The Duration of a Turn Cannot be Used to Predict When It Ends

Charles Threlkeld and **JP de Ruiter**
Tufts University
{charles.threlkeld, jp.deruiter}@tufts.edu

Abstract

Turn taking in conversation is a complex process. We still do not know how listeners are able to anticipate the end of a speaker's turn. Previous work focuses on prosodic, semantic, and non-verbal cues that a turn is coming to an end. In this paper, we look at simple measures of duration—time, word count, and syllable count—to see if we can exploit the duration of turns as a cue. We find strong evidence that these metrics are useless.

1 Introduction

Turn-taking is a fundamental aspect of dialogue. Timing of turn initiation is critical. Sometimes long pauses are socially relevant (Bogels et al., 2015). Sometimes people overlap in conversation without the reason being clear (Heldner and Edlund, 2010). When trouble occurs, people can pause to signal misunderstandings (Mertens and De Ruiter, 2021). But turn taking as a whole is not well understood.

What is known is that the time between successive turns is generally very short—much shorter than can be attributed to simple reactions to a turn ending (De Ruiter, 2019). What this means is that people must anticipate the end of a turn (Ruiter et al., 2006). If we are anticipating the end of a turn, then there must be some features of utterances that we use to predict their ending, enabling fluid turn transition.

In artificial agents that engage in spoken dialogue, turn-taking often falls by the wayside, leading to stilted conversations with long delays between turns or interruptions at inappropriate times (Skantze, 2021). Typical human interactions with current conversational agents work uniformly sequentially, as the agent processes and responds to the human once the end of an utterance has been completed, and it does not expect interruptions or overlaps (Gervits et al., 2020).

In general, computers can process information much faster than humans, but we have not yet developed fluid turn-taking algorithms. Humans prepare a one-word utterance in around 600ms (Indefrey and Levelt, 2004). Computers can perform much faster than this and their speed is still increasing. But if an agent doesn't know when a turn ends, fluidity can be compromised. For smooth turn taking, agents need to know how to time their contributions appropriately.

Previous research looks at lexical (Magyari and de Ruiter, 2012), semantic (Gervits et al., 2020; Riest et al., 2015), prosodic (Bögels and Torreira, 2015), or non-verbal (Roddy et al., 2018) attributes of utterances in order to anticipate turn ends. Each of these has its own merits and drawbacks. Lexical boundaries are relatively easy to compute and reason about. Semantic completion of an utterance makes logical sense for an end-point to a thought. Prosodic cues can be computed quickly from the speech signal, and non-verbal cues are ripe for deep learning techniques (Lala et al., 2019). Turn *duration*, however, has not been studied yet for its use as a cue in anticipating its end, despite its ready availability to any spoken dialogue system.

Intuitively, one would expect that the duration of a turn is a strong cue about its ending. It would be plausible to assume that the longer someone has been talking already, the higher the probability is that the speaker will end their turn. Compare it to waiting for a bus – we tend to assume that the longer we have waited for the bus, the higher the probability that it will finally arrive. But this is only so when the duration of a turn is normatively constrained. However, looking at distribution of a large number of conversational turns in Dutch, De Ruiter (2019) found that the distribution of turn-duration looks suspiciously much like an exponential distribution. And a unique and counter-intuitive property of this distribution is that it has a constant *hazard*

Proceedings of the SIGdial 2022 Conference, pages 361–367
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 ©2022 Association for Computational Linguistics

rate: no matter how long we have waited for the process to complete, the probability of it terminating in the next instant remains constant. If turn durations are in fact exponentially distributed, it would mean that the duration of a turn so far does not contain any information about its projected duration.

However, the observation in De Ruiter (2019) were only for one small corpus in Dutch, and measured in milliseconds. It could be that measuring duration in other units, like words, syllables, or other turn-related units would show a different distribution. In this study we set out to study if this suspected property of turn durations is generalizable to a larger corpus in English, and to other units of duration.

