonly and uncommonly used words, as well as abbreviations, numbers, and dates. The user will then evaluate the reading using a likert scale for understandability, language fluency and pronunciation accuracy for a quantitative analysis, and a comment box will be available for qualitative analysis as well.

3) Ethical Approval

Ethical Considerations

The required ethics forms for this project are included in Appendix A. Annex A is a checklist to investigate whether a study requires ethical approval, Annex B is the completed Ethics Review for, Annex C is the Participant Information Sheet and Annex E is the Participant Consent form which was included as part of the online survey.

Different ethical considerations were concluded for the evaluation in broad areas such as *users selection* which intends to choose the correct users for the evaluation by making sure all the users are older than 18 and ensuring that potentially vulnerable participants including those with cognitive accessibility needs were not included amongst participants. We achieved this through recruiting people via convenience sampling and, thus, no vulnerable people were asked to complete the form.

Protecting the data of the users by not saving, recording, or releasing the users personal information, this is also stated in the agreement for users to read before starting the evaluation, the only needed information from the users is their native language, which is collected anonymously online and not stored anywhere. It is known that the users participating in this evaluation might be from outside the UK and the EU. No social media platforms are used during the evaluation such as Facebook, Twitter, and more.

All the users have a “Participant Information Sheet” available at the beginning of the evaluation where the users choose to agree to it and continue the evaluation. This form explains all the process to the participants such as how the software is used, how their data is safe and automatically deleted after the evaluation, and it also states that users have the right to leave the evaluation survey when they decide to do so. Before users submit their responses, a final confirmation is provided informing them that if they proceed, then their responses will be used in the project, however they can still leave and their responses will not be recorded, and users are required to tick the box to agree before their submission will be accepted.

4) Results

a. Demographics and Qualitative Analysis

Demographics

We received 30 responses in our survey, which were collected from the evening of 2nd November until the morning of 6th November. All of these responses were from the languages that we focused this study on (English, German, Russian, Arabic), although we did provide the option for native speakers of other languages to evaluate TTSReader in a less structured manner.

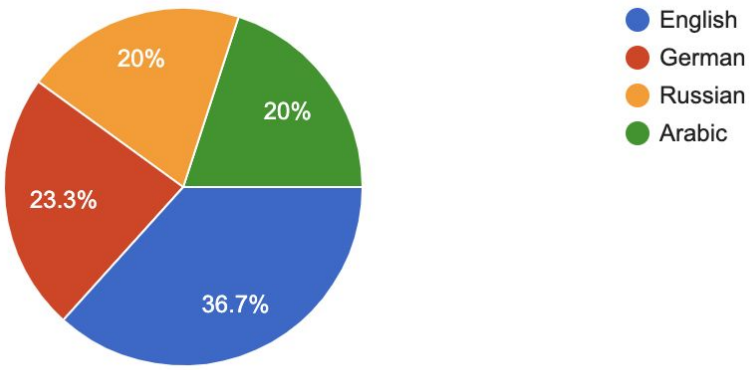

Figure 1 - Responses for “What is your native language?”

As can be seen in figure 1, 11 of our respondents stated that English was their native language, which was the largest individual group in this survey. A majority of our respondents spoke a native language other than English, with approximately equal proportions between German, Russian and Arabic. This provides us with a good number of responses in order to carry out statistical analysis to test our hypotheses.

As mentioned in the experimental design, due to having to carry out this study remotely, the system that participants are using will have an impact on the voice options that TTSReader provides. In order to assess how this could affect the participant’s experience in using TTSReader, we asked respondents to report which operating system and which web browser they were using in order to complete the survey. The results can be seen in figures 2 & 3. The 30% of users who used Chrome as their web browser for this study had the benefit of being able to choose from Google’s TTS voice options, as well as the voice options that are included in the user’s operating system.

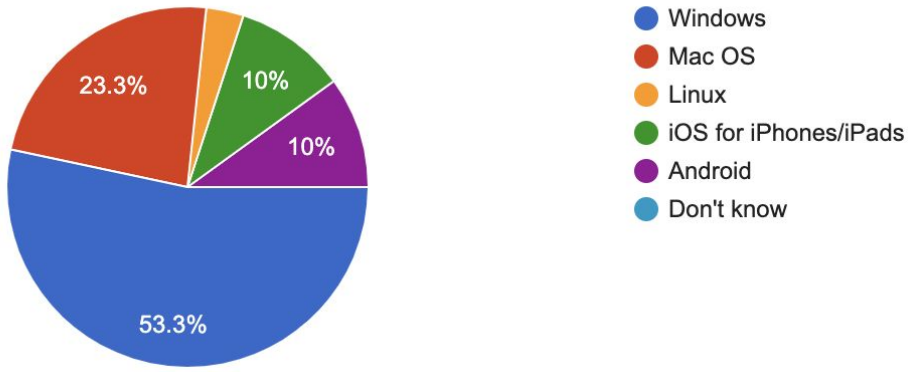

Figure 2 - Responses for “What is your operating system that you are using for this survey?”

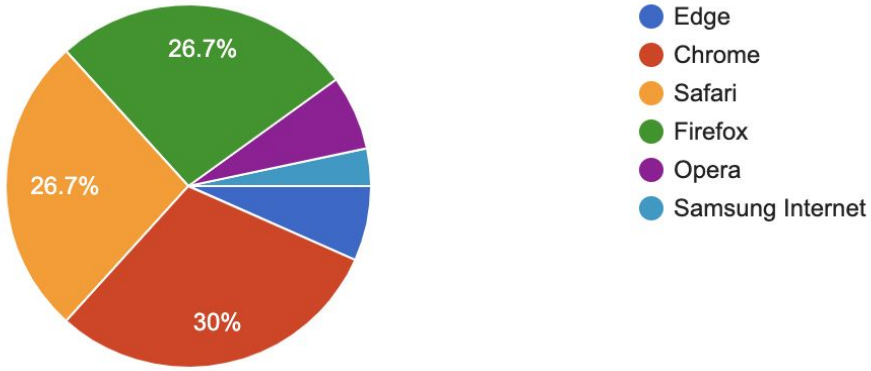

Figure 3 - Responses for “Which web browser are you using?”

English Results

TTSReader provides English users with a wide variety of voice options, even on a single operating system - for example, a user using Chrome on MacOS has 14 different options available for English, including options for India, Australia, and Ireland alongside UK and US generated voices (figure 4). While this has the potential for providing a better user experience through giving users choices to have the system sound more familiar, this can also lead to a wider variety of experiences. Figure 5 shows the voice options that the English native speakers selected during this survey.

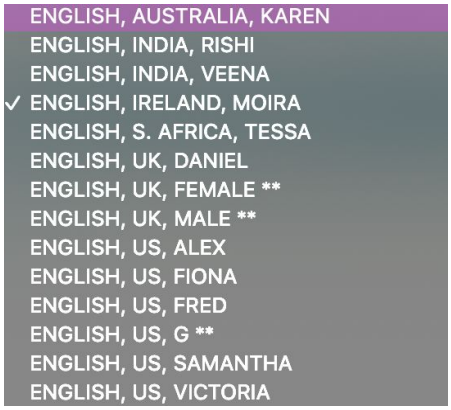

Figure 4 - The available choices for one user on TTSReader for English

The users that provided comments on their experience brought up how the voice sounded, with comments including “very monotonous”, “the ‘Hawkingness’ of the speech”, “a bit robotic” descriptions of English UK Daniel, English UK Hazel, and English UK Received Pronunciation. The Daniel and Received Pronunciation voices were also criticised for the intonation and cadence of the readings, and how this can be off-putting when trying to listen to what the voice is saying. The default English voice on Android was described as “strangely accented” and “nearly incomprehensible”. The users who rated their experience highly were less likely to leave comments, as such the comments are generally negative. The one comment from a participant who rated their experience well was for English UK Hazel, and described the pronunciation and understandability of the reading as “perfect”, however it was still “very obviously unnatural” which detracted from the experience.

These comments illustrate that while our hypothesis was that native English speakers would have a better experience in using TTSReader, this does not mean that the experience is perfect.

