Linguistic Difference of Human-Human and Human-Chatbot Dialogues about COVID-19 in the Russian Language

Aleksandr Perevalov[1], Aleksandr Vysokov[2] and Andreas Both[1]
[1]*Anhalt University of Applied Sciences, 55 Bernburger Str., Köthen, Germany*
[2]*Perm National Research Polytechnic University, 29 Komsomolsky avenue, Perm, Russia*
{*aleksandr.perevalov, andreas.both*}*@hs-anhalt.de, sshvsk@mail.ru*

Keywords: Quantitative Text Analysis, Human-Chatbot Interaction, Human-Computer Interaction, Chatbot Interaction Design, Language Features, Wizard of Oz, COVID-19.

Abstract: This work describes the quantitative analysis of the linguistic difference in human-human and human-chatbot dialogues. The research is based on conducting a set of experiments where respondents communicate with a human or a chatbot in the domain of COVID-19 questions. In the case of the human-human dialogues, the approach of the inverted "Wizard of Oz" experimental setting is used. During the experiments, 35 human-human and 68 human-chatbot dialogues in Russian language were performed. The dialogues were collected during 4 months and thereafter analyzed with a set of quantitative text measures such as descriptive statistics of a text, syntactic complexity, lexical density, and readability. As a result, a set of measures demonstrated a statistically significant linguistic difference between the language structure of questions that were asked to the human and to the chatbot. Specifically, respondents were using shorter sentences and words, simpler syntax while communicating with the chatbot. Moreover, lexical richness of the human-chatbot dialogue data is lower while the readability is higher – these markers indicate that humans use simpler language constructions while speaking with a chatbot.

1 INTRODUCTION

The use of virtual assistants such as chatbots or question answering systems is a fairly relevant topic and develops every year more and more[1]. In today's highly competitive realities, chatbots bring real benefits to society [1, 2, 3]. A chatbot responds instantly in comparison to a human, making it less likely that the user will leave without getting a response and also simplifies real-world problems (e.g., customer service, corporate information search) by doing a lot of routine work.

Researchers and developers strive to bring their chatbot products to such an extent that communication with them could not be distinguished from communication with a real human in the context of fulfilling information needs. However, chatbots may not always sufficiently cover all the data and knowledge which is necessary for successful communication. The process of interaction between a human and a chatbot also has social and psychological aspects, as a result of which a person can somehow change his behavior and adjust his way of communication to a chatbot [4].

This work analyzes the linguistic difference between human-human and human-chatbot dialogues by using the method of quantitative text analysis [5]. To model a real-world scenario, the emergent knowledge domain of COVID-19 pandemic was selected[2]. Thereafter, a chatbot capable of answering frequently asked questions (FAQ) based on public data provided by government was developed. The working language of the chatbot and therefore all the collected data is the Russian language. During the experiment, the respondents were randomly distributed to one of the two groups: Group_1 – respondents communicate with a human-expert, Group_2 – respondents communicate with the developed FAQ chatbot. After that, the corresponding dialogues were collected in the textual form and analyzed using different measures, such as syntactic complexity, readability, lexical diversity and other descriptive statistics (see Section

[1] https://www.gartner.com/smarterwithgartner/chatbots-will-appeal-to-modern-workers

[2] It is worth underlining, that the subject of this study is not to analyze what questions were asked, but how they were asked.

2.2 for details). The experimental results highlight significant difference between the human-human and human-chatbot dialogues w.r.t. the considered measures. Based on the results, the practical recommendations on the chatbots' development and design process were formulated.

In this work, the following research questions are answered: RQ_1 – Are human-human and human-chatbot dialogues different from each other in terms of quantitative textual measures, RQ_2 – If the first is true, what measures distinguish human-human and human-chatbot conversations, and RQ_3 – What recommendations to the chatbot developers can be created in this regard? The scientific novelty of the work consists of the following: a dialogue dataset based on the topic of consulting people on COVID-19 FAQ questions was collected[3]. The dataset consists of human-human and human-chatbot dialogues, (2) the collected data was analyzed using several quantitative text measures, and (3) the corresponding practical recommendations for the developers were formulated. The practical value of the work is that it can help researchers and developers understand how to make the human-chatbot interaction process more natural and useful for both the user and the developer.