Turns in dialogue are composed of turn construction units (TCUs). TCUs are bounded by transition relevant places (Sacks et al., 1978). At each transition relevant place, another person could take the floor or the current speaker could continue. In this study, we will investigate the duration of both TCUs and entire turns. As there may be social preferences regarding the number of TCUs within a turn, we will also examine the usefulness of the number of TCUs per turn in predicting floor transitions.

In the following sections, we outline the data collection and our statistical analyses. Then we will show the distributions of the data and the statistical models describing the data. We will then discuss the implications of our results, and present ideas on how these results can be used to improve spoken dialogue agents.

2 Methods

2.1 Dataset

For this study, we are using the Switchboard corpus (Godfrey and Holliman, 1993). The Switchboard corpus is a large, well-studied corpus of dyadic, open-ended telephone conversations. Its use limits our ability to draw conclusions about face-to-face speech patterns, but extends the work of De Ruiter (2019, p.542–543) — a study of Dutch telephone conversations — to English. Since the corpus is well-studied, we can draw on previous work for transcription, timing, and segmentation.

We used two transcriptions of the Switchboard corpus. First, the Mississippi State University transcriptions[1] were used for word-by-word timing. Second, the Discourse Language Modeling Project transcriptions[2] break the conversation into turn construction units. We are interested in TCUs as the basic building blocks of turns, and to compare that to the analysis of turn duration in De Ruiter (2019, p.542–543) which only looked at duration in seconds.

After merging these two sources, we analyzed only conversations where the word-level exact matches were at least 90% of words in a conversation, and the total error rate of the conversation (that is, words matching none of our word-matching heuristics) was below 2%. Heuristics included accounted for simple, systematic alternative transcriptions, like repeated or omitted words, alternative spellings (“uh-uh” / “uh-huh”), or abandoned words (“ho-” / “how”). The analyses use the resulting 75 conversations with 5,857 turns and 11,796 turn construction units.

2.2 Probabilistic Modeling

For each aspect of the data, we will build two models. The first model will be a best-fit exponential distribution; the second will be a best-fit gamma distribution (except for TCUs per turn; see below). We will show the curves of the data along with curves for each model so that we can quantify and visualize the differences in prediction between the models and in reference to the data. Full descriptive statistics can be found in the appendix.

We chose the exponential distribution as a null model. It is the maximum entropy distribution for positive-domain data with a known mean. It also has the property of being memoryless, or having a constant hazard rate. This means that no matter how long an exponential process has been ongoing, the chance that it will end in the next time step is constant. This makes it a good null model, as there is very little information that can be gained about a distribution via the exponential distribution. Both the mean and the hazard rate are related to the single distribution parameter λ, which is the hazard rate or probability of the process ending in the next time-step. More concretely, λ is the chance of stopping at the next millisecond, word, or syllable, given that the process has not stopped so far.

The gamma distribution is a generalization of the exponential distribution. It is parameterized by

[1] http://www.openslr.org/5/

[2] Available at https://web.stanford.edu/~jurafsky/ws97/

a shape and rate parameter. If the shape parameter is one, the gamma is equivalent to the exponential distribution. Importantly, other shape parameters allow a gamma model to fit many different positive-domain datasets with different modes and varying hazard rates.

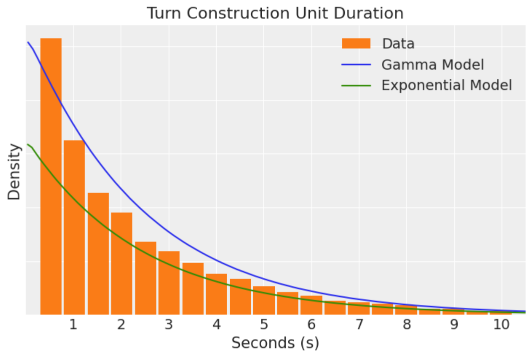
TCUs per Turn data is fit to Geometric and Negative Binomial distributions, which are the discrete forms of the Exponential and Gamma distributions, respectively. The p and n parameters of the discrete distributions mirror the $rate$ and $shape$ parameters of their continuous analogues. This decision was made because the small number of TCUs per turn does not lend itself to an assumption of continuity.