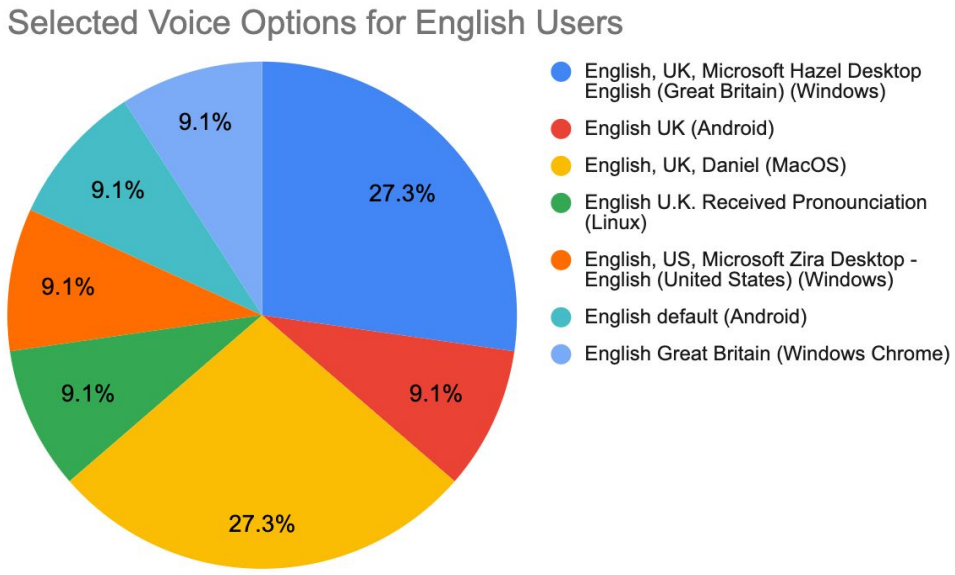

Figure 5 - The voice options that native English participants selected

German Results

Seven participants of the survey selected German as their native language and conducted the survey accordingly.

A German voice option is available on each browser regardless of the os. However, which voice option is available, it’s quality, and how many differs strongly. For example, on Windows using Chrome the participants could have selected between two german options, Mozilla offers only one, and with Edge four German voice options are available.
At first glance it could seem like every participant selected a different voice option as either the os, the browser, or the voice option entered in the survey is different. However, having a closer look at the participants reveals that all the users that used windows and mozilla used the same voice. This is obvious as there is only one option for German for that setting.

Having a look at the entries shows that the differences are capitalization and appending the speed information. This problem of having the user type in their used voice option is further discussed in section 5,a.

The distribution of the voice options after reviewing and adjusting them can be seen in figure 6. It shows that four voice options had only one participant. Only the fifth voice option (Windows, Firefox, DEUTSCH) had more with four people using that environment.

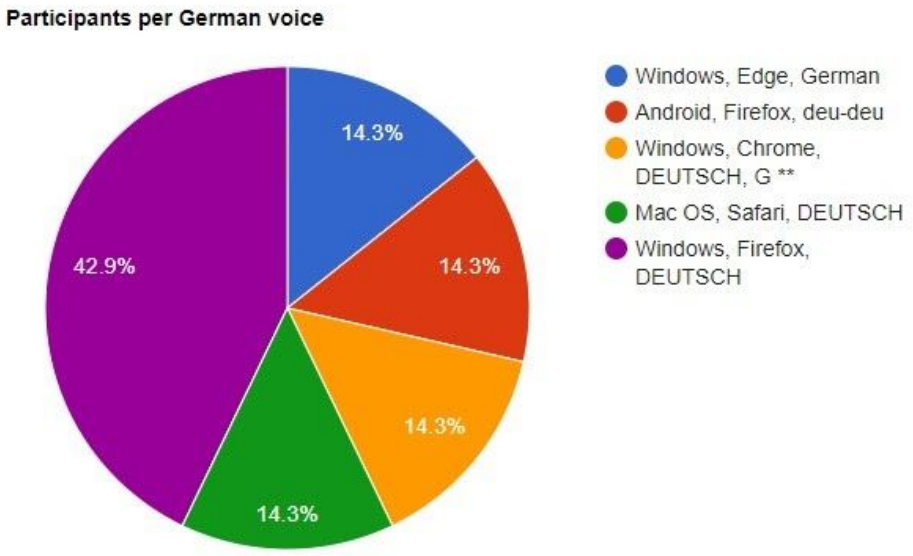

Figure 6 - The voice options that native German participants selected

To obtain a better understanding of the values provided by the users we provided them with the option to leave a comment about their selection. A German participant commented that the voice Windows, Firefox, DEUTSCH failed to, first, recognize the abbreviated company's name and tried to pronounce the abbreviation as a normal word. Secondly, it pronounced the year 1863 as a normal number but in return didn't pronounce 16000 as a number but read each character individually.

Russian Results

The users who used Russian as their native language had very different experiences using the TTSReader system. Depending on the platform and browser different voice options available for different users. The functionality of the system also depends on the user system, which is commented by the users.
The TTSReader by itself does not provide many options for Russian users, for example users have only two options to choose from when using Windows OS with Chrome browser and only three options with Edge.

Six participants used Russian as their native language which is 20% of the total number of participants (see Figure 1).
The voice option selected by the participants is sometimes hard to assess, because some participants did not indicate the full name of the voice. So, we can only guess what voice options they might have had with their system and browser, where option "РОССИИ, G**" is available for many users (need to say that even this one word voice name is used in a wrong

grammatical form). If the web page had the documentation on voice options available for each language we could provide our participants with the selection menu for the voice used instead of the text box we used, which might assist to consider different voice options and help to compare them.

The understandability of the text to speech in Russian is evaluated as good by most of the participants. However, some of the participants could not evaluate the quality of speech as in some operating systems would not play. For example, some iPhone users reported difficulties in playing the sample.

The voice reading the text with the “РОССИИ, G**” option sounds fine and accurate in Russian as it was commented by participant number 29. One of the participants left a comment saying: “Only the first sentence was read in Russian, and then pronounced only the numbers from the text fragment in English”, all these kinds of bugs make the system very hard to use and give the bad impression to the user.

Most of the Russian users also reported issues connected with the page design. The main menu is poor, background images distract from the text, some text on the page is hard to read, there is no enough description for the interface elements, buttons on screen are not responsive when used on the phone, and no option available to change the page language which is very important for the online system that is aimed for multilingual users auditorium (Figure 7).

Figure 7 Fragment of the TTSReader web page.

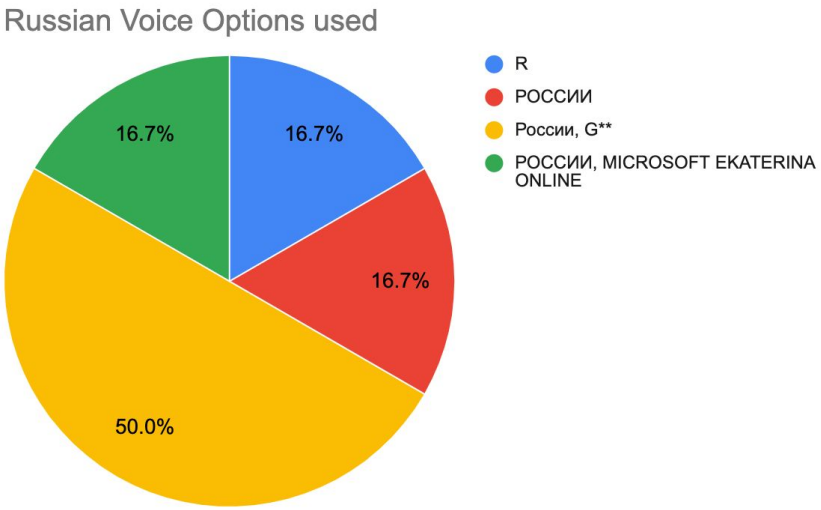

Figure 8 - The voice options that native Russian participants selected

Arabic Results

The users who used Arabic as their native language during the evaluation had different thoughts and results about TTSReader. There are two different Arabic versions in TTSReader depending on the device and the operating system being used. In terms of understanding the Text-To-Speech in Arabic, by looking at the results it can be seen that the users had a good experience with understanding, only one user out of the six users found problems with understanding the speech. The figure below shows the percentage of users who chose different Arabic voices on TTSReader.

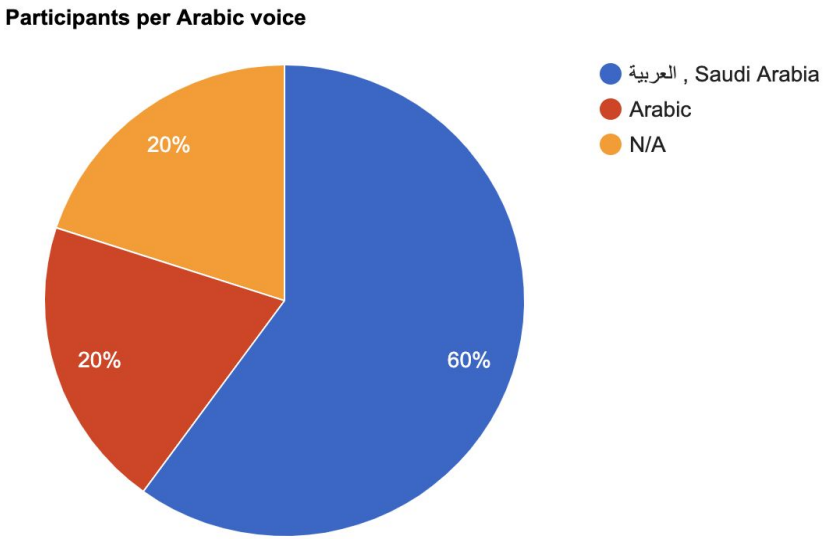

Figure 9 - The voice options that native Arabic participants selected

Unfortunately, the users with Windows as their operating system did not find Arabic in the options. Which affected the evaluation and they were not able to evaluate the system using their native language.

