Zipf's Law in Human-Machine Dialog

Linders, Guido M ; Louwerse, Max M

DOI: https://doi.org/10.1145/3383652.3423878

Zipf's Law in Human-Machine Dialog

Guido M. Linders
Dept. of Cognitive Science & Artificial Intelligence,
Tilburg University
Tilburg
g.m.linders@uvt.nl

Max. M. Louwerse
Dept. of Cognitive Science & Artificial Intelligence,
Tilburg University
Tilburg
m.m.louwerse@uvt.nl

ABSTRACT

Zipf's law is a mathematically relatively simple formula stating that the frequency of a word is inversely correlated with its rank. Zipf's law is well-known in computational linguistics and cognitive sciences alike. In the context of agent development, however, Zipf's law has hardly ever been mentioned. This is surprising as principles regarding language likely benefit the development of conversational agents. This paper serves as a starting point to explore the role of Zipf's law in agent development, showing that Zipf's law also applies to dialog. Moreover, it can shed light on human-machine dialog. In addition to word frequency distributions that demonstrate Zipf's law, we also included frequency distributions of words at specific positions in the sentence as well as turn lengths. Zipf's law was found in the far majority of analyses we conducted. In addition, we investigated whether Zipf's law can be used to detect differences between human and agent-generated speech through correlating the distributions and found that even though both the human and agent frequency distributions follow Zipf's law, these distributions are not necessarily similar, shedding light on where agent dialog may distinguish itself from human dialog. The findings in this paper can thus serve as a way to monitor to what extent ubiquitous patterns in human-human dialog are found in human-machine dialog.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Mathematics of computing** → *Distribution functions.*

KEYWORDS

Zipf's law, Human-machine dialog, Dialog Systems, Agent development

ACM Reference Format:
Guido M. Linders and Max. M. Louwerse. 2020. Zipf's Law in Human-Machine Dialog. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423878

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

ACM ISBN 978-1-4503-7586-3/20/09…$15.00
https://doi.org/10.1145/3383652.3423878

1 INTRODUCTION

Zipf's law is a surprising phenomenon, first reported for language. When the words in a corpus are counted and ordered on frequency, an inverse relationship between the position of the word in the ordered list (rank) and its frequency of occurrence is obtained [45, 46]. This is not the case for a specific written language corpus, but for about any written language corpus. Indeed, Zipf's law has been found across many languages and language groups [19, 34, 36] and has not only been found for word frequencies, but also for other linguistic phenomena, such as part-of-speech tags [36], *n*-grams [18] and number words [11]. What's more, Zipf's law has even been found for non-human communication, such as for dolphin whistles [32] and gesturing in gorillas [16].

The explanations why this pattern exists in language can be roughly classified in cognitive and statistical theories, neither one being mutually exclusive. Statistical theories argue that Zipfian patterns emerge as an artifact from the statistical properties of natural language data [28, 31, 33]. Because many other (non-language-related) systems also have those properties, Zipf's law can be considered (statistically) ubiquitous. Evidence for a statistical explanation of Zipf's law comes from randomly typed text, produced by a (simulated) monkey [28, 33]. These so-called random-typing approaches were taken to be an indication that Zipf's law could emerge without involvement of any cognitive processes. More recently, in [1], a set statistical conditions for a distribution to follow Zipf's law were outlined and the authors showed that it emerges under very mild conditions when the range of frequencies is very large. They also showed that a large range of frequencies can be induced by a latent variable. Specifically for word frequencies, they showed that part-of-speech tags as a latent variable induces a broad range of frequencies, which, in turn, very easily leads to the emergence of Zipf's law.

Many researchers have pointed out that theories, purely based on statistics cannot be the full story. After all, language and communication are intentional with language users not being random sequence generators [13, 36]. Still, the statistical properties of word frequency distributions, as well as of other (linguistic) phenomena following Zipf's law, could still partially explain the emergence of Zipf's law. For example, even though, it is not fully explained in [1] why part-of-speech tags induce a large range of word frequencies, the authors show that favorable statistics of natural language data can be part of the solution to explain the emergence of Zipf's law for word frequencies. If the emergence of Zipf's law is mainly a consequence of the statistical characteristics of the data, the question emerges why these favorable statistics occur in language in the first place. In other words, there must be some cognitive benefits for humans to adhere to Zipfian distributions.

Zipf, himself, framed the cognitive benefits of Zipf's law in terms of the *Principle of Least Effort* [46], which states that both the listener and speaker in a conversation try to optimize their behavior to increase the efficiency of the conversation and decrease the effort that is needed. More concretely, speakers try to minimize effort by using the smallest vocabulary as possible and listeners try to minimize effort by minimizing the ambiguity through using less ambiguous, but also less frequent words. In line with the Principle of Least Effort, word length also follows Zipf's law, meaning that shorter words are exponentially more frequent than longer words [37, 46]. Zipf predicted this would lead to more efficient communication, which was demonstrated in [37]. Many information-theoretic accounts have tried to quantify the optimization benefits of Zipfian distributions in computational models [10, 14]. Also, empirical evidence from language evolution studies have shown how these benefits may have shaped language over time [24, 26, 35]. In fact, behavior optimization in dialog is a well-established phenomenon with dialog participants aligning their speech on different levels to increase communicative efficiency and understanding [15, 29, 38].