We will compare the exponential and gamma models for each dataset using the widely-applicable information criterion (WAIC). The WAIC estimates the effective number of parameters to adjust for overfitting, and gives results similar to a leave-one-out cross-validation for model-fitting. A lower WAIC is a better fit.

3 Results

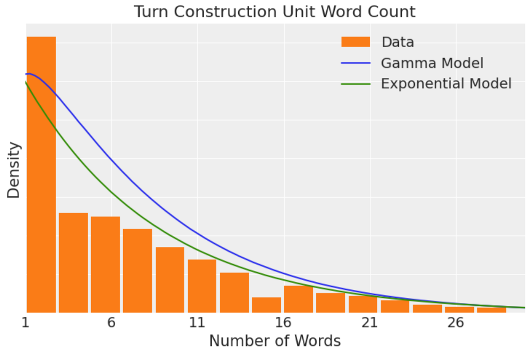
Here we will report the basic findings of our analyses. The interpretation of the findings will be delayed to discussion.

3.1 TCU Duration



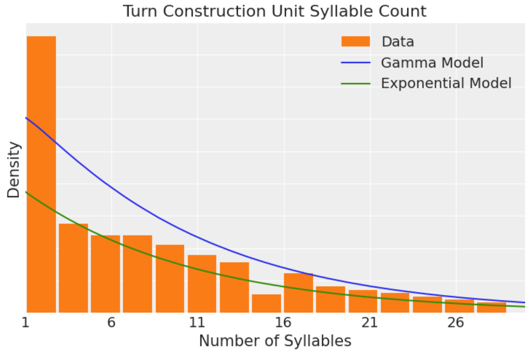
TCU duration exponential and gamma models were very similar, since the best-fit gamma model has a shape of 1.01, which is effectively the same as an exponential distribution. We can see this close fit in the WAIC scores, too, which were identical to the third decimal place.

3.2 TCU Word Count



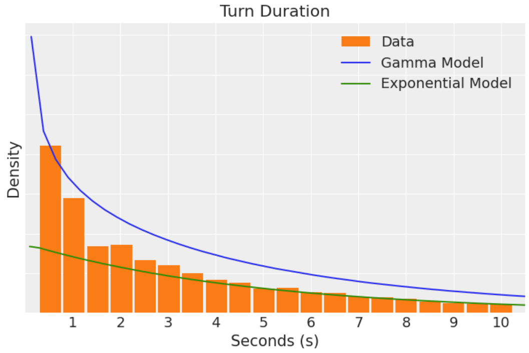
TCU word count was interesting in that it was the only model with a gamma shape parameter substantially above one, at 1.18. We can see this reflected in the concavity of the gamma model curve near zero. WAIC scores were very similar, with the exponential distribution only 0.3% higher than the gamma distribution.

3.3 TCU Syllable Count



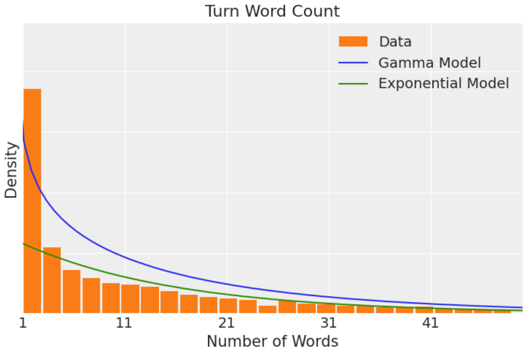
The TCU syllable count gamma model had a shape parameter of 1.05, very nearly identical to an exponential distribution. We can see the similarity in the chart. The WAIC is again only 0.3% higher for the exponential model than the gamma model.

3.4 Turn Duration



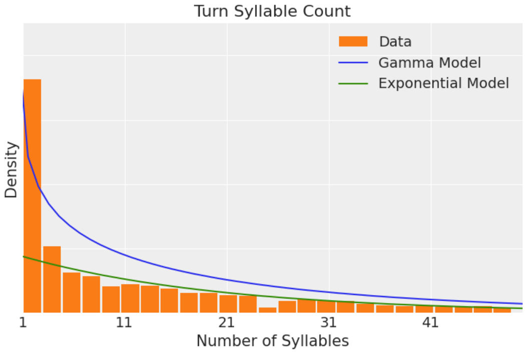
In the turn duration statistics, we see a sizable difference between where the gamma model is positioned and the exponential model, but also the data. The best fit gamma pulls the curve toward the tail in order to accommodate the high number of fast turns in the data. Despite the different orientations to the data, the trade-off between good fit on low values or good fit at high values cancel out and the WAIC is again only 0.3% higher for the exponential model than the gamma model.

3.5 Turn Word Count



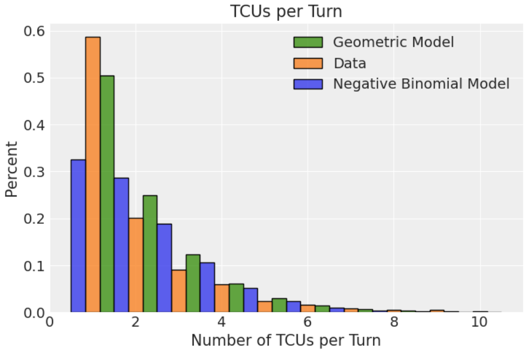
The turn word count models show similar tendencies to the turn duration models—the gamma model better accounts for low values, but the exponential distribution fits better at higher values. The low shape parameter of the gamma distribution (0.718) allows this distortion. The WAIC of the exponential distribution is 1.1% higher than the gamma.

3.6 Turn Syllable Count



Similar to the turn words model, a low gamma shape value—only 0.667—lets the gamma model account for many utterances with very few syllables compared to the expectations in the exponential distribution. Here we have our largest WAIC disparity with at 1.65% higher for the exponential model compared to the gamma model.

3.7 TCUs per Turn



TCUs per turn was fit to geometric and negative binomial models rather than exponential and gamma models. The negative binomial model has underestimated the low numbers, trading off probability mass at low counts for larger predictions at high counts. The WAIC shows that the simpler exponential model is a better fit with a 20% lower score.

4 Discussion

The analyses above, when taken together, suggest that there is little to be learned from examining the length of utterances as a sole heuristic for predicting their end. As was suspected on theoretical grounds (De Ruiter, 2019) there is very little information in simple duration. The work here extended this previous work from timing of TCus to examine semantic content as shown by word counts, phonetic information as shown by syllable counts, or social action as shown by TCUs per turn. None of these linguistics frames showed any substantial departure from the constant hazard-rate distribution.

There may be contexts where utterance length is a useful heuristic for TCU or turn end or situations in which the statistics describe here do not fit well. For example, one would suspect that different dialogue acts may lend themselves to different TCU lengths — short backchannels, for example. Or, particular social situations may lend themselves to fewer TCUs per turn to ensure participants maintain the same mental models. A follow-up study on (e.g.,) the Map Task Corpus might show these deviation, if they exist.

The largest deviation from the exponential model occurred in the turn word and syllable count analyses. These two results reflect the combination of high rates of single TCU turns and large shares of low TCU word and syllable counts. Our TCU per turn analysis shows that single-TCU turns are more common than the exponetial model is able to fit, but the negative binomial distribution moves probability mass to the tail, making it an even worse fit, both visually and statistically. Short turns skew the data in turn-level word and syllable count toward very low counts as compared to the exponential model.