Looking at the language fluency, different users saw different results where the results of the fluency were divided between different outcomes between the users. Some participants thought the fluency was satisfactory and other users did not.

Regarding the Arabic pronunciation in TTSReader, no user found the pronunciation to be very understandable and satisfying, although most users thought it is only acceptable.

As an overall experience when dealing with TTSReader using the Arabic language, most of the users rated the experience to be satisfying and had a good interaction with the system.

Overall Results

Experiences between speakers of the languages included varied widely, however there were some common aspects found between languages on TTSReader. Two participants found similar experiences in TTSReader not being able to read the entirety of the extract in Russian or Arabic, including pronouncing numbers in English. One German-speaking participant also had issues with the pronunciation of numbers, with a year being read as a number, and a number having its digits read individually. This is not something that English-speaking participants commented on in their evaluation.

Native English speakers were more critical of how the voice options sounded compared to speakers of other languages. In fact, a Russian participant commented that the voice option they used was “pleasant to listen to”, which is the opposite reaction compared to the native English speakers’ complaints about the voice sounding “robotic” and “unnatural”.

b. Statistical Analysis

The participants rated the conducted reading on a likert scale from 1 to 5 in the categories understandability, fluency, pronunciation accuracy, and overall experience. The standard deviation is used to measure how the data values dispersed around the mean. The mean and median are used to estimate the average evaluation given by the participants of each language group which we can compare with English language group.

The Mann-Whitney-Wilcoxon test is used to decide whether the population distributions are identical. “Because Likert item data are discrete, ordinal, and have a limited range, there’s been a longstanding dispute about the most valid way to analyze Likert data. The basic choice is between a parametric test and a nonparametric test.”(Agnes Ogee, 2020). The advantage of the nonparametric Mann-Whitney-Wilcoxon test is that we do not need a normal continuous distribution and in our case we have Likert scale with discrete data having a limited range. The parametric tests, such as t-test need a sufficient sample size and assume a normal, continuous distribution, which is not relevant in our case. The advantage was given to the Mann-Whitney-Wilcoxon test because the data set is small and distribution of the values is limited by the Five-Point Likert scale, but both t-test and Mann-Whitney-Wilcoxon test are able to detect the significant difference (de Winter, J.C.F. and D. Dodou, 2010).

The table of results used for the data analysis has been included in Appendix 2, and the R code used to carry out this analysis is included in Appendix 4.

English Results

Having carried out a statistical evaluation on the results of the rating exercise, the table in figure 10 shows the median, mean and standard deviation scores for the English speakers. The various voice options scored well for understandability, with a median score of 4, the mean score was 4.27, and the standard deviation was 0.9 - as the median and mean are very close in value, and the standard deviation is low, this indicates that users generally agreed on this rating. There was 1 user who felt that the understandability was a lot poorer, and this rating was for the English default voice on Android.

There was a bit more of a disagreement in the ratings when describing the voice's language fluency, although the median score is also 4, the mean is lower at 3.73, and the standard deviation is the highest of the categories, indicating that experiences varied between participants. The lowest scoring voice for fluency is English Great Britain (Windows, Chrome), followed by English default (Android) and English Received Pronunciation (Linux), whereas the voices that were rated 5/5 by participants were the Microsoft created ones Hazel (UK) and Zira (US) for Windows. The ratings for pronunciation accuracy are similar, with the standard deviation also being relatively high at 1.17. The English default voice on Android had the lowest single rating again, while Microsoft Hazel was rated 5/5 by all participants who selected that voice.

However, the overall experience received lower ratings compared to the individual category scores, with both the median and mean being closer to 3. As discussed with the qualitative results, this seems to be due to the unnatural qualities of the voice, including intonation and rhythm of the voice while reading. The high standard deviation scores indicate that there is a wide variety of experiences depending on which voice options the user had available to them - with the Microsoft voices for Windows scoring better than the Android and Linux options.

	Understandability	**Fluency**	**Accuracy**	**Experience**
Median	4	4	4	3
Mean (2 d.p.)	4.27	3.73	3.82	3.18
Standard Deviation (2 d.p.)	0.90	1.27	1.17	1.17

Figure 10- Table showing Median, Mean and Standard Deviation values from English native speakers

German Results

In figure 11 statistical values that give some insight in the data gathered are portrayed, i.e. median, mean, and standard deviation.

The various voice options scored decent for understandability, with a median score of 4, the mean score was 3.57, and the standard deviation was 0.79. The standard deviation shows that users generally agreed on this rating (5 of 7 people gave a rating of 4).

In the fluency ratings more of a disagreement can be found, although the mean and median are both 3 the standard deviation is 1.15 and, thus, the highest of all of the categories. In this example it clearly shows how much the rating is dependent on a subjective opinion and why a comparison instead of a simple rating, further discussed in task 5, might have been more beneficial. Here we can observe that different people that chose the same voice option (Windows, Mozilla, DEUTSCH) have rated the fluency completely different, i.e. one person rated the fluency as a 2 and another person as a 5. Because of the low number of participants it is really hard to find a trend in this type of varied answers.

The ratings for the pronunciation accuracy have a median of 3 and a mean of 3.14. The standard deviation of 0.70 is the lowest of all the categories. The results indicate that the users experienced the pronunciation accuracy neither as excellent nor as poor.
Similar results, with a median of 3, a mean of 3.29, and a standard deviation of 0.76, can be seen for the experience using TTSReader in general.

	Understandability	**Fluency**	**Accuracy**	**Experience**
Median	4	3	3	3
Mean (2 d.p.)	3.57	3	3.14	3.29
Standard Deviation (2 d.p.)	0.79	1.15	0.70	0.76

Figure 11 - Table showing median, mean, and standard deviation of the evaluation categories for the German language participants

Russian Results

Statistical values shown in figure 12 give some insight into the data gathered from the Russian speaking participants. The median for understandability and fluency is 4.5, for accuracy is 4, which demonstrates that the Russian speech was understandable, fluent, and accurate for most of the Russian speaking participants. For the overall experience of using the system the median is only 2 which shows that many participants were not very satisfied with their experience interacting with the TTSReader system. The mean value which is similar to the median value: for understandability and fluency is 4, for accuracy 3.5, and for experience 2.333.

The standard deviation shows how the values are dispersed around the mean and calculated as the square root of the variance. The standard deviation for Russian speaking participants shows that understandability has some diverse marks from different users. The value of the standard deviation is 1.549, which demonstrates a big deviation. The standard deviation is also related to the fact that only 6 Russian participants were using the system and one participant marked the understandability with the lowest mark 1 where the rest marked it with the highest marks (3 participants marked 5 and 2 participants marked 4). The

same standard deviation value for the fluency 1.549 can be explained the same way as for the fluency. The standard deviation for the accuracy is 1.378 which is less than for understandability and fluency, which can be explained that more participant evaluation marks are closer to the mean value. The experience standard deviation is 1.366 which means the experience mean value is closest to the most participants' evaluations.

	Understandability	**Fluency**	**Accuracy**	**Experience**
Median	4.5	4.5	4	2
Mean (2 d.p.)	4	4	3.5	2.333
Standard Deviation (2 d.p.)	1.549193	1.549193	1.378405	1.36626

Figure 12 - Table showing median, mean, and standard deviation for the Russian language participants

Arabic Results

Figure 13 shows the results of the statistical analysis carried out on the results provided by the native Arabic speaking participants.It shows the Median, Mean, and Standard Deviation for the Arabic native speakers participants. Regarding the understandability for the Arabic participants, the median score is calculated to be 4, the mean score is calculated to be 3.67, and the standard deviation is 1.510. Having the mean and the median close in values, when the standard deviation is very low in value for the understandability it concludes that the participants had a good experience with the understandability of the arabic text to speech.

Regarding the fluency of the Arabic language, it can be seen in figure 13, the median is rated to be 3.5, the mean for the fluency is given to be 3.33 which is also very close in value to the median, and finally the standard deviation is also in very low value given to be 1.63. As seen in figure 8, shows that there are two options for the Arabic speaking participants, which are Arabic and العربية, Saudi Arabia - it can be developed that when the participants chose العربية, Saudi Arabia, the fluency understand is better and same for the Experience of the pronunciation of Arabic. Speaking of the Accuracy of the pronunciation, it showed a mid range understanding for the participants with a score of 3 for the median, where it also scored a close value of 2.67 for the mean, and finally 1.03 for the standard deviation.

Looking at the overall experience of the system for Arabic native speakers, no user gave full marks for their experience with TTSReader in Arabic. 50% of the users (3 users) gave the system a rating of 4/5 , where 33.3% of the users (2 users) gave a rating of 3/5 for the system. Finally, there is one user who was not satisfied at all and gave the system a rating for 1/5 . Looking at the statistics in the figure below for the experience, it scored a median of 3.5, it scored a mean of 3.17, and finally a standard deviation of 1.17. Overall, the overall statistics for the participants is satisfying while evaluated by users using different operating systems.