Next to optimization benefits in the production and recognition of language, Zipfian distributions might also have other benefits for humans, biologically [8]. For example, in [3], it was argued that learning and memory seem well-adapted to Zipfian patterns of word frequencies. A parallel that was also made by [25], where the authors went a step further and argued that scaling laws, like memory retention functions, frequency distributions of animal foraging moves and Zipf's law are ubiquitous in cognitive systems. They argued that these scaling laws emerge as a consequence of these systems being in a critical state. These critical states provide an optimal balance between change and stability and are therefore favorable in many cognitive systems. Indeed, word frequency distributions that follow a Zipfian pattern, improve the learning of word-meaning mappings [22], linking the cognitive benefits of Zipfian distributions in word learning to memory. Humans are generally better in remembering more frequent words [23], as well as shorter words [4]. As a consequence, very frequent and short words are learned faster, which, in turn, decreases the referential uncertainty for low-frequency words and increases the learnability of them [22]. Finally, Zipfian distributions enhance the segmentation of speech into words, allowing for faster understanding [27].

Although Zipf's law has been found in many corpora, we know of no studies that have measured the presence of Zipf's law in human-machine dialog. This may be no surprise, considering the fact that Zipf's law has never been mentioned in the context of agent development. [9] did mention that the user-generated speech in the Dialogue State Tracking Challenge 2 corpus followed Zipf's law, but the agent-generated speech was not mentioned. Hence, it is unclear whether Zipf's law can be obtained for agent-generated speech.

Following the above mentioned statistical explanation, we would expect agent-generated speech to also follow Zipf's law, since conversational agents also use natural language. Following the cognitive explanation, however, the predictions are less clear. On the one hand, it might be the case that the agent-generated speech is naturally constrained by the topic of conversation, the task and the goals to communicate efficiently and aid the user. As a consequence, agent-generated speech might also follow Zipf's law. On the other hand, one could argue that most of the current dialog systems, especially the systems that are very constrained in their vocabulary and utterance selection, do not have the cognitive capabilities to minimize their speaking effort.

The motivation behind the current study is twofold. It can aid in understanding the phenomenon of Zipf's law as it adds new observations from a different domain (e.g., speech generated by a dialog system) to the debate. However, the current study can also shed light on the potential of using Zipf's law as an objective tool to measure the quality of agent speech. To illustrate this point, a parallel to the application of Zipf's law in the study of animal cognition may be worthwhile. The field of animal communication used Zipf's law to study the communicative ability of certain communication channels in animal behavior and argued that it can be used as quantitative measure for the complexity of the vocabulary [32]. Using an information-theoretic view on Zipf's law, it was argued in [32] that the exponent of Zipfian curve shows the degree of optimization of how much information can be transmitted through the communicative channel. It shows the balance between the repetitiveness and variation of the communicative units. Zipf's law could play a similar role in agent development.

This paper investigates whether Zipf's law is also present in human-machine dialog and whether differences can be observed between different linguistic frequency distributions extracted from the speech generated by a dialog system and speech generated by the user of the system. We will first analyze human-machine dialog corpora on a global level to investigate any differences in the goodness-of-fit to Zipf's law between the user and the dialog system. Next we will look within the frequency distributions of the dialog system and the user and investigate whether any differences on this level can be observed. If any differences between frequency distributions are found, it would mean that Zipf's law could distinguish agent-generated speech from human-generated speech and would show that Zipf's law has potential to be applied as a tool to agent development to monitor the human-like quality of the speech generate by the dialog system.

2 METHOD

2.1 Corpora

Three English corpora were used in the analyses presented below, one corpus with human-human dialog and two with human-machine dialog. For the human-human dialog corpus we selected the task-based Map Task corpus. The HCRC Map Task corpus contains dialogs between two participants that collaboratively solve a road map task [2]. Both participants received a map with or without a route. The instruction giver was asked to provide the route to the instruction follower. However, a slightly different map was used between the two participants who were not able to see each other's map. This generated a dialog negotiating on the route. In each dialog, there was a clear distinction between the roles of the participants, with an instruction giver mainly giving information and an instruction follower mainly receiving information.

In addition, we analyzed two corpora with human-machine dialogs, the Dialogue State Tracking Challenge (DSTC) corpus and the WOCHAT corpus. The Dialogue State Tracking Challenge (DSTC) corpora have been used in a shared task on improving the quality

Corpus	Dlgs	Turns	Words/Turn	Tokens	Types	Types/Tokens	With System	Domain-specific	Task-based	Spoken
Map Task	128	20,715	7.023	145,490	2,158	0.0148	no	yes	yes	yes
DSTC	5,510	87,371	9.341	816,128	2,049	0.0025	yes	yes	yes	yes
WOCHAT	607	19,872	5.452	108,349	6,687	0.0617	yes	no	no	no

Table 1: Overview of the evaluated corpora and their properties.