We expected that the TCUs per turn negative binary model would account for the high number of single TCU turns, much like the gamma model does for the other data views. However, the geometric model outperforms the negative binomial by the largest WAIC difference of any models discussed. Therefore, we must conclude that the geometric model is the superior fit, and the best-fit hazard rate for each TCU—the chance that the speaker's turn is over at the end of any TCU—is 50%, or a coin flip. So, not only is the maximum entropy geometric model a better fit, but there is no reliable bias for whether a turn is over at the end of a TCU.

5 Conclusion

In this paper, we first confirmed the suspicions raised in De Ruiter (2019)—the duration of TCUs follows an exponential distribution. We then extended these findings in several ways. First, TCUs also follow this distribution by syllable or word count. Conversation does not orient to the amount of phonological or semantic information. It follows that if these factors are useful for turn taking, they are useful based on their meaning and structure, not their quantity or base informational load.

Next, we expanded our findings to the turn level, rather than just TCUs. Turn duration, syllable, and word count findings were akin to those at the TCU level, and so we must draw the conclusions that these turn length measurements are not useful either to exploit as information source in turn taking.

Finally, we looked at TCUs per turn for evidence that the number of dialogue acts of which a turn has numerical norms. The TCU per turn analysis showed that the end of a TCU is essentially a coin flip for whether there will be a floor transfer. Not only did the maximum entropy distribution have the best fit, but the hazard rate was very close to 0.5. So, we must conclude that there is no more pragmatic pressure to end one's turn when it is already very long.

Our general conclusion therefore is that, surprisingly, the duration of turns are not useful cues for turn segmentation or turn taking decisions. This is independent of whether we use temporal, phonological, lexical, or TCU-based measures of information. Agents that do turn taking will need to use linguistic or prosodic cues other than duration to achieve accurate timing in their turn taking behavior.

Acknowledgments

This paper was funded in part by a grant from the Data Intensive Studies Center at Tufts University.

References

Sara Bogels, Kobin H. Kendrick, and Stephen C. Levinson. 2015. Never say no ... how the brain interprets the pregnant pause in conversation. *PLOS ONE*, 10(12):e0145474.

Sara Bögels and Francisco Torreira. 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Jan P. De Ruiter. 2019. Turn-taking. *The Oxford Handbook of Experimental Semantics and Pragmatics*, page 536–548.

Felix Gervits, Ravenna Thielstrom, Antonio Roque, and Matthias Scheutz. 2020. It's about time: Turn-entry timing for situated human-robot dialogue. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 86–96, 1st virtual meeting. Association for Computational Linguistics.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62. *Linguistic Data Consortium*.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

P Indefrey and W.J.M Levelt. 2004. The spatial and temporal signatures of word production components. *Cognition*, 92(1–2):101–144.

Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *2019 International Conference on Multimodal Interaction*, pages 226–234.

Lilla Magyari and J. P. de Ruiter. 2012. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, 3.

Julia Beret Mertens and J. P. De Ruiter. 2021. Cognitive and social delays in the initiation of conversational repair. *Dialogue & Discourse*, 12(1):21–44.

Carina Riest, Annett B Jorschick, and Jan P de Ruiter. 2015. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in Psychology*, 6:89.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal continuous turn-taking prediction using multiscale rnns. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 186–190.

Jan-Peter de Ruiter, Holger. Mitterer, and N. J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

A Appendix

A.1 TCU Duration

Data Mean	2.40E+03
Exponential Model Mean	2.40E+03
Gamma Model Mean	2.40E+03
Data Std Dev	3.36E+03
Exponential Model Std Dev	2.40E+03
Gamma Model Std Dev	2.39E+03
Exponential Rate Mean	4.16E-04
Gamma Rate Mean	4.22E-04
Exponential Rate Std Dev	3.79E-06
Gamma Rate Std Dev	6.27E-06
Gamma Shape Mean	1.01E+00
Gamma Shape Std Dev	1.17E-02
Exponential WAIC	2.101E+05
Gamma WAIC	2.101E+05

A.2 TCU Word Count

Data Mean	7.70E+00
Exponential Model Mean	7.70E+00
Gamma Model Mean	7.70E+00
Data Std Dev	7.62E+00
Model Std Dev	7.70E+00
Gamma Model Std Dev	7.10E+00
Exponential Rate Mean	1.30E-01
Gamma Rate Mean	1.53E-01
Exponential Rate Std Dev	1.20E-03
Gamma Rate Std Dev	2.23E-03
Gamma Shape Mean	1.18E+00
Gamma Shape Std Dev	1.38E-02
Exponential WAIC	7.275E+04
Gamma WAIC	7.255E+04

A.3 TCU Syllable Count

Data Mean	9.81E+00
ExponentialModel Mean	9.80E+00
Gamma Model Mean	9.80E+00
Data Std Dev	1.01E+01
Exponential Model Std Dev	9.80E+00
Gamma Model Std Dev	9.56E+00
Exponential Rate Mean	1.02E-01
Gamma Rate Mean	1.07E-01
Exponential Rate Std Dev	9.44E-04
Gamma Rate Std Dev	1.57E-03
Gamma Shape Mean	1.05E+00
Gamma Shape Std Dev	1.21E-02
Exponential WAIC	7.852E+04
Gamma WAIC	7.850E+04

A.4 Turn Duration

Data Mean	4.83E+03
Exponential Model Mean	4.82E+03
Gamma Model Mean	4.83E+03
Data Std Dev	7.93E+03
Exponential Model Std Dev	4.83E+03
Gamma Model Std Dev	5.56E+03
Exponential Rate Mean	2.07E-04
Gamma Rate Mean	1.56E-04
Exponential Rate Std Dev	2.69E-06
Gamma Rate Std Dev	3.49E-06
Gamma Shape Mean	7.55E-01
Gamma Shape Std Dev	1.23E-02
Exponential WAIC	1.116E+05
Gamma WAIC	1.113E+05

A.5 Turn Word Count

Data Mean	1.53E+01
Exponential Model Mean	1.53E+01
Gamma Model Mean	1.53E+01
Data Std Dev	2.34E+01
Exponential Model Std Dev	1.53E+01
Gamma Model Std Dev	1.81E+01
Exponential Rate Mean	6.53E-02
Gamma Rate Mean	4.69E-02
Exponential Rate Std Dev	8.55E-04
Gamma Rate Std Dev	1.04E-03
Gamma Shape Mean	7.18E-01
Gamma Shape Std Dev	1.14E-02
Exponential WAIC	4.390E+04
Gamma WAIC	4.341E+04

A.6 Turn Syllable Count

Data Mean	1.95E+01
Exponential Model Mean	1.95E+01
Gamma Model Mean	1.95E+01
Exponential Model Std Dev	1.95E+01
Data Std Dev	3.02E+01
Gamma Model Std Dev	2.39E+01
Exponential Rate Mean	5.13E-02
Gamma Rate Mean	3.42E-02
Exponential Rate Std Dev	6.72E-04
Gamma Rate Std Dev	7.73E-04
Gamma Shape Mean	6.67E-01
Gamma Shape Std Dev	1.05E-02
Exponential WAIC	4.676E+04
Gamma WAIC	4.600E+04

A.7 TCUs per Turn

Data Mean	1.99E+00
Geometric Model Mean	1.99E+00
Neg Binomial Model Mean	1.99E+00
Data Std Dev	1.90E+00
Geometric Model Std Dev	1.40E+00
Neg Binomial Model Std Dev	1.62E+00
Geometric p Mean	5.03E-01
Neg Binomial p Mean	7.59E-01
Geometric p Std Dev	4.61E-03
Neg Binomial p Std Dev	1.12E-02
Neg Binomial n Std Dev	6.27E+00
Neg Binomial n Std Dev	3.82E-01
Geometric	16209.397714
NegativeBinomial	20315.258775

B Supplemental Material