This paper is organized as follows. In Section 2 the related work connected with this research is presented. In Section 3 the analysis approach is described. Section 4 presents experimental setup. In Section 5 the analysis and discussion of the results is described. Section 6 summarizes the research and explains the limitations and future plans of the work.

2 RELATED WORK

2.1 Human-Chatbot Interaction from Linguistic Perspective

It is well known that chatbots tend to limit free text from the user in order to ensure conversational structure [6]. Developers have to deal with this limitation by hiding it from a user such that one doesn't see explicitly the borders of dialogue. Thus, a chatbot triggers the fact that the language used by a human while interacting with a chatbot differs from a conversation with a human. There are studies showing that a human tends to imitate the vocabulary of a chatbot [7] and to match its language style [8]. The role of used language constructions is not limited to the utterances used by chatbot. If such a system uses machine learning or corpus based methods, its performance is also biased to the available datasets [9]. Hence, it may also negatively influence the quality of machine learning models used in a chatbot (e.g., intent classification) as human-human and human-chatbot language constructions are different. To the best knowledge of the authors, there are only few recent studies that compared human-human and human-chatbot from linguistic perspective [8, 10, 11]. The main research question of these studies is similar to this work: "Do humans communicate differently when they know their conversational partner is a computer as opposed to another human being?". However, this work is different from the mentioned study by a knowledge domain of a chatbot and the language of experiments. *Authors of this is work are not aware of any publications studying differences in human-human and human-chatbot dialogues from the linguistic perspective in Russian language.*

2.2 Measures of Quantitative Text Analysis

The syntactical complexity represents how complex are the sentence structures used in the text. This characteristic is represented by the *Mean Dependency Distance* (MDD) [12]. The MDD is calculated as shown in (1).

$$MDD = \frac{1}{n-s}\sum_{i=1}^{n}|DD_i| \qquad (1)$$

Where n is the total number of words in a document, s is the number of sentences in a document, DD_i is the dependency distance of the i-th syntactic link of the document[4] [13].

The lexical diversity represents how complex and dense is the vocabulary used in the text. This characteristic is represented by the following measures. *Lexical Density* (LD) is calculated according to the Ure's method [14] (see (2)).

$$LD = \frac{N_{lexicalitems}}{N_{words}} * 100 \qquad (2)$$

Where N_x corresponds to number of a corresponding variable in the formula.

Another well-known measure is *Type-Token Ratio* (TTR) [15]. The term "type" corresponds to the number of unique words in a text corpus. The TTR is calculated according to (3).

$$TTR = \frac{V}{N} \qquad (3)$$

[3] Authors will publish the dataset online after the paper acceptance decision.

[4] The connection between words or group of words in a string.

Table 1: Language-specific coefficient values for FRE and FKG metrics.

Language	k_1	k_2
Flesch Reading Ease (FRE)		
English [18]	1.015	84.6
Russian [19]	1.3	60.1
Flesch-Kincaid Grade (FKG)		
English [20]	0.39	11.8
Russian [19]	0.5	8.4

Where V is the number of types and N – number of tokens.

There are also several TTR-related measures such as *Herdan's C* or *LogTTR* [15] (see (4)), *Summer's Index* [16] (see (5)), and *RootTTR* [17] (see (6)).

$$LogTTR = \frac{log(V)}{log(N)} \tag{4}$$

$$S = \frac{log(log(V))}{log(log(N))} \tag{5}$$

$$RootTTR = \frac{V}{\sqrt{N}} \tag{6}$$

The readability measures demonstrate how hard is to read the text. There are several such measures supported in the presented software. The well-known *Flesch Reading Ease* [18] (FRE) depends on the syllables per word ($\frac{n_{sy}}{n_w}$) and words per sentence (*ASL*), see (7).

$$FRE_{lang} = 206.835 - k_1^{lang} * ASL - k_2^{lang} * \frac{n_{sy}}{n_w} \tag{7}$$

The coefficients (k_1^{lang}, k_2^{lang}) mentioned in the Formula are language-specific. The corresponding values are demonstrated in Table 1.