Corpus	Words/Turn	Word Tokens	Word Types	Types/Token
Map Task giver + follower	7.023	145,490	2,158	0.0148
Map Task giver	9.534	100,233	1,845	0.0184
Map Task follower	4.436	45,257	1,404	0.0310
DSTC system + user	9.341	816,128	2,049	0.0025
DSTC system	14.624	648,190	1,091	0.0017
DSTC user	3.901	167,938	1,335	0.0079
WOCHAT system + user	5.452	108,349	6,687	0.0617
WOCHAT system	6.119	61,451	4,928	0.0802
WOCHAT user	4.771	46,898	3,976	0.0848

Table 2: Comparison of the lexical statistics of the users with the dialog systems.

Speaker	Utterance
User	I would like to look for a different pub please
System	The Alma is an outstanding pub
User	I am looking for a pub with an internet connection
System	The Bakers which has an internet connection is an outstanding pub
User	Does it have television
System	I am sorry but there is no information on the Baker's food

Table 3: A sample taken from the DSTC corpus.

Speaker	Utterance
System	What do you like to do for fun?
User	listening to music
System	My favorite kind of it is techno but I also like Opera
User	ok my favourite one is Baroque music
System	And Baroque music is ok one
User	what is your favourite baroque composer?
System	I don't have a favorite baroque composer. My favorite color is unknown

Table 4: A sample taken from the WOCHAT corpus.

of dialog systems [44]. We used the data from the second [20] and third [21] edition. The data from both the second and third edition contain interactions where the user was searching for a restaurant, coffee shop or pub. We combined these corpora together, since they are very similar in the domain and the dialog systems being used, with both being frame-based dialog systems. Furthermore, we combined the training and test sets of both corpora together as well. Note that both the system and user have different roles in the dialog, with the system being the information giver and the user being the information requester.

The WOCHAT corpus includes dialogs between users and chatbots [12]. In this corpus, there was no goal and the users could freely interact with the chatbots about anything they liked. These dialogs where held in chat form and are thus written. The data was collected through a shared task where the goal was to collect dialog data. The whole corpus contains interactions between users and nine different chatbots. We excluded the chatbots that did not reply in English.

Both the dialog systems in the DSTC corpus and in the WOCHAT corpus used a rigid turn-taking system, alternating a system's turn with the user's turn and did not allow for interruptions and barge-ins. The dialog systems in the DSTC corpus were goal-driven and the algorithms designed to best estimate the user goal(s). The dialog systems in the WOCHAT corpus were designed to keep the conversation going for as long as possible. These systems include Eliza [43] and a prototype of Alice [42]. The dialog systems use different strategies to keep the conversation going, ranging from simple transformations of the input to retrieval-based systems that try to find the best reply in a database. Table 3 shows a sample from a dialog from the DSTC corpus and Table 4 from the WOCHAT corpus. These examples show that the sentences are mostly natural in isolation, but are not always natural in the context.

The three corpora were preprocessed by removing all punctuation marks and converting all letters to lower case. For the Map Task corpus, we divided the utterances into those of the instruction giver and those of the instruction follower. For the human-machine dialog corpora, we split the utterances in system-generated and user-generated utterances in the human-machine dialog corpora. We will refer to the former as *system* and to the latter as *user*. We did not modify or remove any further conventions and annotations in the original corpora, such as tokens replacing names for anonymization purposes.

Statistics on the corpora, such as the number of dialogs, turns and words are summarized in Table 1. The dialogs in the DSTC and WOCHAT corpora are very different in their nature. While

the DSTC corpus is domain-specific, task-based and based in a spoken setting, the WOCHAT corpus is task-free, not restricted to any domain and conducted in a chat setting. Despite the DSTC corpus being much larger, it has fewer word types and therefore a lower type-token ratio than the WOCHAT corpus.[1] This can likely be explained by to the task-based nature of the dialogs and the fact that the frame-based dialog systems in the corpus have limited flexibility in constructing utterances. The Map Task corpus is most similar to the DSTC corpus in being task-based, domain-specific and spoken. A comparison of the number of word types and tokens is shown in Table 2. In both human-machine corpora, the system talks significantly more than the user, which means that the conversational turns of the system were generally longer, since the dialog flow follows a rigid pattern of alternating turns between the system and user.

2.2 Linguistic Variables

Different linguistic variables for Zipfian distributions were considered. First, we included the traditional word frequencies. Zipfian distributions have been found for word frequencies within spoken contexts before [5, 41]. Next to word frequencies, we also considered words at a specific position in a sentence. More specifically, we considered the frequency distributions over the first word of a conversational turn. The first word often has the function of coordinating or "gluing" the conversation and might therefore be more relevant for the turn-taking dynamics in dialog [30]. In addition, we also evaluated the frequency distribution over the last word of a conversational turn.

As mentioned earlier, frequency distributions over word length also follow Zipf's law, with shorter words being exponentially more frequent [37, 46], a law commonly referred to as Zipf's Law of Abbreviation. Extending this law and Zipf's Principle of Least Effort [46] to the level of conversational turns, we hypothesized that shorter turns are more frequent than longer turns and suspected that their distribution could also be Zipfian. To quantify this, we counted the number of words in a turn and created a distribution where we kept the order of the number of words in a conversational turn fixed. Note that this introduces a second variable on which the frequencies are ordered. A similar approach was used in [11] to establish Zipf's law for number words, where the number words are not ordered on their frequency, but on their "size". As was the case in [11], this means that potentially the most frequently occurring turn lengths might not be at the top of the distribution. Hence, effectively we are considering a histogram that is ordered on the turn length here.

