Word-Gesture Typing in Virtual Realty

Bachelor thesis

Databases and Information Systems
Department of Mathematics and Computer Science
Databases and Information Systems
https://dbis.dmi.unibas.ch/

Examiner: Prof. Dr. Heiko Schuldt
Supervisor: Florian Spiess

Philipp Weber
phil.weber@stud.unibas.ch
19-051-697

20.06.2022

Acknowledgments

So Long, and Thanks for All the Fish. And the template.

Abstract

Text-entry is one of the most common forms of computer-human interaction and indispensable for many tasks such as word processing and some approaches to multimedia retrieval. The conventional keyboards everybody knows have long-established as the main text input method for desktop and laptop computers and even for touchscreen based devices they are very useful. But when it comes to virtual reality (VR) and augmented reality (AR), conventional keyboards might not be the best solution. With today's technology, VR and AR lack of tactile feedback and accurate finger tracking. As a result, text input for VR and AR is still an area of active research.

In recent years, word-gesture keyboards (also called slide-to-type keyboards) have been introduced in most major smartphone operating systems. Word-gesture keyboards look more or less like a conventional keyboard. But instead of tapping on the different keys, words are written with gestures. Now, the idea is, that they might also work well with VR/AR.

Table of Contents

1

Evaluation

In this section we want to talk about the evaluation as a whole. We want to look at the phrases we took, how we carried it out, the results observed and shortly discuss what all of this means.

1.1 MacKenzie Phrase Set

One precondition for the evaluation is to use the MacKenzie Phrase Set[1]. Basically, this is just a set of 500 phrases. According to the paper [?], such a phrase set should use phrases of moderate length, that are easy to remember and representative for the target language. These phrases do not contain any punctuation. Some of them use uppercase characters, but the authors mention, that participants can also be instructed to ignore the case of the characters.

Some statistics for the whole phrase set, also found in the original paper [?]: The MacKenzie phrase set consists of 500 phrases, that have a minimum length of 16, a maximum length of 43 and an average length of 28.61 characters. On the whole, 2712 words were used, which consist of 1163 unique words. A phrase consists of a minimum of 1, a maximum of 13 and on average of 4.46 words.

1.2 Task of the Participants

The task of the participants is to copy 15 ”random” phrases. They are not really random, but adjacent phrases from the downloadable MacKenzie Phrase Set (http://www.yorku.ca/mack/PhraseSets.ziphttp://www.yorku.ca/mack/PhraseSets.zip). As they are not in a specific order, e.g. alphabetic order, we decide to do it like this.

TODO: PICTURE OF EVALUATION SCENE. The participants see two text fields. On the top is the phrase to copy, on the bottom the words/phrase they write. If the given phrase matches the user inputted phrase, a sound sounds, such that the participants know when they finish one specific phrase. After that, a new phrase appears until 15 phrases

[1] http://www.yorku.ca/mack/PhraseSets.zip

are correctly inputted. If an incorrect word is entered, the user either can use the word suggestions (fig **??**) or delete the wrong word and try to write it again. If a mistake is only noticed later on, the participants have to remove all words and characters up to and including the wrong word by using the backspace button.

After this first step, in the second step, we shortly explain two functions of the keyboard, which they can test afterwards. First the scaling buttons and then the function to add a new word. This is important, because we want to know if they find these functions useful and well implemented.

The last step of the evaluation is to fill a questionnaire. First it has some general questions about the participant's experience in VR. Then there is a block of questions in the form of a system usability scale. Per question, there are five possibilities to set the cross. From 1 (strongly disagree) to 5 (strongly agree). The questions are structured in such a way, that if the user is highly satisfied with everything, they would alternately make a cross at the 5 and 1. TODO: FRAGEBOGEN ALS ANHANG BEIFÜGEN.

1.3 Carry-out

To carry out the evaluation, we used two different VR systems. One was a setup with a HTC vive and HTC vive controllers. The other one included an Oculus Rift headset with corresponding controllers. Even though these are two different systems, it does not change much for the participants. In fact, only the controllers and their buttons differ a bit.