The value of Flesch-Kincaid Grade (FKG) correspond to a U.S. grade level that is required to read a given text [20]. The (8) is dependent on the language-specific coefficients as FRE (see Table 1).

$$FKG_{lang} = k_1^{lang} * ASL + k_2^{lang} * \frac{n_{sy}}{n_w} - 15.59 \tag{8}$$

The *Automated Readability Index* (ARI) [21] and its simplified version – sARI are also supported by the software (see (9) and (10) respectively).

$$ARI = 0.5 * ASL + 4.71 * AWL - 21.34 \tag{9}$$

$$sARI = ASL + 9 * AWL \tag{10}$$

Where AWL corresponds to the average word length. The *Coleman's Readability* [22] that includes number of one-syllable words is shown in (11) (where $n_{wsy=x}$ corresponds to the number of words with x syllables).

$$Coleman's = 1.29 * \frac{100 * n_{wsy=1}}{n_w} - 38.45 \tag{11}$$

The *Easy Listening Score* (ELS) [23] is the ratio between the number of words with 2 syllables or more and number of sentences (see (12)).

$$ELS = \frac{n_{wsy>=2}}{n_{st}} \tag{12}$$

3 APPROACH

In this section, the approach for collecting the dialogue data and therefore its comparison is described. Firstly, the chatbot that is able to answer frequently asked questions has to be developed. The underlying knowledge base D for the chatbot must be taken from the trustworthy and validated sources, and structured as pair $d_i = (Q_i, a_i)$, $d_i \in D$, where Q_i – is the list of possible questions targeting on a similar information need (e.g., "Will a mask protect me from the virus?" and "How helpful are the masks for COVID?"), a_i – is the answer text from a validated data source that is fulfilling the information need of Q_i. The algorithm of the FAQ chatbot works over the data D and selects the most relevant a_i for a given Q_i. Such an algorithm is just represented by a simple multi-class classifier as the number of unique a_i is much lower than a number of unique Q_i. This algorithm for FAQ answering of the chatbot was selected due to the lower quality requirements and implementation simplicity in comparison to the data-driven chatbots (e.g., [24]). The process of the FAQ chatbot development is presented in Section 4 in detail.

Secondly, the respondents has to be collected and assigned randomly to the two different groups: Group$_1$ – respondents communicate with a human-expert, Group$_2$ – respondents communicate with the developed FAQ chatbot. In case of Group$_1$, *a respondent knows that a human-expert is on the other side of the dialogue*. In its turn, the human-expert forwards a question to the chatbot and returns an answer to a user produced by the FAQ chatbot, s.t., it is hidden to a respondent (i.e., this is an inverted "Wizard of Oz" experiment following [25, 26]). The *human-expert does not change neither question nor answer* and is required in the experiment to create a trustworthy UI outlook such that a respondent is sure that a dialogue is conducted with a human. In case of Group$_2$, a user is provided with a link to the FAQ chatbot and prompted to have conversation with the chatbot. These two groups of respondents are independent and are in the same conditions, both utilize the same user interface (UI) while performing the dialogues with either a human-expert or a FAQ chatbot. The only difference is that they know who they are speaking to.

Finally, the collected dialogue data has to be anonymized and therefore analyzed using quantitative text analysis measures. For the analysis all the measures introduced in Section 2.2 were used, in addition such descriptive statistics of text as average- words per sentence, sentence length, word length and syllable per word were computed. The general schema of the approach is demonstrated in Figure 1.

4 EXPERIMENT

The knowledge base was collected from the FAQ section of the official portal of the Russian government "Stop Coronavirus"[5]. Thereafter, it was structured using a parser script as follows $d_i = (Q_i, a_i),\ d_i \in D$ (see Section 3).