2.3 Evaluation Metrics

Historically, Zipf's law was defined as the inverse relationship between the rank and the frequency, modified by an exponent (α) and a scaling parameter (C), as is represented in (1).

$$f(r) = \frac{C}{r^{\alpha}} \tag{1}$$

[1]A comparison of type-token ratio is not fair if the corpora do not have the same size. However, it should be clear that the dialogs in the WOCHAT corpus contain significantly more word types than the DSTC corpus, despite having significantly less word tokens.

Variations have also been proposed, including ones that have more parameters and therefore also naturally provide a better fit. The most well-known version was introduced by Mandelbrot and was proposed as a generalization of Zipf's law [31]. It is shown in (2).

$$f(r) = \frac{C}{(r+\beta)^{\alpha}} \tag{2}$$

This formula contains an additional parameter (β), which can make the approximation more or less skewed, such that the formula can more closely approximate the distribution. Mandelbrot argued that Zipf's original formula could not explain the behavior of the frequency distribution in its entirety [31]. We will refer to the formula as the *Zipf-Mandelbrot formula* and we will call distributions following this formula near-Zipfian, consistent with the terminology used in [36]. We included both the original Zipf formula and the Zipf-Mandelbrot formula in our analysis.

Both formulas in our experiments were fitted by estimating its parameters using the maximum likelihood estimate (MLE). MLE has been shown to give better fits than its alternative, a linear regression on a log-log plot, which in turn, produces biased and inaccurate fits [17]. Finally, the R^2 determination coefficient was utilized to quantify the variance that the fit of the Zipf and Zipf-Mandelbrot formula to each frequency distribution can explain. This metric was used, since it is intuitive and has been used before in validating Zipf's law, for example in [36, 40].

3 RESULTS & DISCUSSION

We will first analyze the frequency distributions on a global level to investigate whether they follow Zipf's law. More concretely, we will measure the goodness-of-fit of the distributions on the whole corpus and on the system's and user's frequency distributions separately. Tables 5 and 6 show the goodness-of-fit, measured with the R^2 coefficient, of the Zipf and Zipf-Mandelbrot formulas respectively, on the frequency distributions from the Maptask corpus. Similarly, the results for the DSTC corpus are shown in Tables 7 and 8 and the results for the WOCHAT corpus in Tables 9 and 10. First, note that the goodness-of-fit of the Zipf-Mandelbrot formula on the distributions is generally higher than of the Zipf formula. This can be explained by the fact that an additional parameter is used, which should lead to a better fit. However, apart from absolute differences in Zipf's law, the tendencies of the results for both evaluation metrics are fairly similar. We can observe that, overall, almost all analyzed frequency distributions follow at least a near-Zipfian distribution. Except for the system's turn length distribution from the DSTC corpus, even the turn length frequency distributions adheres to Zipf's law, despite there being a second variable that is the substitute for the actual rank.[2]

Particularly noteworthy are the differences in goodness-of-fit scores between the DSTC corpus and the WOCHAT corpus. While on the DSTC corpus the system's turn length distribution has a low goodness-of-fit, the reverse is true on the WOCHAT corpus with the user having a lower goodness-of-fit. One possible explanation is that turn length distributions are simply more sensitive to less optimized distributions, because the frequencies are not ranked on their "size".

[2]Because Zipf's law orders the frequencies on their "size", the goodness-of-fit scores are automatically lower than when this order is not preserved, such as for turn lengths.

Variable	Giver + Follower	Giver	Follower
Word	0.911	0.905	0.883
First Word	0.934	0.911	0.953
Last Word	0.956	0.917	0.950
Turn length	0.938	0.775	0.977

Table 5: Goodness-of-fit (R^2) of the Zipf formula with the frequency distributions from the Map Task corpus.

Variable	Giver + Follower	Giver	Follower
Word	0.883	0.890	0.918
First Word	0.943	0.981	0.957
Last Word	0.986	0.957	0.981
Turn length	0.856	0.905	0.920

Table 6: Goodness-of-fit (R^2) of the Zipf-Mandelbrot formula with the frequency distributions from the Map Task corpus.

Variable	System + User	System	User
Word	0.718	0.724	0.637
First Word	0.849	0.934	0.882
Last Word	0.787	0.784	0.727
Turn length	0.659	0.352	0.701

Table 7: Goodness-of-fit (R^2) of the Zipf formula with the frequency distributions from the DSTC corpus.

Unit	System + User	System	User
Word	0.973	0.976	0.966
First Word	0.954	0.981	0.967
Last Word	0.989	0.973	0.987
Turn length	0.867	0.401	0.893

Table 8: Goodness-of-fit (R^2) of the Zipf-Mandelbrot formula with the frequency distributions from the DSTC corpus.

Unit	System + User	System	User
Word	0.862	0.835	0.897
First Word	0.931	0.938	0.901
Last Word	0.871	0.880	0.929
Turn length	0.624	0.718	0.524

Table 9: Goodness-of-fit (R^2) of the Zipf formula with the frequency distributions from the WOCHAT corpus.