To find participants, we wrote an email to students from our university, and asked family members and friends. All in all, eleven people got in touch with us and participated at our evaluation. Every participant got the same explanation to give everybody the same foundation of knowledge.

We told them that if they are close enough to the keyboard, then the color gets a bit brighter, and they are in the keyboard's hitbox. We said, the keyboard is movable, if they press and hold the controller's grip button in the hitbox of the keyboard. If they release it, the keyboard gets static again and stays where it got put.

To write, they do also have to be in the hitbox of the keyboard but not pressing and holding the grip button, but the trigger button. Then they had to make a gesture over the characters of the keyboard to write a word. We also told them, that if they do a full gesture and a word longer than one character is written, a space is automatically put behind the word. We also said to them, that single characters could be inputted by clicking on a key of the intended character. If they did so, no space is put, and they have to do it their own. In the English language, this is particularly important for the words ”I” and ”a”.

We also told them, that if they made a gesture and a word was written, there may be one to four other choosable words. They could pick from them, if the word written in the text field is the wrong one. We especially mentioned the word “the”. All the time “thee” would be written as the best match, therefore they would have to correct it every time.

They were also informed about the backspace. So, that if they use the backspace button after writing a word, the whole word gets deleted and afterwards only single characters get deleted.

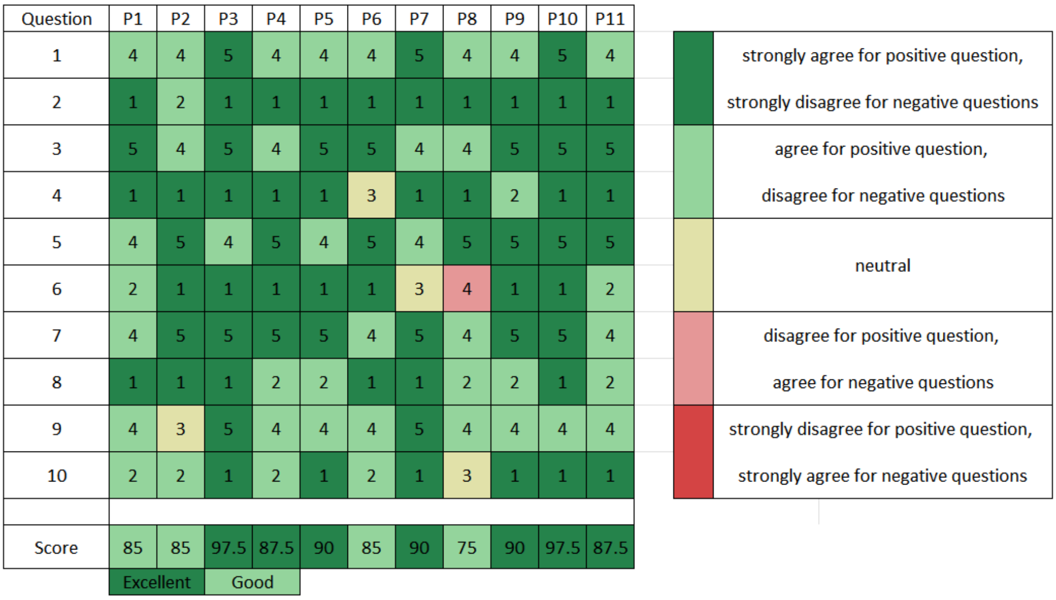
We also told the participants, that we have enough time and that they should not hurry, but rather look, that the inputted words are correct. Because if they are not correct, they have to use the backspace a lot of times.

1.4 Results

Now, we want to talk about the results and some statistics we gained through the evaluation.