	Understandability	Fluency	Accuracy	Experience
Median	4	3.5	3	3.5
Mean (2 d.p.)	3.67	3.33	2.67	3.17
Standard Deviation (2 d.p.)	1.510	1.63	1.03	1.17

Figure 13 - Table showing Median, Mean and Standard Deviation values from Arabic native speakers

Overall Results

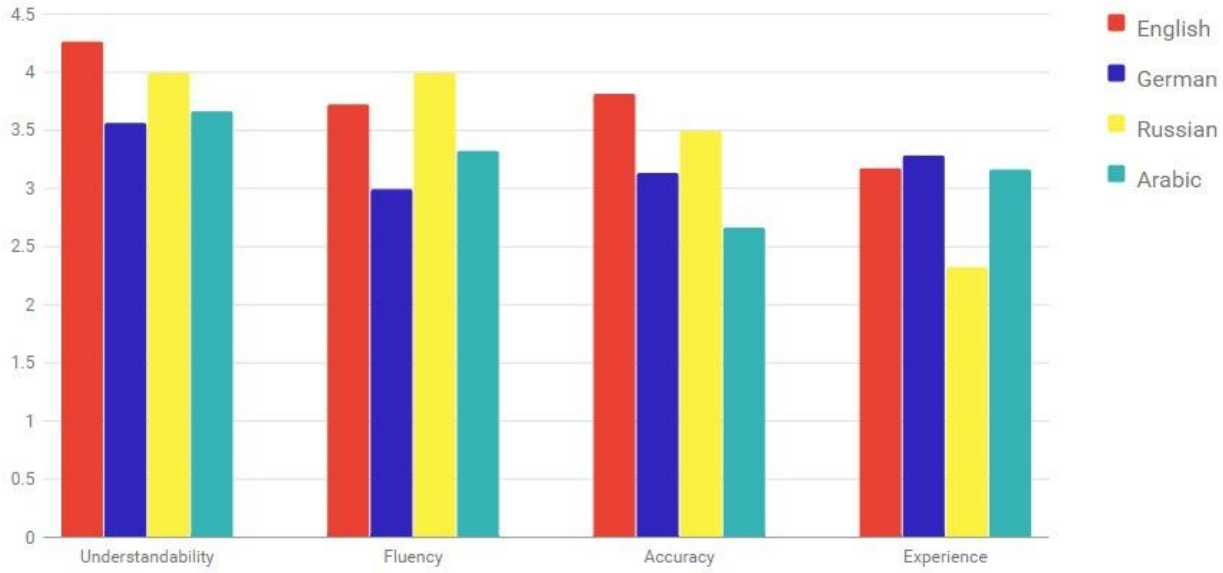

Figure 14 - Mean values for all four languages regarding all four categories

Wilcox p-value	Understandability	Fluency	Accuracy	Experience
German	0.05559	0.2071	0.08607	0.77
Russian	1	0.488	0.668	0.2335
Arabic	0.4171	0.7151	0.04174	0.7891

Figure 15 - Table showing p-value from the independent Mann-Whitney U Test comparing the English group with the other language groups

The independent Mann-Whitney U comparison test between English and non-English mean vectors shows non significant difference between the data (p-value = 0.11). The statistical R code can be found in Appendix 4: R Code for Statistical Analysis.

5) Conclusion

a. Limitations of the Evaluation

A major challenge that we were facing was the amount of independent variables, especially the number of different voice options available. The voice options vary strongly in their quality. The difficulty was that each operating system had a different set of voice options. Additionally, the browser used also influenced the number of voices.

So, we had to make the decision to either restrict the number of voices to make the evaluation more comparable. This would introduce a number of problems such as limiting the evaluating users to one specific operating system and one specific browser which would significantly reduce the number of people that would be able to do the survey. With this restriction we would, furthermore, only have representative data for this small group of possible users. Another problem with this approach would be the establishment of a bias towards whatever subset of voices we would have selected. Which might skew the result significantly, e.g. the English voices on windows chrome might be better than the German ones but on Mac it might be the other way around.

For this reason we decided to leave the decision to the user. Which has other disadvantages. First, we now have a large amount of different voices. With the amount of survey participants some voices might be underrepresented and, therefore, fluctuate strongly based on a singular opinion. And some voice options might not be covered at all. This can be problematic as the quality between the different voice options might vary significantly.

Secondly, to differentiate between the different voices we have to ask the survey's participants which operating system and browser they used. Additionally, we had to ask the participant to write down which voice they used (as some operating system and browser combinations have different voices but the same name). Some voices, e.g. DEUTSCH, MICROSOFT HEDDA DESKTOP - GERMAN, have a really long name and it might be tedious and error prone to type it and there is no copy option out of the drop-down menu in the TTSReader interface. So, we have to rely on the user to accurately type it and not just for convenience reasons write Deutsch (which might be another voice).

After conducting the experiment we can observe exactly that behaviour. For example, for the combination Windows and Edge the participant entered "German" which is not a listed option. So it is not possible to find out which voice option exactly was used. For other participants the capitalization differed, or additional information like selected speed option was given. So, we have to make assumptions on which voice option they used.

Another issue might have been evolved due to the experiment design. As we have chosen a between-subjects design, i.e. each participant only evaluates one voice. Therefore, the participants do not have a comparison between different languages or voices. Therefore, our experiment might be prone to subjectivity in a sense of some users being more critical in general than others.

Furthermore, we noticed that some voices were not working. For example, Russian on iPhone devices was available but not reading out the text.

b. Reflection of Results

The standard deviation data analysis shows that German and English native speakers evaluations are not as dispersed as the Russian and Arabic. This can be explained by the fact that participants from Russian and Arabic groups experienced more problems using the TTSReader and placed lowest possible marks available to evaluate the system, where German and English participants were able to find the voice options that would work for them.

The amount of the participants is relatively small and some inaccuracies in participants' evaluation are possible, but still we can see that the mean value for the English group is higher than for other language groups, which means that English participants could appreciate the TTSReader better than participants from other groups.

The significant difference is found for the Accuracy of Arabic and English text comparison (p-value:0.04174), another big difference (p-value:0.05559) is in case of comparison of understandability of the German and English text reading. The p-value is not 0 for any of the tests, which means that there is a difference between English and non-English group evaluations, but this difference is not significant.

The comparison test between English and non-English mean vectors shows non significant difference between the data (p-value = 0.11). The p-value = 0.11 > 0.05 demonstrates that we have failed to reject our null hypothesis. There is no significant difference in system evaluations between English speakers and non-English speakers means we cannot approve our main hypothesis, but based on this research we can make another experiment. We need to note that the mean value for English speakers is 3.75 and for non-English speakers is 3.25 which is 10% difference, but this is still not enough to prove our hypothesis. The result may encourage us to do the experiment again with lessons learned from this experiment.

c. Potential Improvements for Future Work

The limitations that had a large impact regarding the set-up of the study was the number of independent variables that we came across while creating the survey, mainly the variety in the availability of voice options across devices. This has made it very difficult to accurately compare the user experience - as noticed by our participants, Arabic was not available as a voice option on Windows, but was available on MacOS, and English UK Hazel on Windows received higher ratings compared to English UK Daniel on MacOS. This is due to participants having to use their own devices to take part in this study due to Coronavirus limitations.

A future study could involve a pilot study, in order to find out which voice options were highly rated by native speakers, before conducting our study using these voices. Ideally, this study would also be carried out in person, using the same web browser and operating system for every participant (depending on what results the pilot study provides) which will also reduce

the variables involved - including ensuring that the system would work for everyone, since Russian had issues with it appearing as an option but not working on iPhones. Instead of an online survey being used, in-person interviews would provide more opportunities for qualitative data on the participants' experiences using TTSReader - including more comments on the useability of the web app, as well as being able to press for further comments regarding the understandability and fluency of the available voices.

As carrying out rating tasks is still important for carrying out a quantitative analysis using statistics, the fact that a 5-point likert scale was used could reduce the number of options that participants will generally choose to 3, due to tendencies to not select the highest or lowest options. For a follow-up study, a likert scale with 7 options will be used to effectively provide participants with 5 options that are in the middle of the scale, to get more variation in the ratings assigned.

Participants were also only asked to rate a single voice option for their native language in this study - this was decided due to the potential limitations of voice availability, and that some participants may not have a choice available. The ratings can be affected by participant subjectivity, as some participants may rate voices harsher than others. By providing participants with a comparison exercise, this can result in more effective ratings, as well as potentially receiving more information as to why they rated the voice option in that way.

An issue was brought up regarding the title of our survey, which was our research question "Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language?". The fact that we were including effectively our hypothesis in the title of the survey, with the inclusion of the word "better", could potentially shape our participants' experience using TTSReader. However, as this was brought up only after we'd received some responses, it was decided that to change the title during the experiment would have an effect on subsequent results. More consideration would be put towards the wording of the title, as well as the rest of the information provided to participants to lessen the impact of unconscious biases.