Unit	System + User	System	User
Word	0.943	0.947	0.962
First Word	0.956	0.922	0.976
Last Word	0.944	0.976	0.984
Turn length	0.893	0.931	0.783

Table 10: Goodness-of-fit (R^2) of the Zipf-Mandelbrot formula with the frequency distributions from the WOCHAT corpus.

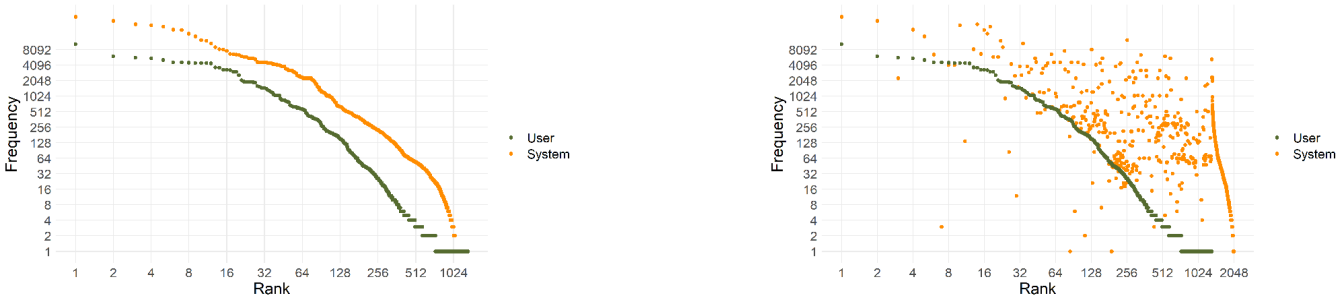
Overall, these results show Zipf's law is present in human-machine corpora too, but they do not show consistent major differences between the frequency distributions of the system and the user or between the human-machine and human-human dialog, at least not globally. One possible explanation for not observing any consistent differences on the goodness-of-fit with Zipf's law between system- and user-generated utterances is that both frequency distributions are relatively similar or exhibit similar statistical tendencies. With the distributions being approximations over many dialogs, we expect them to be relatively robust for individual differences. Again, arguably this comparison is not fair on the Map Task corpus and the DSTC corpus, since the dialog participants have different roles in the dialog. This might influence their word choice or the number of words they use in a conversational turn and, in turn, their frequency distribution, but we leave that issue aside for now.

In Figure 1, we show a selection of the graphs in which we compare the frequency distributions of the system and the user for the different linguistic units. For visualization purposes, we included only the instances that occurred at least once in both the system- and user-generated speech[3]. This means that ranks with a frequency count of zero are left empty in the graphs. Since the distributions are plotted on a log-log plot, a straight line would indicate a good fit with the Zipf formula. As expected, the left part of the distributions is usually curved, an effect which is generally present in Zipfian linguistic distributions [36]. This means that the highest frequencies are lower than the estimations given by the original Zipf formula.

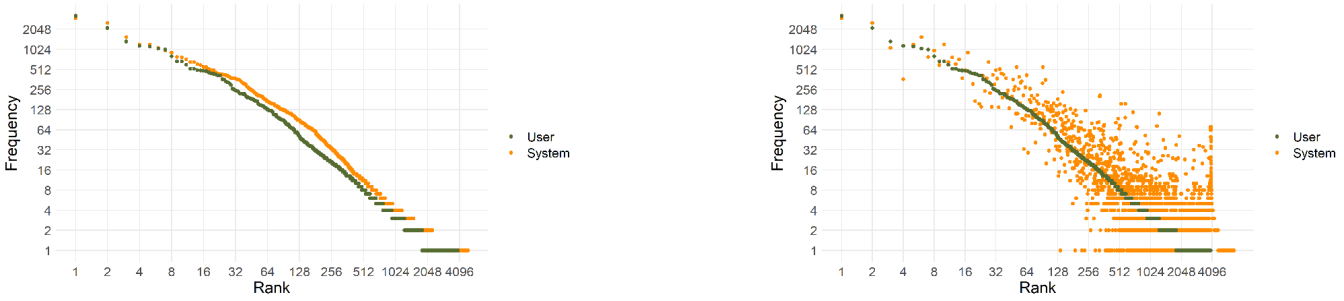
Figure 1a shows the word frequency distributions from the DSTC corpus, where the frequencies are ranked according their own respective ranks and the same frequencies are ranked both according to the user's ranks. Figure 1b shows the same, but with the distributions from the WOCHAT corpus. As the figures show, the system's distribution of the WOCHAT corpus seems closer to the user's distribution than when comparing the system's distribution on the DSTC corpus.

Figure 1c shows the first and last word frequency distributions from the DSTC corpus, respectively, where this time both distributions use the ranks from the system. As can be observed with the word frequency distribution of the system in the DSTC corpus, for both the first and last word frequency distributions, the tail (e.g., the distribution at the lower ranks) gets increasingly curved. This is an indication that the system has a limited vocabulary and set of phrases to utter, and thus repeats them. This effect was, to a lesser extent, also observed on the word, first word and last word frequency distributions of the system on the WOCHAT corpus, of which the latter are not shown here. This is an effect that is generally not observed in human word frequency distributions [36].