To develop the chatbot, the Google Dialogflow[6] framework was used. Each Q_i was considered as "intent" in the framework. Therefore, the knowledge base D was loaded into the Google Dialogflow platform to train the intent recognition model. After a particular intent i is recognized, the chatbot returns a_i as a response.

The chatbot was deployed as a Bot in the Telegram Messenger[7]. Consequently, it was accessible from any kind of device (e.g., mobile, desktop, tablet etc.). This option enables to use the same UI for both respondent groups and exclude the UI from the possible threats of validity of this study.

The respondents were attracted to the study via social media announcements. There were 103 respondents involved in our experiment in total. 35 respondents were part of the Group$_1$. The other 68 respondents were part of the Group$_2$. The first dialogue was conducted on March 16, 2021 and the last dialogue was conducted on July 29, 2021. Hence, the dialogue data collection process continued for more than 4 months. All the dialogues were carried out in Russian language. The following features were collected for each message: chat id, user id, data, first name and last name (anonymized), question, answer. The example of a collected dialogue translated to English is demonstrated in Table 2. Thereafter, a set of quantitative text measures was calculated on both datasets from Group$_1$ and Group$_2$ using the LinguaF Python package[8].

5 RESULTS AND DISCUSSION

The measures on which the calculation and comparison will be made are presented in Table 3. The columns "p-value" and "Is significant" correspond to the result of two-sample unequal variance two-tailed T-test with significance level (α) 0.01. If the result of this statistical test is "Yes", then the difference is significant.

5.1 Descriptive Statistics and Syntactical Complexity

It is evident from the Table 3 that in case of human-human dialogues, respondents use more words and therefore sentences are longer (see metrics 1, 2). For instance, in a human-human, the respondent uses 1.8 times more words than in a human-chatbot dialogue. There were also significant differences seen on measures 3 and 4. Hence, *respondents use longer or more complex words*. Since the average distance between dependent words in human-human dialogues (2.210) is greater than in human-chatbot dialogues (1.651), the Syntactic complexity will also be higher (see Metric 5).

5.2 Lexical Diversity

Lexical density and Type-Token Ratio (cf., Section 2.2) values demonstrate that the differences between human-human and human-chatbot dialogues are insignificant. However, considering such measures as Log Type Token Ratio, Root Type Token Ratio and Summer Index, it is obvious that Lexical diversity in human-human dialogues is much higher than in human-chatbot ones. Thus, in the dialogue with the chatbot, the respondents choose simpler phrases. Consequently, *when developing such a chatbot at the stage of preparing the training data, it is worth considering the fact that dialogues with such systems are not always similar to human ones*, and, if necessary, make adjustments to the data (e.g., simplification of the questions).

5.3 Readability

The metric Flesch Reading Ease also signals the use of more complex language constructions in "human" dialogues. According to the obtained values of the Flesch Reading Ease metric, in the case of a dialogue with a human, the respondents operate close to the language level of a university student, while *in a dialogue with a chatbot, the level of language constructions is two steps lower, which corresponds to the 7th*

[5] https://stopkoronavirus.rf/faq

[6] https://cloud.google.com/dialogflow

[7] https://core.telegram.org/

[8] https://github.com/Perevalov/LinguaF

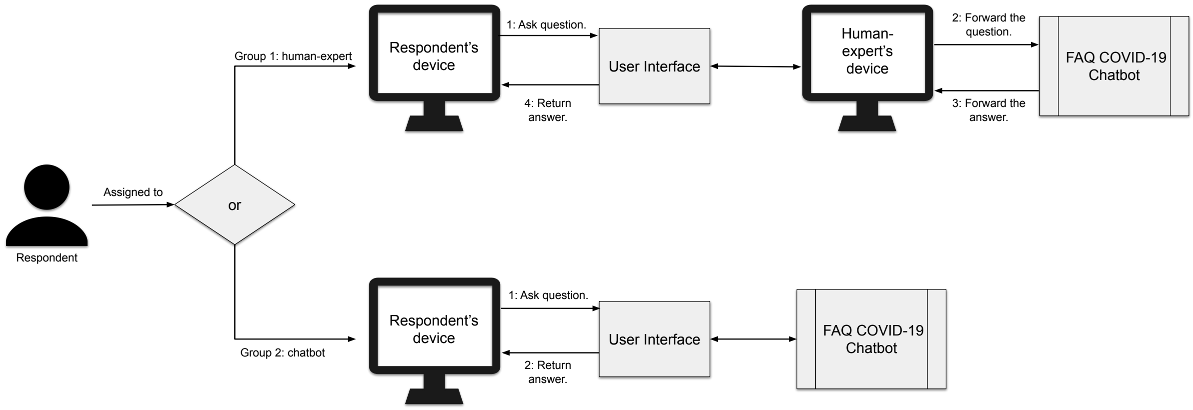

Figure 1: The general approach schema. The human-expert is just "playing" the role of an expert, while actually forwarding all the questions and answers to/from the FAQ chatbot (inverted "Wizard of Oz"). In this setting, the respondent thinks that the communication partner is a real human-expert, however, the human-expert is fully controlled by the chatbot.

Table 2: The examples of collected questions and their intentions. For each intention, an answer in the knowledge base is defined.

№	Question	Intention	Group
1	What vaccines there are?	Information on vaccines	2
2	Can I put Pfizer in a private clinic, for example?	Information on vaccines	2
3	What should I do if I see signs of Coronovirus?	Symptoms	1
4	What antibodies can be detected and what does this tell me?	Antibodies	1

Table 3: Results of the measures' calculation. The column "Avg. Human Data" contains the values of the human-human dialogues (i.e., Group_1). The column "Avg. Chatbot Data" contains the values of the human-chatbot dialogues (i.e., Group_2). The values of the "Is significant" column are calculated according to the "p-value" results and the significance level $\alpha = 0.01$.

№	Measure Name	Avg. Human Data	Avg. Chatbot Data	p-value	Is significant
	Descriptive Statistics				
1	Avg Words Per Sentence	6.287	3.370	0.000	Yes
2	Avg Sentence Length	37.692	18.894	0.000	Yes
3	Avg Word Length	6.144	5.812	0.007	Yes
4	Avg Syllable Per Word	2.348	2.160	0.000	Yes
	Syntactical Complexity				
5	Mean Dependency Distance	2.210	1.651	0.000	Yes
	Lexical Diversity				
6	Lexical Density	69.902	67.946	0.237	No
7	Log Type Token Ratio	0.986	0.806	0.000	Yes
8	Root Type Token Ratio	2.509	1.774	0.000	Yes
9	Type Token Ratio	0.997	0.998	0.189	No
10	Summer's Index	0.986	0.806	0.000	Yes
	Readability				
11	Automated Readability Index	10.650	7.627	0.000	Yes
12	Automated Readability Index Simple	61.580	55.674	0.000	Yes
13	Coleman Readability	41.439	51.837	0.000	Yes
14	Easy Listening	1.998	1.252	0.000	Yes
15	Flesch-Kincaid Grade	7.279	4.239	0.000	Yes
16	Flesch Reading Ease	57.532	72.639	0.000	Yes

grade student. The Flesch-Kincaid Grade indicates the level of readability as 7th grade in human-human dialogues and 4th grade in human-chatbot dialogues. While *the absolute values may be debatable, the relation between them shows that the language used in the human-chatbot dialogues is simpler*. Also, the Automated Readability Index and Simplified Automated Readability Index metrics show that in human-human dialogues, the perception of text is more complex than in human-chatbot dialogues. Based on the calculation of all the measures in this section, the complexity of readability in human-human dialogues appears to be several levels higher than in human-chatbot dialogues. It is worth noting that the "complexity" may be given by the subject of the dialogues, as in this case numerous medical terms are used.

5.4 General Statistics and Qualitative Analysis on Dialogues

The average number of messages in human-human dialogues is 9.4. The average number of messages in human-chatbot dialogues is 5.3. The total average number of messages in all dialogues is 7.35. The duration of the dialogues also varied. In the case of human-human, the dialogues ranged up to 30 minutes. In the case of human-chatbot dialogues, they are no more than 15 minutes long. The human-human dialogues also appear to be longer w.r.t. the number of questions that were asked on average (9.4 vs 5.3). This difference may be due to the social and psychological aspects of human-chatbot interaction.

The qualitative analysis of the dialogue data showed that in human-human dialogues, in most cases, respondents used more complex sentence structures and gave more clarifying information to get an answer to similar questions than in human-chatbot dialogues. *In the case of human-chatbot dialogues, respondents often ask personal questions*, such as: "How are you?", "What do you know?", "Who is your creator?", whereas *when talking to a human, only questions on the subject were used*. Moreover, in human-chatbot dialogues, the respondents frequently use obscene language.

5.5 Summary

In this section, two dialogues types – human-human and human-chatbot– were analyzed. The content of human-human dialogues is predominantly more complex and rich than in the case of human-chatbot dialogues (RQ_1). It is confirmed by the set of measures related to the Descriptive Statistics of text, Syntactic Complexity, Lexical Diversity, and Readability (RQ_2). The respondents, when communicating with chatbots, construct their speech intentionally or subconsciously in a simpler and clearer way. Therefore, chatbots and other dialogue systems should be designed so that they are prepared to work effectively with the simplified language input. To achieve this, it is proposed to use simplification techniques on the training data that is used to build such systems (RQ_3).

6 CONCLUSION

In this work, the differences between human-human ($Group_1$) and human-chatbot ($Group_2$) dialogues were analyzed. For this purpose, the respondents of the experiment were randomly assigned to one of the two aforementioned groups. The domain of the dialogues was focused on frequently asked questions about COVID-19 and the language of the dialogues was set to Russian. To conduct the experiment, the FAQ chatbot was created based on publicly available data, provided by the Russian government.

In total, 103 respondents were involved in the experiment. The scope of the analysis contained different quantitative text measures related to syntactical complexity, lexical diversity, readability and others. Given the experimental results, it was possible to identify significant difference w.r.t. the considered measures between human-human and human-chatbot dialogues. Specifically, respondents in the human-chatbot dialogues used shorter and simpler language constructions, which is reflected in the experimental results. Based on this, the authors of this work recommend researchers and developers of chatbots and dialogue systems to consider simplifying the training data used for the systems, as the majority of users are not asking well-formed questions. To summarize the discussion, the research questions of the study were fully answered.

However, several limitations may be identified in this work. Firstly, the selection process of the respondents might be biased due to the used social media connections of the authors of this work. In addition, only one UI was used in the experiments, hence, it may have added some bias into the dialogues as well. Finally, the work is limited by the used language of dialogues and knowledge domain.

For the future work, it is worth considering extending the analysis towards more languages, especially, low resource ones. In order to reduce the bias of the respondents, the selection process should be aligned it the way it ensures maximal diversity of the respondents and statistical stability of the results. Moreover, different chatbot user interfaces should be used to reduce a possible corresponding bias. Finally, after each dialogue a (short) feedback from a respondent should be collected (e.g., “Do you think that you were talking to a chatbot?” , “Were the answers helpful?”, etc.) to determine the current impressions of the users. In the same context, It would be very interesting at what frequency of such questions the users measurably change their behavior.

REFERENCES

[1] U. Gnewuch, S. Morana, and A. Maedche, “Towards designing cooperative and social conversational agents for customer service.” in Proceedings of the 38th International Conference on Information Systems (ICIS), 2017.

[2] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Deliv-ering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully au-tomated conversational agent (woebot): a randomized controlled trial,” JMIR mental health, vol. 4, no. 2, p. e7785, 2017.

[3] A. Androutsopoulou, N. Karacapilidis, E. Loukis, and Y. Charalabidis, “Transforming the communication between citizens and government through ai-guided chatbots,” Government Information Quarterly, vol. 36, no. 2, pp. 358-367, 2019.

[4] A. P. Chaves and M. A. Gerosa, “How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design,” International Journal of Human–Computer Interaction, vol. 37, no. 8, pp. 729-758, 2021.

[5] M. R. Mehl, “Quantitative text analysis.” 2006.

[6] D. Duijst, “Can we improve the user experience of chatbots with personalisation,” Master’s thesis. University of Amsterdam, 2017.

[7] M.-C. Jenkins, R. Churchill, S. Cox, and D. Smith, “Analysis of user interaction with service oriented chatbot systems,” in International Conference on Human-Computer Interaction. Springer, 2007, pp. 76-83.

[8] J. Hill, W. R. Ford, and I. G. Farreras, “Real con-versations with artificial intelligence: A compari-son between human–human online conversations and human–chatbot conversations,” Computers in human behavior, vol. 49, pp. 245-250, 2015.

[9] A. Schlesinger, K. P. O’Hara, and A. S. Taylor, “Let’s talk about race: Identity, chatbots, and ai,” in Proceed-ings of the 2018 chi conference on human factors in computing systems, 2018, pp. 1-14.

[10] Y. Mou and K. Xu, “The media inequality: Comparing the initial human-human and human-ai social interactions,” Computers in Human Behavior, vol. 72, pp. 432-440, 2017.

[11] E. Silkej, “Linguistic differences in real conversations: Human to human vs human to chatbot,” 2020.

[12] M. Oya, “Syntactic dependency distance as sentence complexity measure,” Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics, 01 2011.

[13] H. Liu, “Dependency distance as a metric of language comprehension difficulty,” Journal of Cognitive Science, vol. 9, pp. 159-191, 09 2008.

[14] J. Ure, “Lexical density and register differentiation,” Applications of Linguistics, pp. 443–452, 1971.

[15] G. Herdan, Quantitative Linguistics. London: Butterworth, 1960.

[16] H. H. Sommers, “Statistical methods in literary analysis,” The computer and literary style, pp. 128-140, 1966.

[17] P. Guiraud, Problèmes et Méthodes de la Statistique Linguistique. Paris: Presses universitaires de France, 1960.

[18] R. Flesch, “A new readability yardstick,” Journal of Applied Psychology, pp. 221–233, 1948.

[19] V. Solovyev, V. Ivanov, and M. Solnyshkina, “Assessment of reading difficulty levels in russian academic texts: Approaches and metrics,” Journal of Intelligent and Fuzzy Systems, vol. 34, pp. 1-10, 04 2018.

[20] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, “Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel,” Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.

[21] R. Senter and E. A. Smith, “Automated readability index,” CINCINNATI UNIV OH, Tech. Rep., 1967.

[22] E. B. Coleman, “Developing a technology of written instruction: Some determiners of the complexity of prose,” Verbal learning research and the technology of written instruction, pp. 155-204, 1971.

[23] I. E. Fang, “The “easy listening formula”,” Journal of Broadcasting & Electronic Media, vol. 11, no. 1, pp. 63-68, 1966.

[24] A. Both, A. Perevalov, J. R. Bartsch, P. Heinze, R. Iudin, J. R. Herkner, T. Schrader, J. Wunsch, A. K. Falkenhain, and R. Gürth, “A question answering system for retrieving german covid-19 data driven and quality-controlled by semantic technology,” in Joint Proceedings of the Semantics co-located events: Poster&Demo track and Workshop on Ontology-Driven Conceptual Modelling of Digital Twins co-located with SEMANTiCS 2021, ser. CEUR Workshop Proc., vol. 2774, CEUR-WS.org, 2021.

[25] L. D. Riek, “Wizard of oz studies in hri: a systematic review and new reporting guidelines,” Journal of Human-Robot Interaction, vol. 1, no. 1, pp. 119-136, 2012.

[26] J. Newn, R. Singh, F. Allison, P. Madumal, E. Velloso, and F. Vetere, “Designing interactions with intention-aware gaze-enabled artificial agents,” in IFIP Conference on Human-Computer Interaction. Springer, 2019, pp. 255-281.