Another lesson learned by our team is to plan the hypothesis testing when writing our hypotheses down at the start of the process. As we did not have a concrete plan for the statistical analysis to be carried out in advance of the results being collected, this made conducting a statistical analysis more difficult than if this had been discussed and finalised earlier. For future studies this would be decided in advance, and would not change once the survey closes.

References:

1. Agnes Ogee et al., The Minitab Blog. Best Way to Analyze Likert Item Data: Two Sample T-Test versus Mann-Whitney., 2020, available at: https://blog.minitab.com/blog/adventures-in-statistics-2/best-way-to-analyze-likert-item-data-two-sample-t-test-versus-mann-whitney
2. de Winter, J.C.F. and D. Dodou, Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon, *Practical Assessment, Research and Evaluation*, 2010

Appendices

Appendix 1: Ethics Forms

Annex A - Ethics Checklist

RESEARCH ETHICS REVIEW CHECKLIST

The University Research Ethics & Governance Framework applies to all aspects of research undertaken within the University, including research undertaken by undergraduate and postgraduate students as a part of their dissertation, thesis or coursework. All academic staff, PGT, PGR students and UG students therefore should consider the ethical dimensions of their research using the self-assessment checklist below and, where necessary, seek ethical review and approval.

Ethical review must be sought for any project that answers **YES** to any one or more of the questions on this page:

Applicant Name: Jasmin Uhlhorn, Jess McGowan, Vladimir Yesipov, Yassin Dinana	**Project Title:** Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language?

	YES/NO
Does the project involve human participants?	YES
Does the project involve personal data? (Choose NO if the only personal data involved is from an existing dataset and: (a) the participants who originally provided the data have given permission for it to be used in further research, or (b) the dataset is publicly available.)	YES
Does the project involve human remains?	NO
Does the project involve surveys or questionnaires?	YES
Will the participants of the project include staff or students of the University, or colleagues or clients in a work environment?	YES
Does the project involve children (under 18 years) or vulnerable adults?	NO
Does the project involve any clinical procedure or involve clinical populations?	NO
Could participants experience physical or psychological harm or discomfort?	NO
Does the project involve the collection of material that could be considered of a sensitive personal, medical or psychological nature, or is constrained by other data protection requirements?	YES
Does the project involve concealment or deception or deliberately misleading participants?	NO
Does the project involve face-to-face interviews or the collection, preservation or use of sound and/or video material involving human participants?	NO
Does the project involve collecting personal data from websites or from social media (e.g., Facebook, Twitter)?	NO

Is there any potential for conflict of interest between research funder, investigators and/or participants that may affect funding, dissemination or other research outcomes?	NO
Could the project lead to financial gain for funders, investigators or participants?	NO
Is the research likely to have any significant detrimental or lasting impact on the environment? *This includes the natural environment but also buildings and structures created by people, especially ones of historical or archaeological importance.*	NO
Does the project give rise to a realistic risk to the national security of any country?	NO
Does the project involve the collection of genetic resources (Nagoya Protocol)?	NO

CONFIRMATION BY APPLICANTS:

I confirm that I have discussed this checklist with my supervisor. *(For students only)* ☑

I understand that if during my project the answers to any of the above questions change, I must complete a new checklist and seek ethical review if necessary. ☑


FOR FUNDED RESEARCH ONLY

Do you have, or are you applying for, external funding?

YES/NO

If Yes, give the name of the funder.

Does the research funder require the project to undergo ethical review?

YES/NO

Do you need written evidence that your project has undergone ethical review?

YES/NO

If Yes, please fill out these details:

FOR STUDENTS ONLY

In the event an ethical review is required, please complete the Physical Sciences & Engineering research Ethics Review and Application Form (Annex B). If you are a student then you should do this in consultation with your dissertation supervisor or course co-ordinator. You may also need to attach other documents such as a Participant Information Sheet, a Consent Form or a schedule of interview questions.

Most reviews will be undertaken by circulation to appropriate reviewers within the University. The outcome of the review will be communicated to you by the Clerk of the Ethics Board as soon as practicable.

For further information please contact the Clerk to the Ethics Board in the first instance, copsethics@abdn.ac.uk.

Physical Sciences & Engineering Ethics Review and Approval Form

IMPORTANT NOTE: Research projects cannot begin until ethical approval has been granted.

Please complete the relevant sections of this form if, after filling out the relevant ethical review checklist (Annex A), you have identified a potential ethical issue. Please send the completed form and supporting documentation to copsethics@abdn.ac.uk.

Name: **ID number:** ***(for students only)***	Jasmin Uhlhorn, Jess McGowan, Vladimir Yesipov, Yassin Dinana 52095681, 52093344, 52093351, 52094480
School: **Department or discipline:**	Natural Sciences & Computing Computing Science
Programme (e.g., PhD, MSc): ***(for students only)***	MSc Artificial Intelligence
[1]Project Title: **Course Number and Name:** ***(for students only)***	Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language? CS5063 Evaluation of AI Systems
Names of other individuals involved in the research/project?	
Name and email address of main supervisor: ***(for students only)***	Ehud Reiter & John Vargheese & James Robert Forest
Application date:	29th October 2020

Please note:

1. Research involving NHS staff, patients, facilities and premises is subject to ethical review by the NHS North of Scotland Research Ethics Service. This includes research involving individuals when their status as NHS staff or patients is relevant to the research, even when a medical condition is not the subject of the research. Research involving adults who do not have the capacity to consent is also subject to these ethical review procedures.

2. Research involving animal and biological materials is subject to Home Office regulations. Forms and guidance can be obtained from the university's Research Governance Section (researchgovernance@abdn.ac.uk).

[1] Project = the **particular piece of work for which you are applying for ethical approval** (not your overall programme of research)

3. Research involving the collection of genetic resources (organisms, microorganisms, DNA, RNA, proteins, small molecules) from signatories to the Convention on Biodiversity/Nagoya Protocol requires a formal agreement to be in place before this research can begin. Contact your Business Development Officer for further guidance (https://www.abdn.ac.uk/staffnet/secure/research-grant-funding-2405.php#business-development-team-)

CHECKLIST

The purpose of this checklist is to make sure no information has been inadvertently left out and to allow reviewers to assess the application more quickly. If you do not complete the checklist and attach a completed Annex A, the application will be returned to you.

I confirm that if my project changes significantly then I will notify the Ethics Board. ☒

I have **attached** a completed checklist (Annex A). ☒

I confirm that I have discussed this application with my supervisor. ☒
(for students only)

I have completed the University's online ethics training. ~~YES/~~NO
(for Staff and PGR students only. NB: PGT students will undertake this training at the discretion of their Programme Coordinator.)

This project requires me to **travel outwith the UK** ~~YES/~~NO
If YES, please provide the following confirmation:-

- I will comply with the requirements of the University's Overseas Travel Policy, including obtaining permission to travel (where required by the policy), completion of a risk assessment and will obtain University travel insurance cover. ☐

Other Attachments (delete YES/NO as appropriate):

I have attached a Participant Information Sheet. YES/~~NO~~

I have attached a Consent Form. YES/~~NO~~

I have attached a schedule of questions for surveys and/or interviews. YES/~~NO~~

Section 1: Research projects involving human participants (not NHS staff or patients)

If you answered `No' to Q1 and/or Q2 then omit Q 3 – 18 and proceed straight to Q19. In this case, please explain your answers to Q1 and/or Q2 in Section 7.

Recruitment Procedures

		Yes	No
1	Does your project involve human participants? This includes use of surveys, questionnaires, on-line surveys and tests, focus groups and workshops where human participants provide information or data to inform the research.	**Yes**	
2	(a) Does your project involve human remains? (b) If so, does your work conform with the Historic Environment Scotland guidelines?		**No**
3	Does your project involve people less than 18 years of age?		**No**
4	Does your project involve people with learning or communication difficulties?		**No**
5	Is your project likely to involve people involved in illegal activities?		**No**
6	Does your project involve people belonging to a vulnerable group, other than those listed above?		**No**
7	Does your project involve people with whom you have, or are likely to have, a working or professional relationship: for instance, staff or students of the university, professional colleagues or clients?	**Yes**	
8	Does your project involve people who do not have English as their first language?	**Yes**	
9	Does your project require the recording of audio or video of participants or of others not involved in the research?		**No**
10	Do you plan to conceal your own identity during the course of your project?		**No**

*Please explain in **Section 7** how you will recruit your participants. If you answered '**Yes'** to any of the above questions, please give details.*

*If you answered '**Yes'** to **Q1** then you must provide a Participant Information Sheet and a Consent Form. For web-based research, screenshots of the appropriate web pages suffice.*

If your project involves surveys or interviews then you must provide a schedule of questions.

*If you answered '**Yes'** to **Q3, Q4** or **Q6** then you may need to apply for disclosure through Disclosure Scotland.*

Consent Procedures

	Yes	No

11	Do you have set procedures that you intend to use for obtaining informed consent from all participants, including (where appropriate) parental consent for children?	**Yes**	
12	Will you tell participants that their participation is voluntary?	**Yes**	
13	Will you obtain written consent for participation, including for audio and/or video recording?	**Yes**	
14	Will you tell participants that they may withdraw from the research at any time and for any reason?	**Yes**	
15	Will you give potential participants a period of time to consider participation?	**Yes**	
16	Does your project involve concealment or deliberately misleading participants?		**No**

*Please explain in **Section 7** how you will obtain consent from participants. If you answered '**Yes'** to **Q16** or '**No'** to any of the other questions, please give details.*

Possible Harm to Participants

		Yes	**No**
17	Is there any realistic risk of any participants experiencing physical or psychological harm or discomfort?		**No**
18	Is there any realistic risk of any participants experiencing a detriment to their interests as a result of participation?		**No**

*If you answered '**Yes'** to either question, please explain in **Section 7** how this risk was assessed and how you propose to manage it.*

Section 2: Data protection, handling and storage

IMPORTANT NOTE:

The General Data Protection Regulation imposes a number of obligations for the use of **personal data** (defined as any information relating to an identified or identifiable living person), or including the use of personal data in research.

If you are using personal data, you should consider whether your research requires a Data Protection Impact Assessment and complies with the University Data Protection policy.

If you are, you now need to see the Data Protection Checklist for Researchers[2] for guidance.

If you then feel that a DPIA may be required or you need data protection advice, then you should contact the Data Protection Officer dpa@abdn.ac.uk.

Please provide the following confirmation:

I have read the above guidance and have met the relevant data protection obligations.

☒ **Please tick the box to confirm**

		Yes	No
19	(a) Will any non-anonymised and/or personalised data be generated and/or used? (b) Will you use an existing dataset in your research? (c) If 'yes', do you have permission to do so?		**No***
20	Will any data be stored (temporarily or permanently) anywhere other than on password-protected University computers or servers?	**Yes***	
21	Will you gain access to sensitive[3] data about living individuals or organisations that is not already publicly available elsewhere? If 'Yes', will you gain the consent of the individuals concerned?		**No**
22	Does your project require access to personal data about participants from other parties (e.g., teachers, employers), databanks or files? *If yes, please explain in Section 7 how you will gain the consent of these participants.*		**No**
23	Does the project involve collecting personal data from websites or from social media (e.g., Facebook, Twitter)?		**No**
24	Will the data be stored, collected or accessed from: - outside the UK? - outside the EU?	**Yes***	
25	Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory or extremist?		**No**

[2] Click on 'Guides' to find the checklist

[3] Sensitive data includes data that relates to racial or ethnic origin, political opinions, religious beliefs, trade union membership, physical or mental health, sexual life, actual and alleged offences.

* Details are specified in section 7

	If yes, see here for an explanation of the obligations of the researcher and the university under the Prevent duty.		
26	Are there any contractual conditions attached to working with or storing the data? (E.g., an HSCIC data sharing agreement.)		**No**
27	Could working with this data damage the University's reputation? (E.g., bad press coverage, public protest.)		**No**
28	Could working with this data cause an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.)		**No**

For further advice on Data Protection, please refer to www.abdn.ac.uk/dataprotection**.**
For further advice on GDPR, please refer to
https://www.abdn.ac.uk/staffnet/governance/academic-research-data-protection-7195.php

Please provide details in **Section 7** of how you intend to ensure that data is stored securely and in line with the requirements of the Data Protection Act and funding bodies (if applicable). Please give specific consideration to whether any non-anonymised and/or personalised data will be generated and/or stored and what precautions you will put in place regarding access.

If you answered '**Yes'**, to any of the questions above, please give details in **Section 7**.

Section 3: Research involving possible harm to the environment

		Yes	**No**
29	Is the research likely to have any significant detrimental or lasting impact on the environment? *This includes the natural environment but also buildings and structures created by people, especially ones of historical or archaeological importance.*		**No**

If you answered '**Yes'**, please explain in **Section 7** how this risk was assessed and how you propose to manage it. Say whether relevant guidelines exist in your discipline, and whether you intend to follow them.

Section 4: Research which may have an adverse impact on national security

		Yes	**No**
30	Does your project give rise to a realistic risk to the national security of any country?		**No**

If you answered '**Yes'**, please give details in **Section 7**. Explain how this risk was assessed and how you propose to manage it.

Section 5: Funding and conflict of interest

		Yes	No
31	Is your project funded by the university or an outside organisation, or have you applied for funding?		**No**
32	Is there any potential conflict of interest between research funder and researchers or participants and researchers which may potentially affect the research outcome or the dissemination of research findings?		**No**
33	Might the project lead to financial gain to funders, investigators or participants?		**No**

If you answered '**Yes'** to any question, please give details in **Section 7.** Explain any potential conflict of interest and how you propose to manage it.

Section 6: Collection of genetic resources

		Yes	No
34	Does the project involve the collection of genetic resources (organisms, microorganisms, DNA, RNA, proteins, small molecules) from signatories to the Convention on Biodiversity/Nagoya Protocol?		**No**

If you answered '**Yes'**, then a relevant agreement must be in place before the research can begin. This agreement must provide prior informed consent with mutually agreed terms and the must be in keeping with the Convention on Biodiversity/Nagoya Protocol and be obtained via the national focal point of the provider country. Please explain in **Section 7** how you propose to arrange the agreement. Please indicate the need for confidentiality where appropriate.

Section 7: Additional Information

All questions must be answered fully in the space provided (however each box can be expanded as necessary). Incomplete or incorrectly completed forms will be returned to the applicant, delaying the process of obtaining ethics approval.

7.1	**Project description** *Please attach a project descriptor or summary document (where available)*	The project will be an evaluation on the TTSReader (https://ttsreader.com/). The evaluation will be conducted via a user survey and focuses on the research question of whether the tool provides a better experience for native English speakers than non-native English speakers using their native tongue. The project includes an interpretation and a statistical analysis based on the gathered data.
7.2	**Start date and duration**	Start Date: 2nd November 2020 Duration: 7 days (2nd November 2020 - 8th November 2020)
7.3	**Methodology**	The participants will be asked to evaluate TTSReader with text in their native language, and will be rating the output based on accuracy and fluency. This rating exercise will be completed on

		Google Forms*, using a likert scale for ratings, and the form will not require users to log in or provide personal information outside of their native language, which is required as the topic of this study. *Alternatives that can be used if Google Forms is not appropriate are Microsoft Forms or Survey Monkey.
7.4	**Recruitment of participants**	Convenience sampling techniques will be used, with recruitment from friends and acquaintances to fill out an online survey. Due to the nature of this study, these users may not be native English speakers, however we expect them to be able to complete this study. Due to this recruitment of participants, we expect that some participants will come from outside of the UK and/or outside of the EU.
7.5	**Consent**	Using university consent forms, this will be displayed before participants can fill out the online survey, and participants will be able to withdraw from the survey at any point prior to submission, this will be made clear on the form.
7.6	**Harm to participants**	We do not foresee any realistic risk of harm to participants. We will be selecting text for participants to use as input for TTSReader, to mitigate the potential for any offensive text being used.
7.7	**Data storage**	Data capture will be anonymised and stored on 3rd party servers temporarily during data collection, before being moved onto university servers, and will be deleted after the study is concluded.
7.8	**Ethical considerations** *Concise statement of the ethical considerations raised by the project and how you intend to deal with them. Include details related to Sections 3-6 above, if applicable.*	As we are using human participants in this study, we have had to carefully examine the ethics of this study. The only personal information that we require from participants is their native language, as this is the focus of the study, and this will be collected anonymously online. The survey will contain the Participant Information sheet as well as a consent form which will be shown at the start of the survey for each participant, and participants will be given the option to opt out of submitting before the survey is completed. The data collected will not be stored for longer than is necessary, once data collection is complete it will be stored on university servers, and then deleted upon completion of the study. A sample of the text that we will be including for participants to use is included below (https://wordpress.com/create/): "WordPress.com allows you to build a website that meets your unique needs. Start a blog, business site, portfolio, online store, or anything else you can imagine. With built-in optimization and responsive, mobile-ready themes, there's no limit to who you can reach with your new website. Create a simple website for your family or sell products around the world—it's up to you." Mit WordPress.com kannst du eine Website erstellen, die deinen ganz persönlichen Bedürfnissen entspricht. Starte ein Blog, eine Firmenwebsite, ein Portfolio, einen Onlineshop oder etwas ganz anderes – deiner Fantasie sind keine Grenzen gesetzt.

		Dank integrierter Optimierung und responsiven, mobilfreundlichen Themes kannst du mit deiner neuen Website jede Zielgruppe erreichen, die du willst. Erstelle eine einfache Website für deine Familie oder verkaufe Produkte auf der ganzen Welt – es liegt ganz bei dir. WordPress.com позволит создать веб-сайт для решения именно ваших задач. Вы сможете сделать из него всё, что захотите: блог, корпоративный ресурс, онлайн-магазин или что-то ещё, что вам подскажет воображение. Встроенные средства оптимизации и гибкие темы для мобильных устройств сделают ваш сайт доступным для всех пользователей. Вы сможете создать простой сайт для своих близких или продавать товары пользователям со всего мира. Что делать с этими возможностями — решать вам. تسمح لك ببناء موقع إلكتروني يلبي احتياجاتك الفريدة. ابدأ مدونة أو موقع أعمال أو محفظة أو متجر على الإنترنت أو أي شيء آخر يمكنك تخيله. مع إضفاء الطابع الأمثل على المواضيع والاستجابة الجاهزة للتنقل، لا يوجد حد للذين يمكنك الوصول إليه مع موقعك الشبكي الجديد. قم بإنشاء موقع شبكي بسيط لعائلتك أو بيع منتجات في جميع أنحاء العالم، الأمر متروك لك

For any contractual or intellectual property questions, please contact the business development team in Research & Innovation (june.middleton@abdn.ac.uk).

Make sure you have completed the checklist on page 2.

FOR STAFF AND PGR STUDENTS: Please send your completed application form and any supporting documentation to copsethics@abdn.ac.uk.

FOR PGT STUDENTS: Please refer to the ethical review procedures outlined in your Project Guidelines or contact your Postgraduate Programme Coordinator for further advice.

Annex C - Participant Information Sheet

The School of Natural Sciences & Computing
computingscience@abdn.ac.uk
University of Aberdeen, King's College, Aberdeen, AB24 3FX

PARTICIPANT INFORMATION SHEET

Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language?

Principal Investigators: Jasmin Uhlhorn, Jess McGowan, Vladimir Yesipov, Yassin Dinana
Other researchers: n/a
Supervisors: Ehud Reiter & John Vargheese & James Robert Forest

We are MSc Artificial Intelligence students in the Department of Natural Sciences and Computing. We would like to invite you to consider participating in the research project: Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language?. Below is some information about the project, to help you decide whether you would like to take part.

Participation in the research project is completely voluntary. You can withdraw from the project at any time, without having to give a reason.

AIMS

The aim of the project is to test the hypothesis that native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language. We suppose that our hypothesis is true because the TTSReader interface is in English and all tools available on the TTSReader web site are described in English. Thus, the information gathered aims towards the goal of evaluating that hypothesis. No personal information will be gathered except for the native language of the participants. As the outcome of the research we expect to get the TTSReader system evaluation with the questionnaire provided from the participants and analyse this data according to the native language group of each participant.

WHAT YOU WILL BE ASKED TO DO

We would like to ask each participant to answer a questionnaire available online on the Google forms. The survey will be open from 2nd November 2020 to 5th November 2020, and we do not expect the survey to take longer than 15 minutes.

RISKS

We do not foresee any realistic risk of harm to participants. We will be selecting text for participants to use as input for TTSReader, to mitigate the potential for any offensive text being used. If at any point you no longer wish to continue in this study prior to completing this survey, you can close this browser tab and any information that you have entered so far will be deleted.

DATA MANAGEMENT AND STORAGE

Data capture will be anonymised and stored on 3rd party servers temporarily during data collection, before being moved onto university servers, and will be deleted after the study is concluded.

CONFIDENTIALITY AND ANONYMITY

The University's Privacy Notice for Research Participants is available here

Raw data and the identity of participants will not be released to anyone outside the research team. The data you provide will be analysed and may be used in publications, dissertations, reports or presentations derived from the research project, but this will be done in such a way that your identity is not disclosed. This survey will not record any personal information outside of what native language you speak, and this will be collected anonymously through this survey.

CONSENT

If you agree to take part in the research, you will be asked to indicate your consent by ticking a box on an online Consent Form.

SPONSORS

n/a

Thank you for considering taking part in this research.

If you have any questions about this research please contact us:

Jess McGowan		j.mcgowan1.20@abdn.ac.uk
Jasmin Uhlhorn		j.uhlhorn.20@abdn.ac.uk
Vladimir Yesipov		v.yesipov.20@abdn.ac.uk
Yassin Dinana		y.dinana.20@abdn.ac.uk

For any queries regarding ethical concerns you may contact the Convener of the Physical Sciences & Engineering Ethics Board at the University of Aberdeen:

Email: copsethics@abdn.ac.uk

This research project was approved by the Physical Sciences & Engineering Ethics Board on 30th October 2020.

Annex E - Participant Consent Form (Online)

Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language?

Consent form for participation in the research project *(Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language?).*

Please read the statements below and tick the final box to confirm you have read and understood the statements and upon doing so agree to participate in the project.

I confirm that the research project ***(Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language?)*** has been explained to me. I have had the opportunity to ask questions about the project and have had these answered satisfactorily.

I consent to the material I contribute being used to generate insights for the research project (***Do native English speakers have a better experience with TTSReader compared to non-native English speakers using TTSReader in their native language?***).

I understand that my participation in this research is voluntary and that I may withdraw from the research at any time.

I consent to allow the fully anonymised data to be used for future publications and other scholarly means of disseminating the findings from the research project.

I understand that the information/data acquired will be securely stored by researchers, but that appropriately anonymised data may in future be made available to others for research purposes. I understand that the University may publish appropriately anonymised data in its research repository for verification purposes and to make it accessible to researchers and other research users.

- I confirm that I have read and understood the above statements (check the box).

Appendix 2: Resulting table

Column1	OS	Browser	Language	Voices	Understand	Fluency	Accuracy	Experience	Comments
1	iOS for iPhones/ iPads	Safari	Russian	R	5	5	5	2	No comments
2	Mac OS	Safari	Arabic	العربية saudi arabia	5	5	4	4	No comments
3	Windows	Firefox	English	Microsoft Hazel	5	5	5	5	No comments
4	iOS for iPhones/ iPads	Safari	Russian	РОССИИ	1	1	1	1	No comments
5	Windows	Edge	German	German	4	4	3	4	No comments
6	Mac OS	Chrome	Arabic	Arabic, Saudi Arabia	5	4	3	4	No comments
7	Android	Samsung	English	English UK	5	4	3	3	No comments
8	Android	Firefox	German	deu-deu	4	2	2	3	No comments
9	Mac OS	Safari	English	English, Daniel	4	4	3	3	Voice very monotonous and I thought it treated a comma in a way I would have expected a full stop which could lead to misunderstanding.
10	iOS for iPhones/ iPads	Safari	Arabic	arabic	4	2	2	3	The arabic pronunciation was very vague and it was not very well understood, some words were pronounced wrong.
11	Windows	Chrome	German	DEUTSCH, G **	3	2	3	3	No comments

12	Windows	Opera	English	ENGLISH, UK, MICROSOFT HAZEL DESKTOP - ENGLISH (GREAT BRITAIN)	5	5	5	4	The pronunciation and understandability of the reading was perfect but the voice I selected sounded very obviously unnatural which partially takes away from an otherwise perfect experience.
13	Windows	Chrome	Arabic	English, UK	1	1	1	1	I copied in the text, the system only read '1863' in english. The system was unable to read any of the Arabic text. Furthermore, the option of choosing Arabic as the system's language was not available.
14	Linux	Firefox	English	English U.K. Received Pronounciation	4	3	4	2	While the speech is understandable the 'Hawkingness' of the speech is alienating and offputting. The speed ((fast) and cadence of the speech is really weird, normally speech doesn't move at a constant speed.
15	Windows	Opera	English	English, US, Microsoft Zira Desktop - English (United States)	4	5	4	3	Very poor audio quality
16	Windows	Firefox	English	English, UK, Microsoft Hazel Desktop English (Great Britain)	4	4	5	5	No comments
17	Mac OS	Safari	German	DEUTSCH	4	3	4	4	No comments
18	Windows	Firefox	German	DEUTSCH	4	5	3	4	Firmenname als Wort, 1863 nicht als Jahr und 16000 nicht als Zahl
19	Windows	Chrome	Arabic	n/a	3	3	3	3	Could not find Arabic voice

20	Android	Chrome	English	English default	2	2	1	1	Very poor - voice was strangely accented that made it nearly incomprehensible. Some words were unrecognisable
21	Mac OS	Safari	English	English, UK, Daniel	5	4	4	3	Microsoft voice was not available. Daniel (+Default) Voice same and a bit robotic.
22	Mac OS	Safari	Arabic	(Saudi Arabia), العربية	4	5	3	4	No comments
23	Windows	Firefox	German	Deutsch, normal speed	4	3	4	3	It is barely possible correctly understand numbers or abbreviations; besides that understandability is good
24	Mac OS	Firefox	English	English UK Daniel	5	4	4	3	The intonation and rhythm/stress patterns are off, which makes it difficult to parse, and very unlike naturally spoken language. Pausing to consider what the inflection *should* be, when listening to a complex text, makes it easy to miss the total meaning.
25	Windows	Chrome	Russian	России, G**	5	5	4	4	No comments
26	Windows	Edge	Russian	РОССИИ, MICROSOFT EKATERINA ONLINE	4	5	4	2	Only the first sentence was read in Russian, and then pronounced only the numbers from the text fragment in English
27	Windows	Chrome	Russian	РОССИИ **	5	4	3	1	No comments
28	Windows	Chrome	English	English Great Britain	4	1	4	3	No comments

29	Windows	Chrome	Russian	Roccia G	4	4	4	4	1.It was pleasant to listen to, didn't seem bad. 2. It was quite fluent and went from word to word well 3.The reading was fine. 4.TTS reader seems to be a fine program.
30	Windows	Firefox	German	Deutsch	2	2	3	2	No comments

Appendix 3: Survey Questions

Introduction section:
Includes Annex C, followed by Annex E and this confirmation:
I confirm that I have read and understood the above statements (check the box). ▢

About you & your system
Which of these languages is your native language?

- English
- German
- Russian
- Arabic
- Other

Due to variances in text to speech options, the following information is requested:
Which operating system are you using?
List: Windows, Mac, Linux, Android, iOS

Which browser are you using?
List: Edge, Chrome, Firefox, Safari, Opera, Other

Example section: English

For this exercise, we would like you to use ttsreader.com, on one of the English voice options - e.g. English UK Daniel, or ENGLISH, UK, MICROSOFT HAZEL DESKTOP - ENGLISH (GREAT BRITAIN).

Please enter which voice option you selected (if there is none available for your native language please enter 'n/a':

Please copy the following paragraph into the text box on TTSReader.com, and press play:

Established in 1863, the ICRC operates worldwide, helping people affected by conflict and armed violence and promoting the laws that protect victims of war. An independent and neutral organization, its mandate stems essentially from the Geneva Conventions of 1949. We are based in Geneva, Switzerland, and employ over 20,000 people in more than 80 countries. The ICRC is funded mainly by voluntary donations from governments and from National Red Cross and Red Crescent Societies.

In terms of understandability, how would you rate the reading?
Poor Bad Okay Good Excellent

In terms of language fluency, how would you rate the reading?
Poor Bad Okay Good Excellent

In terms of pronunciation accuracy, how would you rate the reading?
Poor Bad Okay Good Excellent

How would you evaluate your experience using TTSReader?
Poor Bad Okay Good Excellent

Feel free to add any comments on how you found TTSReader below:

Final page before submission:

I confirm that once I submit this form, the data that I have provided can be used as part of the study and I can no longer withdraw. ▫

Sample text in German:
Das im Jahre 1863 gegründete IKRK ist weltweit tätig. Es leistet von Konflikt und bewaffneter Gewalt Betroffenen Hilfe und fördert die Rechtsvorschriften, welche die Kriegsopfer schützen. Das IKRK ist eine unabhängige und neutrale Organisation, dessen Mandat im Wesentlichen in den Genfer Konventionen von 1949 festgelegt wurde. Sein Sitz befindet sich in Genf, Schweiz, und es beschäftigt rund 16 000 Personen in über 80 Ländern. Das IKRK finanziert sich hauptsächlich aus freiwilligen Zuwendungen von Regierungen und Nationalen Rotkreuz- und Rothalbmondgesellschaften.

Sample text in Russian:
МККК, который был основан в 1863 году, работает по всему миру, оказывая помощь людям, пострадавшим в результате конфликтов и вооруженного насилия, а также распространяя знания о законах, защищающих жертв войны. Являясь независимой и нейтральной организацией, он обладает мандатом, предусмотренным, главным образом, Женевскими конвенциями 1949 года. МККК находится в Женеве, Швейцария, его сотрудниками являются примерно 16 тысяч человек более чем в 80 странах мира. Финансируется он, в основном, за счет добровольных пожертвований правительств и национальных обществ Красного Креста и Красного Полумесяца.

Sample text in Arabic:
منذ نشأتها عام 1863، كان هدف اللجنة الدولية للصليب الأحمر الوحيد هو حماية ضحايا النزاعات المسلحة والاضطرابات ومساعدتهم . وذلك عن طريق عملها المباشر عبر أنحاء العالم، وكذلك من خلال تشجيع تطوير القانون الدولي الإنساني وتعزيز احترامه من قِبَل الحكومات وجميع حاملي السلاح. وتعكس قصة اللجنة الدولية تطور العمل الإنساني واتفاقيات جنيف وحركة الصليب الأحمر والهلال الأحمر.

Sourced from: https://www.icrc.org/en/who-we-are, https://www.icrc.org/de/wer-wir-sind, https://www.icrc.org/ru/who-we-are, https://www.icrc.org/ar/who-we-are.

Appendix 4: R Code for Statistical Analysis

```
#Exporting table shown in Appendix 2 into the RStudio as ‘results_All’
#Creating subsets for each language
English <- subset(results_All, results_All$Language=='English', select = Understand:Experience)
Arabic <- subset(results_All, results_All$Language=='Arabic', select = Understand:Experience)
German <- subset(results_All, results_All$Language=='German', select = Understand:Experience)
Russian <- subset(results_All, results_All$Language=='Russian', select = Understand:Experience)

#Standard Deviation English
sd(English$Understand) 0.904534
sd(English$Fluency) 1.272078
sd(English$Accuracy) 1.167748
sd(English$Experience) 1.167748

#Standard Deviation German
sd(German$Understand) 0.7867958
sd(German$Fluency) 1.154701
sd(German$Accuracy) 0.6900656
sd(German$Experience) 0.7559289

#Standard Deviation Russian
sd(Russian$Understand) 1.549193
sd(Russian$Fluency) 1.549193
sd(Russian$Accuracy) 1.378405
sd(Russian$Experience) 1.36626

#Standard Deviation Arabic
sd(Arabic$Understand) 1.505545
sd(Arabic$Fluency) 1.632993
sd(Arabic$Accuracy) 1.032796
sd(Arabic$Experience) 1.169045

#Median English
median(English$Understand) 4
median(English$Fluency) 4
median(English$Accuracy) 4
median(English$Experience) 3

#Median German
median(German$Understand) 4
median(German$Fluency) 3
median(German$Accuracy) 3
median(German$Experience) 3
```

#Median Russian
median(Russian$Understand) 4.5
median(Russian$Fluency) 4.5
median(Russian$Accuracy) 4
median(Russian$Experience) 2

#Median Arabic
median(Arabic$Understand) 4
median(Arabic$Fluency) 3.5
median(Arabic$Accuracy) 3
median(Arabic$Experience) 3.5

#Mean English
mean(English$Understand) 4.272727
mean(English$Fluency) 3.727273
mean(English$Accuracy) 3.818182
mean(English$Experience) 3.181818

#Mean German
mean(German$Understand) 3.571429
mean(German$Fluency) 3
mean(German$Accuracy) 3.142857
mean(German$Experience) 3.285714

#Mean Russian
mean(Russian$Understand) 4
mean(Russian$Fluency) 4
mean(Russian$Accuracy) 3.5
mean(Russian$Experience) 2.333333

#Mean Arabic
mean(Arabic$Understand) 3.666667
mean(Arabic$Fluency) 3.333333
mean(Arabic$Accuracy) 2.666667
mean(Arabic$Experience) 3.166667

independent 2-group (English-German) Mann-Whitney U Test
wilcox.test(German$Understand,English$Understand) p-value:0.05559
wilcox.test(German$Fluency,English$Fluency) p-value:0.2071
wilcox.test(German$Accuracy,English$Accuracy) p-value:0.08607
wilcox.test(German$Experience,English$Experience) p-value:0.77

independent 2-group (English-Russian) Mann-Whitney U Test
wilcox.test(Russian$Understand,English$Understand) p-value:1
wilcox.test(Russian$Fluency,English$Fluency) p-value:0.488
wilcox.test(Russian$Accuracy,English$Accuracy) p-value:0.668
wilcox.test(Russian$Experience,English$Experience) p-value:0.2335

```
# independent 2-group (English-Arabic) Mann-Whitney U Test
wilcox.test(Arabic$Understand,English$Understand) p-value:0.4171
wilcox.test(Arabic$Fluency,English$Fluency) p-value:0.7151
wilcox.test(Arabic$Accuracy,English$Accuracy) p-value:0.04174
wilcox.test(Arabic$Experience,English$Experience) p-value:0.7891

# Comparing English with non-English
eng <- c(mean(English$Understand),mean(English$Fluency),
        mean(English$Accuracy),mean(English$Experience))
nonEng <- c(mean(mean(German$Understand),
              mean(Russian$Understand),
              mean(Arabic$Understand)),
          mean(mean(German$Fluency),
              mean(Russian$Fluency),
              mean(Arabic$Fluency)),
          mean(mean(German$Accuracy),
              mean(Russian$Accuracy),
              mean(Arabic$Accuracy)),
          mean(mean(German$Experience),
              mean(Russian$Experience),
              mean(Arabic$Experience)))

mean(eng) 3.75
mean(nonEng) 3.25

wilcox.test(eng, nonEng)
        Wilcoxon rank sum test
        Data: eng and nonEng
        W = 14, p-value = 0.1143
        Alternative hypothesis: true location shift is not equal to 0
```