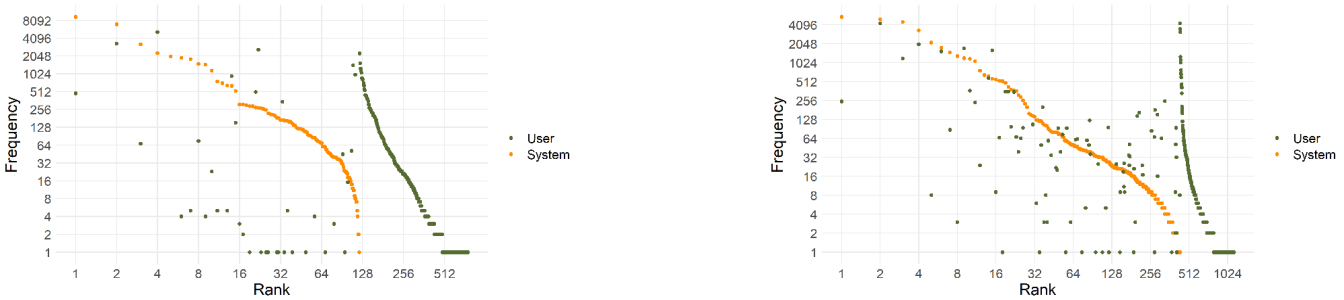
[3]Since we can best visualize Zipf's law on a log-log plot and since the logarithm of zero is not defined, we removed all zero counts.



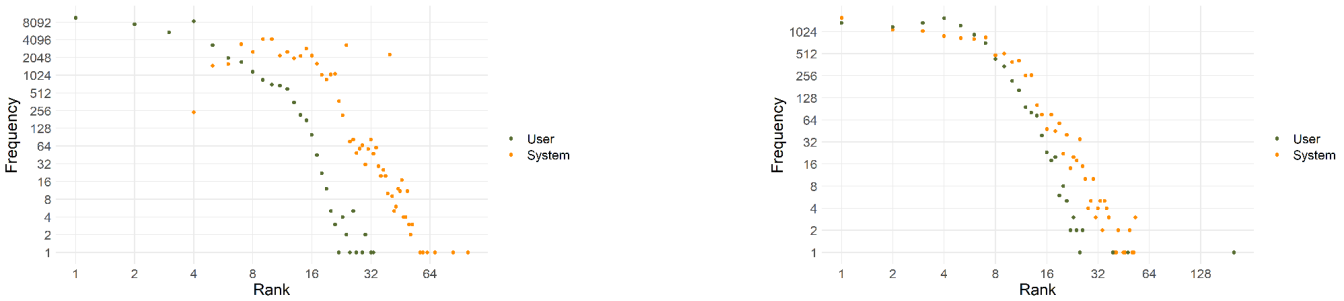
(a) Word frequency distributions from the DSTC corpus, both ranked according to their respective ranks (left) and the same distributions, ranked according to the user's ranks (right).



(b) Word frequency distributions from the WOCHAT corpus, both ranked according to their respective ranks (left) and the same distributions, ranked according to the user's ranks (right).



(c) First word (left) and last word (right) frequency distributions from the DSTC corpus, both ranked according to system's ranks.



(d) Turn length frequency distributions from the DSTC corpus (left) and the WOCHAT corpus (right). Note that these distributions cannot be ranked according to the system's or user's ranks, because they are ordered on the turn length.

Figure 1: A selection of graphs, where we compare the system's and user's frequency distributions.

Corpus	Word	First Word	Last Word	Turn Length
Map Task	0.864	0.871	0.874	0.859
DSTC	0.741	0.371	0.449	-0.017
WOCHAT	0.936	0.923	0.788	0.945

Table 11: Pearson's correlations between the frequency distributions of the system and user.

We can furthermore observe in Figure 1c a batch of points from the user's distribution after the last ranks of the system. These are the words that were not uttered by the system and thus do not have a rank assigned to them when we rank the frequencies according to the system's frequencies. Instead for visualization purposes, they have been ranked on their frequency. The effects are larger for the DSTC corpus because the system's and user's vocabularies show less overlap.

Finally, Figure 1d shows the turn length frequency distributions of the system and user on both human-machine corpora. It is important to note that the ranks of the turn length distributions are not the actual ranks, but the turn lengths themselves. This means that the distributions are not necessarily ordered from largest to smallest frequency. As was confirmed by the goodness-of-fit scores on a global level, the user's turn length distributions also follow Zipf's law. We can observe that these distributions seem more curved than the original Zipfian distributions. This means that human utterances and, to some extent, the system's utterances in a human-machine interaction are generally short and often contain only one word, perhaps contrary to what we would expect in an interaction. The system's utterances from the DSTC corpus are not shorter than four words. This could again be a consequence of the task-based scenario and different speaker roles in the dialogs.