1.4.1 System Usability Scale

First, we begin with the results of the SUS questions. These ten questions were:
1. I think that I would like to use this system frequently when I work in VR
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I need the support of a technical person to be able to use this system
5. I found the various functions in this system were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people would learn to use this system very quickly
8. I found the system very cumbersome to use
9. I felt very confident using the system
10. I needed to learn a lot of things before I could get going with this system



Question	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
1	4	4	5	4	4	4	5	4	4	5	4
2	1	2	1	1	1	1	1	1	1	1	1
3	5	4	5	4	5	5	4	4	5	5	5
4	1	1	1	1	1	3	1	1	2	1	1
5	4	5	4	5	4	5	4	5	5	5	5
6	2	1	1	1	1	1	3	4	1	1	2
7	4	5	5	5	5	4	5	4	5	5	4
8	1	1	1	2	2	1	1	2	2	1	2
9	4	3	5	4	4	4	5	4	4	4	4
10	2	2	1	2	1	2	1	3	1	1	1
Score	85	85	97.5	87.5	90	85	90	75	90	97.5	87.5
	Excellent		Good								

🟩	strongly agree for positive question, strongly disagree for negative questions
🟢	agree for positive question, disagree for negative questions
🟨	neutral
🟧	disagree for positive question, agree for negative questions
🟥	strongly disagree for positive question, strongly agree for negative questions

Figure 1.1: System Usability Scale with all the given points per question from every participant

Question 9 got the worst score, but with it being a 4.09 out of 5, it is still pretty good. From these values, we can calculate a usability score. Every question with an even number

is a negative one. That means, a score of one or “strongly disagree” is the highest possible. For the other questions, a score of five or “strongly agree” is the best possible score. So, from the odd numbered questions we have to subtract 1 from the average score. And for the even numbered questions we have to subtract their average score from 5. At the end, we have to sum up these ten newly calculated values and multiply them by 2.5. Our calculated usability score is 88.18. This is a high score, because from a score of 85.5 points, one talks about an excellent system usability. Therefore, we are really satisfied with the results of this.
We do also have two other questions about the scale and the “add word” function:
11. The function to add words is well implemented and easy to use
12. The function to scale the keyboard is unnecessary
Question 11 got a score of 4.55 out of 5 and question 12 got a score of 2.18, whereby 1 would be ideal. We conclude from these two questions, that the “add word” function makes a good impression whereas the scale function does not perform so well.

1.4.2 Writing Speed

One important thing of our evaluation is to find out, how fast users can write with our word-gesture keyboard. As unit to measure this values, we take the “words per minute” wpm. We calculate the wpm with following formula:

$$WPM = \frac{\mid T \mid}{S} \times 60 \times \frac{1}{5} \tag{1.1}$$

where T is all the phrases a participant had to write, hence $\mid T \mid$ is the number of characters a participant had to write. S is the time in seconds they used to write all 15 phrases.

participant	average WPM	lowest WPM	highest WPM
1	11.457	-	-
2	12.19	-	-
3	13.055	-	-
4	11.609	5.3	25.5
5	12.578	6.83	21.65
6	10.285	5.27	19
7	12.423	6.1	24.41
8	16.056	8.28	30.74
9	13.363	7.96	24.15
10	17.118	7.98	24.45
11	10.067	4.71	14.55
average	12.75	6.55	23.06

Table 1.1: average wpm, lowest wpm and highest wpm per participant. For the first three, we failed to get this data.

In Table **??** you can see how fast in average the participants were able to write their 15 phrases. We do also list the lowest and highest value. Everything is measured in words per minute.

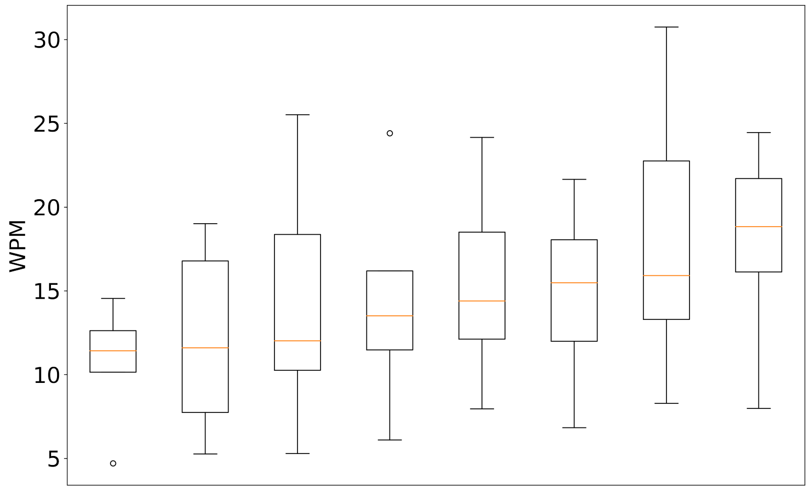


Figure 1.2: for participants 4-11: lowest wpm, 25% quantile, average wpm, 75% quantile and highest wpm

To understand the values of Table **??** a bit better, we make a so-called boxplot for every participant, for whom we have the data. We can see in fig **??**, the higher the participant's median, most of the time they do also have higher lowest wpm value. The lowest wpm values mostly come about because a participant made a mistake and had to delete a lot and basically write the phrase two times. On the other hand, the highest WPM values come about because a participant made no mistake in writing the phrase. The rectangle in the middle of the two bars shows how consistent or inconsistent a participant's writing speed is. The lower bound is the 25% quantile, the upper bound the 75% quantile. This means, if the rectangle is less high, the writing speed is more consistent.
We cannot really find anything that combines writing speed and consistency.

Next, we want to look at the wpm values of the users and their experience in, on one hand VR writing and on the other hand word-gesture keyboards.

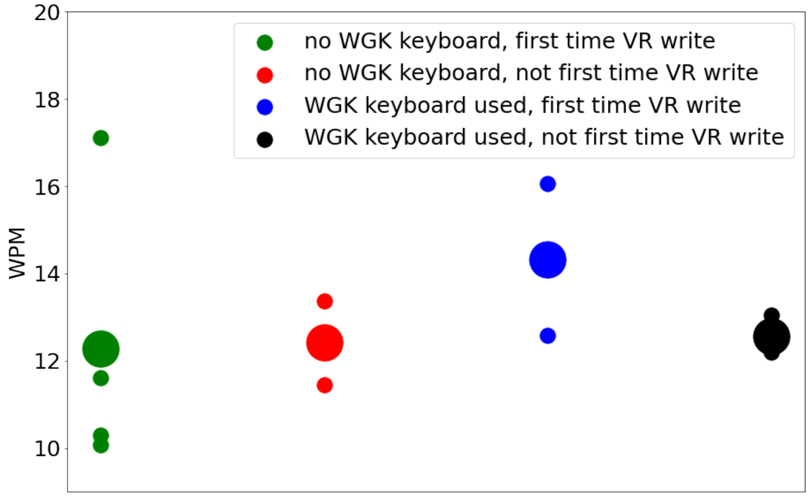


Figure 1.3: Average WPM per group of participant, grouped by their experience in VR writing and word-gesture keyboards

In Fig **??** we can see, that the prior knowledge, that some participants have, did not really help them to write faster. In fact, the fastest group was the one, that is experienced with word-gesture keyboards, but not with writing in VR. But we think this is due to the small sizes of the different groups.

Another thing we wanted to analyze is the WPM value compared to the age:

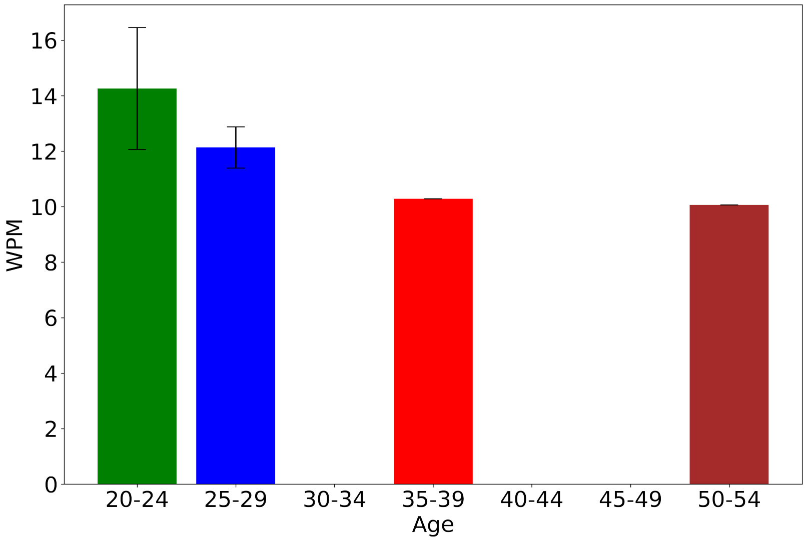


Figure 1.4: wpm values sorted by age groups of five years. The 20-24 and 25-29 groups also have a standard deviation error bar, because they contain of five and four participants. The 35-39 and 50-54 groups do not have an error bar, since each of them only contains of one participant

We can observe in Fig **??**, that at least among our participants, the wpm value decreases with the increase in age. This means, the older participants were a bit slower than the younger ones.

1.4.3 Error Rate

For the measurement of the error rate, we calculate a user error rate and a system error rate. For the user error rate, we look at the amount of backspaces a participant had to use and the amount of words/character that were written wrong in all 15 phrases. For the system error rate we look at the words the system did not recognize at all, which means not as best match, nor as one of the four choosable suggestions and count their number of characters.

1.4.3.1 Most Frequent Error Words

We looked into all the words that were not the best match, so all the words a participant either corrected or not and also the ones that were not even in the suggestions. The top ten such words are:

Table 1.2: most frequent error words

word	times wrong
the	53
is	17
to	14
of	12
in	9
for	8
do	5
all	3
more	3
see	2

As we can see in Table **??**, the word, that is responsible for most of the errors, is “the”. It caused 53 errors. We can also see in this list other words, that could have been avoided by a better implementation. For example “ìn” and “more”. The best match for these words were “thee”, “inn” and “moore”, which are words, that do not appear that much in plain language. But we want to go more in depth on this topic later on.

1.4.3.2 User and Backspace Error Rate

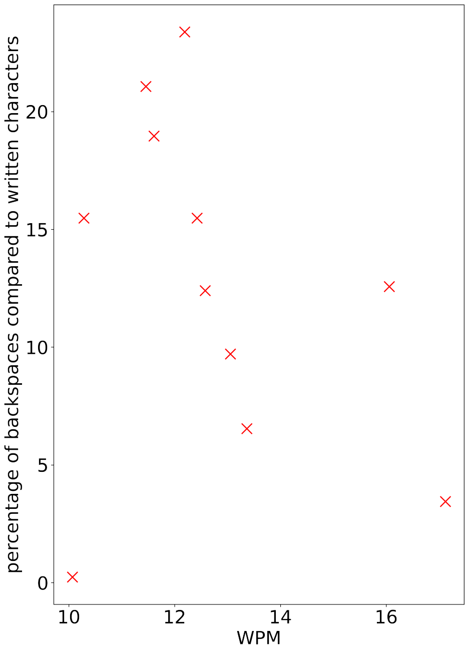
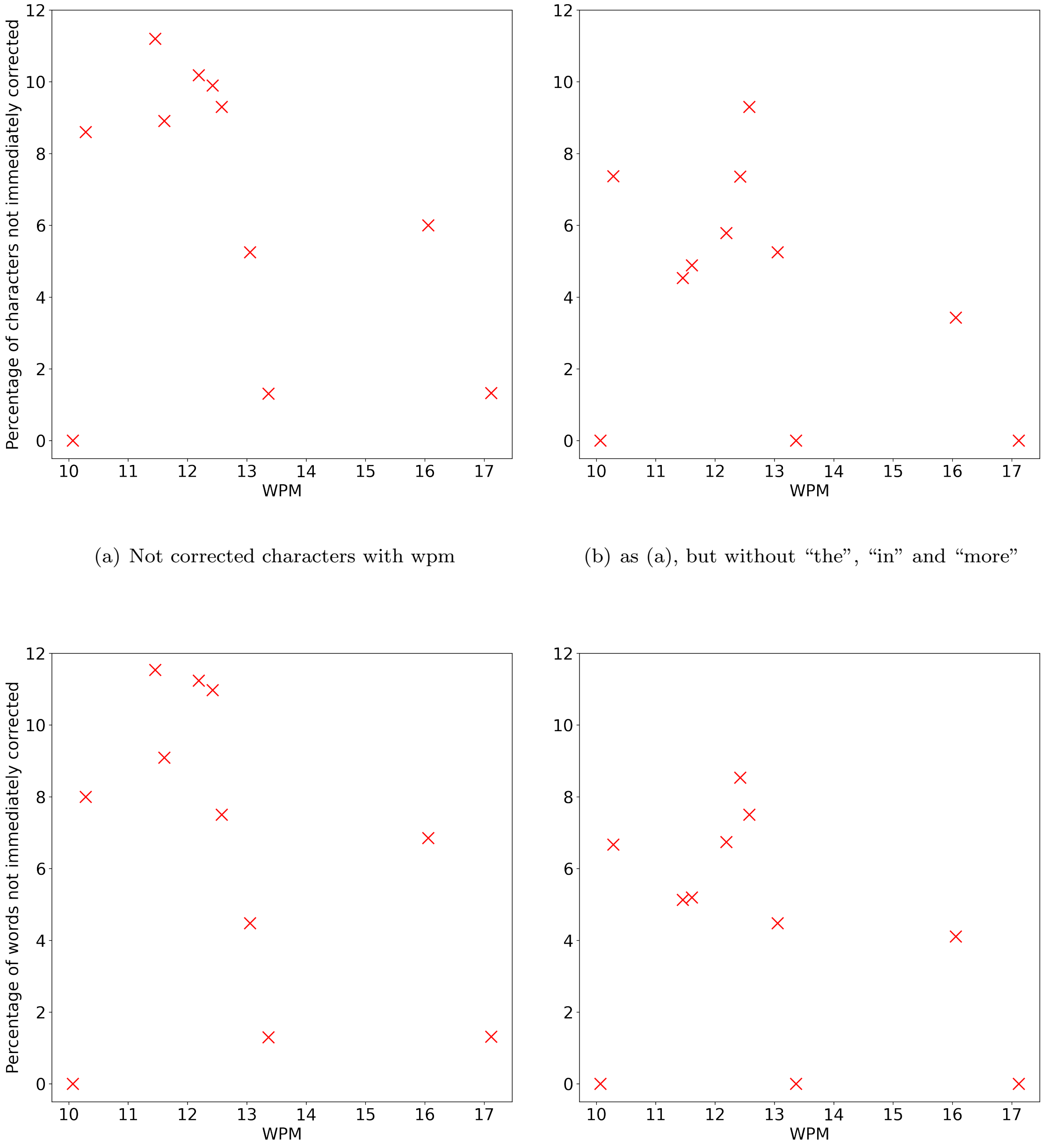


Figure 1.5: percentage of backspaces in relation to all characters in phrases compared to the wpm

In Fig **??** we show how many backspaces were used at which wpm value. We can not tell really much about it. But one thing that can be seen is, that the participant with the lowest wpm seemed to be very careful not overlooking wrong words that could be corrected by choosing the right suggestion. The participant with the highest wpm has the second-lowest usage of backspaces. Therefore, they seem to get used to the system and its suggestions words very fast and good.
Overall, there seems to be a little trend, that participants with higher wpm values had to use fewer backspaces than participants with lower wpm values.



(a) Not corrected characters with wpm

(b) as (a), but without “the”, “in” and “more”

(c) Not corrected words with wpm

(d) as (c), but without “the”, “in” and “more”

Figure 1.6: Plots of four Turing machines

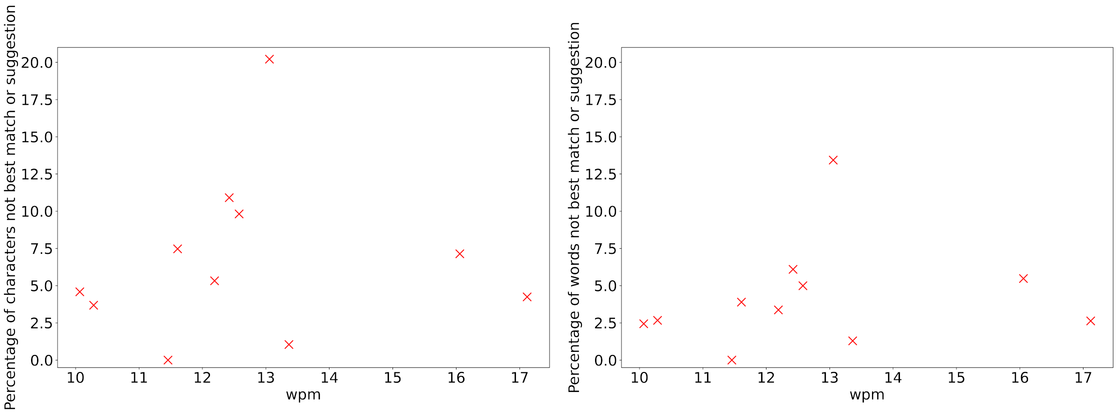
In Fig **??** and Fig **??** we count the characters of the words, a participant did not correct, although they had the chance to do it with the suggestions. On the y-axis, we look at the percentage, that these characters make up in comparison to all characters the participant had to write in all 15 phrases. In Fig **??** we look at all characters. But a lot of the errors came from the words “the”, “in” and “more”, that could have been prevented as mentioned in Section TODO: MENTION TO SECTION THAT HAS TO BE WRITTEN. Therefore, in **??** we look at the errors without consider these three words.

In the figures **??** and **??** we do the same, but without counting the characters of the "error words" but just look at the number of them.

One thing that can be observed is, that in the two lower figures of Fig **??** the percentage values are almost the same as in the two upper ones. This means that the words that were not corrected, have on average about the same length as the average length of all words from the respective 15 phrases is.

One thing that can be observed at the left figures of Fig **??** is that apart from one exception the percentage rate of not corrected words/characters is higher at lower wpm than it is at higher wpm. One obvious reason for this could be, that if words early in phrases were not corrected, a user had to delete the whole phrase back to the wrong word, and had to write all again. Therefore, it would make sense, that the faster participants made fewer errors, hence their risk of not correcting a word in the beginning of a phrase is lower. And then, they would not decrease their average wpm so much.

1.4.3.3 System Error Rate



(a) with characters

(b) with words

Figure 1.7: percentage of characters/words that were not found by the system neither as best match nor as suggestion

In **??** we can see what percentage of the characters/words were not found by the system, neither directly as best match, nor as word suggestions, which a participant could choose. We decide to name this error the system error, although a participant could have done a really bad gesture which then would not really be the system's fault to not find a word.

We can see, that in **??** the percentages are much higher than in **??**. This means that the words, which the system did not find a word, need to be longer than an average word in the written phrases. A reason for this might be, that often gestures for long words did not succeed. If a participant made some little curves too much in a gesture for a long word, the chances are lower, that the system can detect the right word, whereas for a small word the

chance is higher, that the word gets found. Because the gesture can be much shorter and therefore fewer inaccuracies from a participant will happen.
For the wpm and the percentages we can not detect any correlation. It seems, that these two values are not connected with each other.

1.4.4 Feedback

The full list including all verbatim feedback from every participant can be found in the appendix. Here, we just want to highlight the most frequently addressed points.
Some participants find the best matches and suggestions sometimes confusing. The most mentioned example is that "thee" is preferred over "the". The spaces/ current position are another thing that is frequently addressed. For some participants, it was unclear where they were to write. They suggest to implement some kind of visual indication, that indicates, where the cursor currently is. We also got a lot of praise and most of the time we saw a cheerful face when the VR headset got taken off.

1.5 Discussion

In this section we want to compare our results to other works' results and talk about some changes we did because of the feedback and results.

1.5.0.1 Total ER

As last error, we want to calculate the Total ER.

$$Total\ ER = \frac{INF + IF}{C + INF + IF} \times 100\% \tag{1.2}$$

C is the number of correct keystrokes. Here we take the number of characters the 15 phrases a participnt had to input had. INF denotes the number of incorrect and not fixed characters in the transcribed text. Because we only allowed to go to the next phrase, if the last one was fully correct, we take a value of 0. IF denotes incorrect but fixed characters. For this we take the number of characters of the words a user corrected with the suggestion words, because the system did not recognize the right word as best match. The following table shows our calculated results:
TODO: TALK ABOUT TOTAL ER

1.5.1 Result Comparison

First, we begin with the wpm. Boletsis and Kongsvik [?] evaluate in their paper four different VR input methods, a raycasting, drum-like, head-directed input and a split keyboard. The first one is a keyboard where a user can select letter by pointing a ray with a controller on it. For the second one the controllers simulate drum sticks in VR and letters have to be pressed by them. For the third keyboard a user has to aim with the head for the letters and press a button on the controller to input. The last keyboard is one, that is split into two halves, one assigned to each controller.

Table 1.3: Total ER per participant

participant	total ER
σ	
1	value
value	
2	value
value	
3	value
value	
4	value
value	
5	value
value	
6	value
value	
7	value
value	
8	value
value	
9	value
value	
10	value
value	
11	value
value	

Table 1.4: most frequent error words

text input method	wpm
Raycasting keyboard	16.65
Drum-like keyboard	21.01
Head-directed input keyboard	10.83
Split keyboard	10.17
Word-gesture keyboard	12.75

In Table **??** we listed the results of Boletsis and Kongsvik [**?**] and our measurement of the wpm value. They had a similar approach to the evaluation as we did, with one difference. They did use the same ten phrases for every participant and keyboard type.
As we can see, our keyboard lines up in the middle. It is not the one with the lowest wpm, but also not the one with the highest.
Another comparison we can do with the results of Boletsis and Kongsvik [**?**] is the total ER value. They got a value of TODO: INSERT THEIR TOTAL ER VALUE. Our calculated value is TODO: INSERT OUR TOTAL ER VALUE. TODO: COMPARE RESULTS.

1.5.2 New Implementations

2

Conclusion

This is the body of the thesis.

2.1 Results Discussion

2.2 Future Work

A

Appendix

University of Basel

Faculty of Science

Declaration on Scientific Integrity

(including a Declaration on Plagiarism and Fraud)
Translation from German original

Title of Thesis:

Name Assesor: ____________________

Name Student: ____________________

Matriculation No.: ____________________

With my signature I declare that this submission is my own work and that I have fully acknowledged the assistance received in completing this work and that it contains no material that has not been formally acknowledged. I have mentioned all source materials used and have cited these in accordance with recognised scientific rules.

Place, Date: ____________________ Student: ____________________

Will this work be published?

□ No

□ Yes. With my signature I confirm that I agree to a publication of the work (print/digital) in the library, on the research database of the University of Basel and/or on the document server of the department. Likewise, I agree to the bibliographic reference in the catalog SLSP (Swiss Library Service Platform). (cross out as applicable)

Publication as of: ____________________

Place, Date: ____________________ Student: ____________________

Place, Date: ____________________ Assessor: ____________________

Please enclose a completed and signed copy of this declaration in your Bachelor’s or Master’s thesis .

August 2021