Finally, we computed the correlations between the different system's and user's frequency distributions. In Table 11, we have summarized the correlations between the distributions of the instruction giver and follower from the Map Task corpus and between the system's and user's frequency distributions from the human-machine corpora. First, the correlations confirm our visual analysis. We see higher correlations on the word frequency distribution from the WOCHAT corpus than the distribution from the DSTC corpus. In fact, all correlations are higher on the WOCHAT corpus than on the DSTC corpus. These correlations are comparable to or even higher than the correlations on the Map Task corpus. No correlation was found between the turn length distributions of the system and the user. We suspect the role of the additional variable (e.g., the distributions not being ranked on their "'size", but on the turn length) might have played a role here, since the system's frequencies increase for the highest ranks, rather than decrease, as opposed to the user's distribution. This was also confirmed in the analysis of the global results.

Overall, the correlations are very high for the distributions from the WOCHAT corpus and substantial for the word frequency distribution from the DSTC corpus. This means both the system and user utter the same words roughly equally often. This might explain why we do not find any differences in Zipf's law between those distributions. Still, we observe lower correlations on the DSTC corpus for the first word, last word and turn length distributions, which could be caused by the nature of the dialogs and system and user having different roles. Interestingly, those differences were not visible in the goodness-of-fit estimations of Zipf's law between the systems's and user's first and last word frequency distributions. These findings suggest that the distributions between the system and user are not similar in the DSTC corpus and, in turn, that similar goodness-of-fit scores between the systems's and user's distributions cannot be explained by the system and user having similar linguistic distributions.

The high correlations on the WOCHAT corpus are perhaps surprising, because the chatbots in the WOCHAT corpus have a very limited notion of understanding and certainly do not have the capacity to choose or optimize their word use or turn lengths. The most likely explanation is that many chatbots reuse words or phrases from the user in their response, which could result in high correlations. Furthermore, participants in a dialog align their behavior at different levels, including at the level of words [15, 29, 38]. This effect has even been shown to hold for participants conversing with a dialog system [6, 7]. What's more, even a correlation between the degree of alignment between a dialog system and the user and task success has been found [39]. It might be the case that similar word frequency distributions have been found (partially) as a consequence of the user aligning their word frequency distributions with those of the dialog system. Finally, the speaker role could also have a significant impact on language use. Dialog participants in the WOCHAT corpus do not have speaker roles.

These findings suggest that Zipf's law is also present in the language generated by dialog systems. This does not however mean that the frequency distributions of the user and the dialog system are exactly the same, which was shown by the graphs and the correlations.

4 CONCLUSION

In the current paper, we have investigated the presence of Zipf's law in human-human and human-machine dialog. Results showed that in both human-human dialog and human-machine dialog alike, frequency distributions follow Zipf's law. We have not only shown this for word frequencies, traditionally being used to demonstrate Zipf's law, but also for the word frequencies at specific positions in the conversational turn, both at the start and end of a conversational turn. Furthermore, we have extended Zipf's law of abbreviation to the level of conversational turns and shown even turn length distributions follow a near-Zipfian distribution. We also compared the distributions of the dialog system and the user separately and found that human-generated and agent-generated could differ, even if both follow Zipf's law.

Despite the fact that Zipf's law has been reported so widely, it is somewhat surprising that it has not been demonstrated or used in the context of agent development. This is surprising as quantitative linguistic laws like Zipf's law can be used as a tool to objectively evaluate the quality of human-machine dialog and rate the extent to which a human-machine dialog is human-like or not. Morever, it can be used as a dialog management tool automatically monitoring whether agents use distributions of language features that show to be human-like. If agent-generated language in a dialog starts to deviate from human-like frequency distributions, regardless of

whether it is the word frequency distribution, the distribution over the first or last words of a conversational turn or even the turn length, the system can correct the deviation and bring the dialog back to a human-like distribution. This study has paved the way for such tools. In addition, it has added new observations to the debate of whether Zipf's law has a statistical or cognitive origin in language.

5 ACKNOWLEDGMENTS

This research has been funded by a grant No.: PROJ-007246 from the European Union, OP Zuid, the Ministry of Economic affairs, the Province of Noord-Brabant, and the municipality of Tilburg awarded to the second author. The usual exculpations apply.

REFERENCES

[1] Laurence Aitchison, Nicola Corradi, and Peter E Latham. 2016. Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS Computational Biology* 12, 12 (2016), e1005110.
[2] Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC Map Task Corpus. *Language and Speech* 34, 4 (1991), 351–366.
[3] John R Anderson and Lael J Schooler. 1991. Reflections of the environment in memory. *Psychological Science* 2, 6 (1991), 396–408.
[4] Alan D Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior* 14, 6 (1975), 575–589.
[5] Jaume Baixeries, Brita Elvevåg, and Ramon Ferrer-i-Cancho. 2013. The evolution of the exponent of Zipf's law in language ontogeny. *PloS ONE* 8, 3 (2013), e53227.
[6] Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9 (2010), 2355–2368.
[7] Susan E Brennan. 1996. Lexical Entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialogue*. 41–44.
[8] Morten H Christiansen and Nick Chater. 2008. Language as shaped by the brain. *Behavioral and Brain Sciences* 31, 5 (2008), 489–509.
[9] Andrei C Coman, Koichiro Yoshino, Yukitoshi Murase, Satoshi Nakamura, and Giuseppe Riccardi. 2019. An Incremental Turn-Taking Model for Task-Oriented Dialog Systems. In *Proceedings of Interspeech*. 4155–4159.
[10] Bernat Corominas-Murtra, Jordi Fortuny, and Ricard V Solé. 2011. Emergence of Zipf's law in the evolution of communication. *Physical Review E* 83, 3 (2011), 036115.
[11] Stanislas Dehaene and Jacques Mehler. 1992. Cross-linguistic regularities in the frequency of number words. *Cognition* 43, 1 (1992), 1–29.
[12] Luis F D'Haro, Bayan Abu Shawar, and Zhou Yu. 2016. REWOCHAT 2016–Shared Task Description Report. In *Proceedings of the Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents–Development and Evaluation*. 39–42.
[13] Ramon Ferrer-i-Cancho and Brita Elvevåg. 2010. Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE* 5, 3 (2010), e9411.
[14] Ramon Ferrer-i-Cancho and Ricard V Solé. 2003. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences* 100, 3 (2003), 788–791.
[15] Simon Garrod and Martin J Pickering. 2009. Joint action, interactive alignment, and dialog. *Topics in Cognitive Science* 1, 2 (2009), 292–304.
[16] Emilie Genty and Richard W Byrne. 2009. Why do gorillas make sequences of gestures? *Animal Cognition* 13, 2 (2009), 287–301.
[17] Michel L Goldstein, Steven A Morris, and Gary G Yen. 2004. Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems* 41, 2 (2004), 255–258.
[18] Le Quan Ha, Philip Hanna, Ji Ming, and F. J. Smith. 2009. Extending Zipf's law to n-grams for large corpora. *Artificial Intelligence Review* 32, 1 (2009), 101–113.
[19] Le Quan Ha, Elvira I Sicilia-Garcia, Ji Ming, and Francis Jack Smith. 2002. Extension of Zipf's law to words and phrases. In *Proceedings of the 19th International Conference on Computational Linguistics*, Vol. 1. Association for Computational Linguistics, 1–6.
[20] Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*. 263–272.
[21] Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 324–329.
[22] Andrew T Hendrickson and Amy Perfors. 2019. Cross-situational learning in a Zipfian environment. *Cognition* 189 (2019), 11–22.
[23] Charles Hulme, Steven Roodenrys, Richard Schweickert, Gordon DA Brown, Sarah Martin, and George Stuart. 1997. Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23, 5 (1997), 1217–1232.
[24] Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 2017. Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition* 165 (2017), 45–52.
[25] Christopher T Kello, Gordon DA Brown, Ramon Ferrer-i-Cancho, John G Holden, Klaus Linkenkaer-Hansen, Theo Rhodes, and Guy C Van Orden. 2010. Scaling laws in cognitive sciences. *Trends in Cognitive Sciences* 14, 5 (2010), 223–232.
[26] Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition* 141 (2015), 87–102.
[27] Chigusa Kurumada, Stephan C. Meylan, and Michael C. Frank. 2013. Zipfian frequency distributions facilitate word segmentation in context. *Cognition* 127, 3 (2013), 439–453.
[28] Wentian Li. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory* 38, 6 (1992), 1842–1845.
[29] Max M Louwerse, Rick Dale, Ellen G Bard, and Patrick Jeuniaux. 2012. Behavior matching in multimodal communication is synchronized. *Cognitive science* 36, 8 (2012), 1404–1426.
[30] Max M Louwerse and Heather Hite Mitchell. 2003. Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse Processes* 35, 3 (2003), 199–239.
[31] Benoît Mandelbrot. 1953. An Informational Theory of the Statistical Structure of Language. In *Communication Theory*, Willis Jackson (Ed.). Butterworths Scientific Publications, London, 486–502.
[32] Brenda McCowan, Sean F Hanser, and Laurance R Doyle. 1999. Quantitative tools for comparing animal communication systems: Information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour* (1999).
[33] George A Miller. 1957. Some effects of intermittent silence. *American Journal of Psychology* 70, 2 (1957), 311–314.
[34] Géza Németh and Csaba Zainkó. 2002. Multilingual statistical text analysis, Zipf's law and Hungarian speech generation. *Acta Linguistica Hungarica* 49, 3-4 (2002), 385–405.
[35] Matjaž Perc. 2012. Evolution of the most common English words and phrases over the centuries. *Journal of The Royal Society Interface* 9, 77 (2012), 3323–3328.
[36] Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21, 5 (2014), 1112–1130.
[37] Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108, 9 (2011), 3526–3529.
[38] Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27, 2 (2004), 169–190.
[39] David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language* 76 (2014), 29–46.
[40] Arjuna Tuzzi, Ioan-Iovitz Popescu, and Gabriel Altmann. 2010. *Quantitative analysis of Italian texts*. RAM-Verlag, Lüdenscheid.
[41] Marolein Van Egmond. 2018. *Zipf's law in aphasic speech: An investigation of word frequency distributions*. Ph.D. Dissertation. Utrecht University.
[42] Richard S. Wallace. 2009. *The Anatomy of A.L.I.C.E.* Springer Netherlands, Dordrecht, 181–210.
[43] Joseph Weizenbaum. 1966. ELIZA —- a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.
[44] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*. Association for Computational Linguistics, Metz, France, 404–413.
[45] George Kingsley Zipf. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, London.
[46] George Kingsley Zipf. 1949. *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge.