Analyzing contingent interactions in R with `chattr`

Marisa Casillas (mcasillas@uchicago.edu)
Comparative Human Development, 1101 E. 58th St.
Chicago, IL 60637 USA

Camila Scaff (camila.scaff@iem.uzh.ch)
Institute of Evolutionary Medicine, University of Zurich
Zurich, Switzerland CH-8057

Abstract

The `chattr` R package enables users to easily detect and describe temporal contingencies in pre-annotated interactional data. Temporal contingency analysis is ubiquitous across signal system research, including human and non-human animal communication. Current approaches require manual evaluation (i.e., do not scale up), are proprietary/over-specialized (i.e., have limited utility), or are constructed ad-hoc per study (i.e., are variable in construct). `Chattr`'s theoretically motivated, customizable, and open source code provides a set of core functions that allow users to quickly and automatically extract contingency information in data already annotated for interactant activity (via manual or automated annotation). We demonstrate the use of `chattr` by testing predictions about turn-taking behavior in three language development corpora. We find that the package effectively recovers documented variation in linguistic input given both manual and automatically created speech annotations and note future directions for package development key to its use across multiple research domains.

Keywords: Contingency; interaction; turn taking; LENA; communication; R; software.

Introduction

`Chattr` is an R package that facilitates the detection and analysis of temporally contingent interactions in pre-annotated data (github.com/marisacasillas/chattr-basic).[1] Its utility extends across studies of human interaction, non-human animal communication, and contingencies within multi-modal signals. Despite significant common conceptual ground between these domains, definitions of contingency phenomena and implementations of contingency detection remain inconsistent, foregoing critical opportunities for the accumulation of shared construct validity. Such divergences are partly due to a lack of flexible contingency analysis tools: existing systems are either constructed ad-hoc, limited in use, or proprietary. `Chattr` improves this situation by: (1) taking inspiration from conversation analysis, psycholinguistics, and language development to provide theoretically sound, but customizable measurements of temporally contingent interaction at scale and (2) accepting a handful of generic formats as input, opening up its analytical framework to broad application (e.g., child language input, multi-party conversation, non-human animal signaling, event contingencies, etc.). Here we review `chattr`'s theoretical basis, describe the package's core functions, and demonstrate its use in three existing datasets.

[1] All documentation and scripts are available at the URL; the fully packaged version will be available before May 2021.

Contingent interaction

Joint coordination of action by two or more agents usually involves temporal contingencies. Whether we are making music with others, crossing a busy intersection, or chatting with a friend, the timing of our contributions to a coordinated event is crucial to its success. Optimal strategies for coordination often involve turn taking, that is: typically, only one interactant makes their contribution at a time, and decisions about who contributes when can be determined flexibly (as in conversation) or in a pre-defined manner (as in a debate). This sequential structure enables interactants to adapt each contribution such that it relevantly progresses the joint activity and to initiate unplanned sub-sequences (e.g., repairing misunderstandings) without breaking progress toward the larger goal.

Turn-taking (and similar) interactions are essential for communication across the animal kingdom (Fröhlich et al., 2016; Pika, Wilkinson, Kendrick, & Vernes, 2018), as well as AI systems interacting with human users. In humans, turn-taking interactions may be the only reliable source of language universals (Levinson, 2019). Traditionally, these kinds of interactional contingencies have been studied using careful inspection and analysis, both qualitative and quantitative, of manual measurements from video and audio recordings. However, recent advances in recording devices and automated annotation software (e.g., for voice detection) have created a growing need for new analytical approaches that can capitalize on very large, but relatively noisy datasets that cannot feasibly be assessed by hand.

Current contingency detection approaches (and their limitations)

At present, the most widely used tool for automated contingency analysis of human interaction is the LENA system (Greenwood, Thiemann-Bourque, Walker, Buzhardt, & Gilkerson, 2011), which was built for use with young children, but has also been employed to capture adult language environments (e.g., Rodríguez-Arauz, Ramírez-Esparza, García-Sierra, Ikizer, & Fernández-Gómez, 2019). The system includes both a recording device and a set of proprietary software tools that enable the user to collect long-format

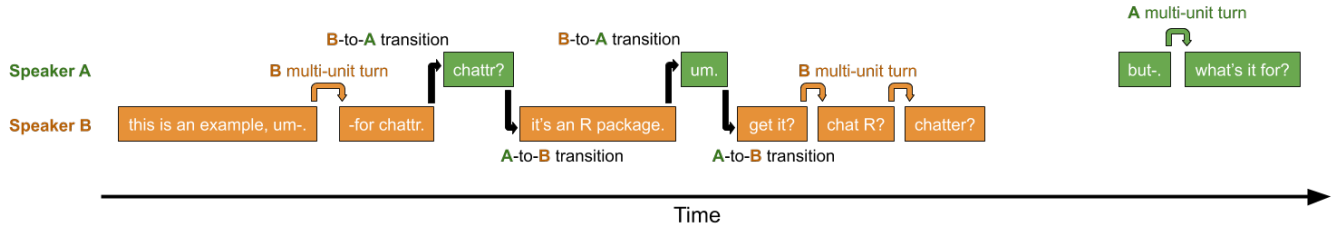

Figure 1: An example of a brief dyadic interaction between two English speakers: A (green) and B (orange). The producers here use both single- and multi-unit turns. There are 6 turns (3 from each producer), 4 turn transitions (two each from B to A and vice versa; black arrows), and one interactional sequence (the contiguous block of producer continuation/transition marked with green/orange arrows; the other turn ('but-. what's it for?') has no transitions and so is not in an interactional sequence).

(16-hour) participant-centric audio recordings and then automatically analyze them for a range of properties, including when vocalizations occur by speakers of different types (e.g., near/far female adult vocalizations). The software then uses the detected vocalizations to find candidate regions of vocal exchange (VABs; Vocal Activity Blocks) between the target child and nearby adults and calculates the estimated number of speaker exchanges that involve the child. It uses temporal contingency to associate speaking turns from different speaker types (i.e., <5 seconds of silence between child and woman/man vocalizations or vice versa). This convenient automated annotation system has been critical to spurring on new research on language development and turn-taking (e.g., Romeo et al., 2018) but has a few unfortunate drawbacks. Reliability estimates for turn count estimates are between 0.3 and 0.6 (Cristia, Bulgarelli, & Bergelson, 2020), with systematically worse errors for younger infants (Ferjan Ramírez, Hippe, & Kuhl, 2021).[2] The system is also proprietary, expensive, and can only be used with recordings made with LENA hardware. Research groups who lack generous funding or who have unique hardware and storage requirements will struggle to enjoy its benefits. Lastly, LENA is designed for child-centric recordings, which improves the accuracy of its application in the developmental language context, but offers minimal utility for those working in other domains.

Beyond LENA, approaches to extracting temporal contingencies have been much more variable. For example, in studies of adult conversation, researchers vary in what timing windows qualify as contingent, what types of contributions count toward turn taking, the modality in which communication is taking place, in how many interactants are considered to be involved (or are of interest), and so on, as is suitable to the research question (e.g., Ten Bosch, Oostdijk, & Boves, 2005; Fröhlich et al., 2016; Heldner & Edlund, 2010; Pika et al., 2018; Roberts, Torreira, & Levinson, 2015). These studies, while heterogeneous in data types and determinants for how and when to count turn-taking exchanges, have typically been inspired by the same set of core concepts from conversation analysis, building up significant theoretical common ground for understanding moment-to-moment processes of interactant coordination. Much of the work on language development, by contrast, has inherited the somewhat idiosyncratic concepts and terminology introduced by the LENA system, leaving a conceptual disjunct between work on turn-taking behaviors in children, adults, and non-human animals. Given the various restrictions on existing tools and free variations in analysis across studies, there is a clear need for a free, flexible, and theoretically grounded tool that can extract temporal contingencies at scale; `chattr` fills this need.

The `chattr` system

In brief, `chattr` is an R package that gives both summary and detailed data on temporal contingencies in pre-annotated data. To keep things simple, it has a single core function for each type of input that it takes: (a) LENA .its files; (b) tab delimited .txt tables with one production/utterance per row (e.g., exported from Praat, ELAN, etc., Figure 2; Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006; Boersma & Weenink, 2021); and (c) .rttm tables, a common output format used with automated speech diarization systems.[3] Users can use the default settings for each function—including limits on the relevant temporal windows, potential interactants, and which productions are considered—or can customize as desired. More advanced users can capitalize on the numerous sub-functions utilized by the core input-type functions to tailor `chattr`'s functions to their unique needs. All settings, output information types, and theoretical background is thoroughly summarized in the online documentation on the project's GitHub repository.

Core concepts We encourage users to first evaluate how well `chattr`'s concepts of 'turn', 'transition', and 'interactional sequence' fit those of the study context; our default definitions differ from those typically used in the language development literature and are restricted compared to their full (and human conversation-specific) meanings in conversation analysis (Sacks, Schegloff, & Jefferson, 1974; Schegloff, 2007). We briefly summarize these core concepts here

[2] CTC error estimates inherit error from earlier steps in the processing pipeline (e.g., misidentifying speech as silence).

[3] If interested in a fully open-source pipeline for child language environments, try Lavechin et al.'s (2021) voice type classifier.

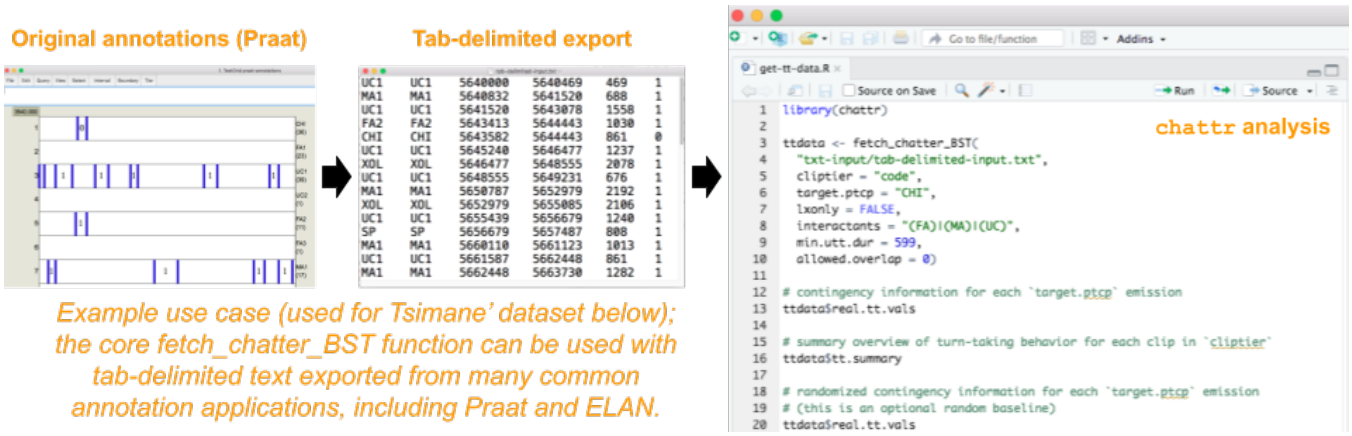

Figure 2: Example workflow for an annotation file using 'chattr'.

(also illustrated in Figure 1). We use the terms 'producer' and 'recipient'/'addressee' rather than 'speaker' and 'listener' to underscore the utility of these concepts across modalities, species, and interactional contexts:

A *'turn'* comprises one or more closely occurring emissions by the same producer. That is, a turn can be formed of multiple complete emissions (e.g., utterances/communicative acts) that may be separated by pauses in production so long as (a) there is no intervening emission from another producer and (b) the pause in production is short. An example of a single-unit turn in English is "Jane is the one in the hat.". An example of a multi-unit turn in English is "Jane is the one in the hat [pause] third from the left."

A *'turn transition'* occurs when one producer's turn stops and another producer's turn begins. Every turn transition has a pre-transition producer and a post-transition producer—these must be different individuals. The transition *begins* when the first turn ends and *ends* when the second turn starts. Therefore, if the second turn starts before the first turn ends, the transition time is negative ('transitional overlap'). If the second turn starts after the first turn ends, the transition time is positive ('transitional gap').

An *'interactional sequence'* is an unbroken turn-taking sequence between the target interactant and one or more of their interactional partners. Interactional sequences likely index more structurally complex, engaged interactional behaviors than single turn transitions do—akin to conversational bouts (or LENA VABs) during which participants can more substantially build on joint goals.

The `chattr` default settings are designed for human spontaneous conversation, including child conversation, which demonstrates fairly robust timing patterns (with some systematic variation) across the signed and spoken languages that have been analyzed (Levinson, 2019). The three most critical default settings are that: (a) up to 2000 ms of transitional gap or up to 1000 ms of transitional overlap is allowed between turns, (b) transitions can occur between turns of any duration, content, and from any potential interactional partner, and (c) when there are multiple potential prompts or responses (e.g., two interactants answer a question nearly simultaneously), `chattr` picks the production that occurs closest to the present one. Users interested in emulating LENA's CTC measure with their .its files can use a specialized function in which the target producer is assumed to be "CH" (target child), potential interactants are limited to "FA"/"MA" (female and male adult), and analyzed turns contain some linguistic material.

Example use case Suppose that I am interested in investigating how adult turn-taking varies in a dataset across semi-structured contexts (e.g., during board game play). I would ensure that the annotations are formatted as tab-delimited text (e.g., Figure 2). Then I would use the core Basic Speech Table call `fetch_chatter_BST()` to fetch turn-taking information. I might also want to, e.g., define a minimum utterance duration and a more strict temporal window for contingency, as well as calculate 10 randomized simulations of turn-taking rates to assess the baseline likelihood contingency: `fetch_chatter_BST(filename, min.utt.dur = 1500, allowed.gap = 1000, allowed.overlap = 600, n.runs = 10)`. This call yields detailed tables of detected turn-taking behavior ready for the author's statistical analysis of choice.

Pilot analysis

We demonstrate the use of `chattr` with three child language environment datasets from unrelated rural Indigenous communities: specifically, we sanity check `chattr`'s performance on corpora for which we have strong a priori hypotheses about basic turn-taking patterns. The analyzed recordings document children's verbal interactional patterns over full days at home in understudied and rural populations, for which use of a tool like LENA would be challenging. `Chattr` allows us to examine interactional patterns at scale in these corpora, evading months of manual annotation that would achieve the same, and making it easy to do so for both conventional (child-adult interaction) and non-conventional (child-child interactional) categories relevant to development in these con-

texts. The first two corpora, Tseltal (Mayan; Chiapas, Mexico; N = 10) and Yélî Dnye (isolate; Milne Bay, Papua New Guinea; N = 10), come from the Casillas HomeBank repository (Casillas et al., 2017) and were made with near parallel methods: children under age 3;0 wore an Olympus WS-832/853 audio recorder at home for 8–11 hours. The third corpus, Tsimane' (Tsimane'; Bolivia; 40 recordings from 27 children) features children under 6;0 who wore one of multiple recording devices (LENA, Olympus, or USB) at home for 4–21 hours (Scaff, Stieglitz, Casillas, & Cristia, n.d.); we focus here on the subset of those 17 recordings made with LENA (from 13 children). In each dataset we assess the baseline turn-taking rate over age and the frequency of interactions with other children. For the Tsimane' corpus can also compare `chattr` estimates on both LENA (automated) and manually created annotations of the same recording minutes. This pilot studies thus test whether previously documented patterns in these children's linguistic input are recapitulated in their turn-taking behavior, as detected by `chattr`.

Study 1. Tseltal and Yélî Dnye

We analyze interactional behavior in 20 clips for each recording: 9 randomly selected clips (5 min for Tseltal and 2.5 min for Yélî Dnye), 5 clips manually selected for day-peak turn-taking behavior of the target child with one or more interactants (each 1 min), 5 clips manually selected for day-peak vocal activity by the target child (each 1 min), and one 5-minute expansion on the most active turn-taking/vocal-activity clip. Each clip was manually annotated for all hearable speech, including addressee coding (e.g., target-child-directed vs. other-directed; see Casillas et al. (2020b, 2020a)). Despite documented differences in caregiver-child interactional style, day-long linguistic input estimates show similar patterns in these two communities. While female adult speech constitutes the majority of linguistic input in both communities, Yélî children show a marked increase in directed speech from other children with age. This pattern appears more weakly in the Tseltal data. We therefore expected to find that: (1) turn-taking rates are higher in turn-taking and vocal activity clips than in random clips, (2) rates are similar between the two communities, and (3) interactional sequences involving other children increase with age, particularly for Yélî children.

Methods We use `fetch_chatter_AAS()`, which is specifically designed for those using the ACLEW[4] Annotation Scheme (Casillas, Bergelson, et al., 2017). It allows 2000 ms of gap and 1000 ms of overlap at turn transitions and searches over all annotated utterances (any duration, content, and from any speaker). We limit our analysis to utterances directed exclusively to the target child. We also indicate the annotated regions by using the `cliptier` argument.

Results The mean rate of turn transitions in the Tseltal corpus was 11.8 and 3 transitions per minute for the active (turn taking and vocal activity) and random clips, respectfully, and 12.8 and 2.4 transitions per minute for Yélî Dnye. The distribution of turn taking rates across annotated clips was similar between the two sites (Table 1). A linear mixed effects regression of transitions per minute with predictors of clip type, corpus, and their interaction, and a random intercept for child reveals that random clips indeed have significantly lower transition rates (B = -8.78, SE = 1.2, t = -7.31). There is no evidence for a significant difference in rates between languages (t = 0.54) and no evidence for a clip type-language interaction (t = -0.74).

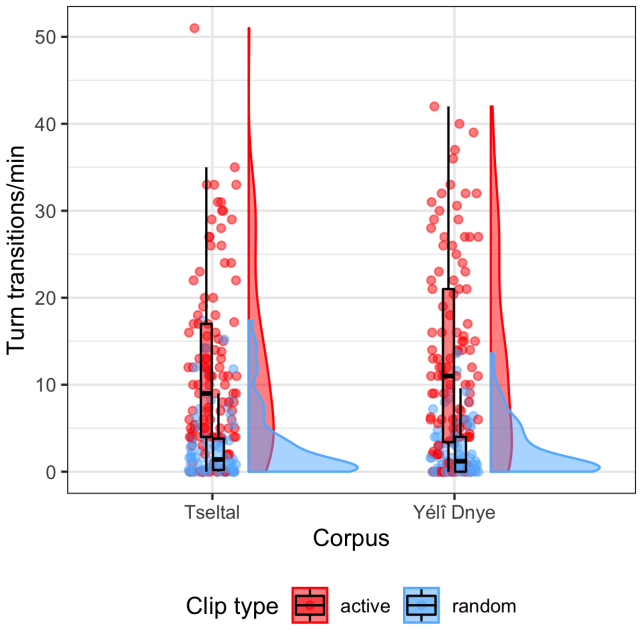

Figure 3: Turn transition rate by corpus, divided across manually selected turn-taking/high-vocal-activity clips (red) and random clips (blue).

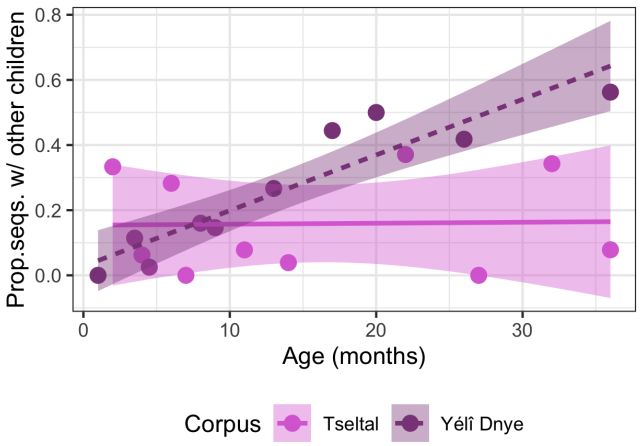

Figure 4: Proportion of interactional sequences involving at least one non-target child across age, by language.

[4] sites.google.com/view/aclewdid/.

A second linear mixed effects regression of the proportion of interactional sequences featuring at least one non-target child with predictors of age (in months), corpus, and their interaction, and a random intercept for child reveals that there is indeed a significant age-by-corpus interaction by which Yélî children show a larger increase in other-child interactional sequences with age compared to Tseltal children (B = 0.01, SE = 0.01, t = 2.47). There is no evidence for simple effects of age (t = 0.2) or language (t = -0.99).

Study 2. Tsimane'

These Tsimane' recordings were first automatically analyzed with LENA and then subsequently (and independently) manually annotated in 1-minute clips every 60 minutes, starting at the 34th minute (min 34-35, min 94-95, min 154-155, etc.; Scaff et al., n.d.). Both annotation types (LENA and manual) encode (a) when speech was occurring and (b) what type of speaker produced it (i.e., the target child, a nearby woman/man/other child, or other) for each of the hand-annotated minutes. Prior analysis shows comparably low rates of directed speech in these Tsimane' data to the Tseltal and Yélî Dnye recordings, again with a high proportion of directed input coming from other children (Scaff et al., n.d.). We therefore expected to find that: (1) despite their slightly different operationalizations, turn-taking rates are overall similar to what we found in the random samples of the other two communities , (2) turn-taking sequences involving other children are comparable to or more frequent than those in the random samples of the other two communities, (3) interactional sequences involving other children increase with age, and (4) manual and automated speech annotations of the same audio clips result in similar turn-taking estimates.

Corpus	Clip type	mean (sd; range), median
Tseltal	active (manual)	11.8 (4.8; 4.5-20.1), 12.3
Tseltal	random (manual)	3 (3.1; 0.4-10.6), 2.3
Yélî Dnye	active (manual)	12.8 (6.5; 3.9-22.2), 10.8
Yélî Dnye	random (manual)	2.4 (1.6; 0.5-6), 2.2
Tsimane'	random (LENA)	3.2 (1.1; 1.2-5.1), 3.1
Tsimane'	random (manual)	3.2 (1.2; 1.3-6), 3

Methods We use `fetch_chatter_BST()` with the manually annotated data, matching conditions of the call as closely as possible to what can be compared in the LENA output files, that is: include woman, man, and other-child speech, both linguistic and non-linguistic, with a minimum utterance duration of 600ms (the LENA lower limit) and no overlap allowed (meaningful overlap is not possible in LENA). With the automatic LENA annotations on the same recordings (the same 1-minute segments) we adjust the default settings on `fetch_chatter_LENA()` to reflect these same restrictions.

Results A linear mixed effects regression of transitions per minute with a fixed effect of annotation type (LENA vs. manual) and a random intercept for child reveals that turn-transition rates are similar between the two annotation methods (B = -0.09, SE = 0.41, t = -0.23). As expected, turn-transition rates are similar to what we found in the Tseltal and Yélî Dnye random clips, at 3.2 transitions per minute, though with fewer instances of rates above 10/min (Table 1).

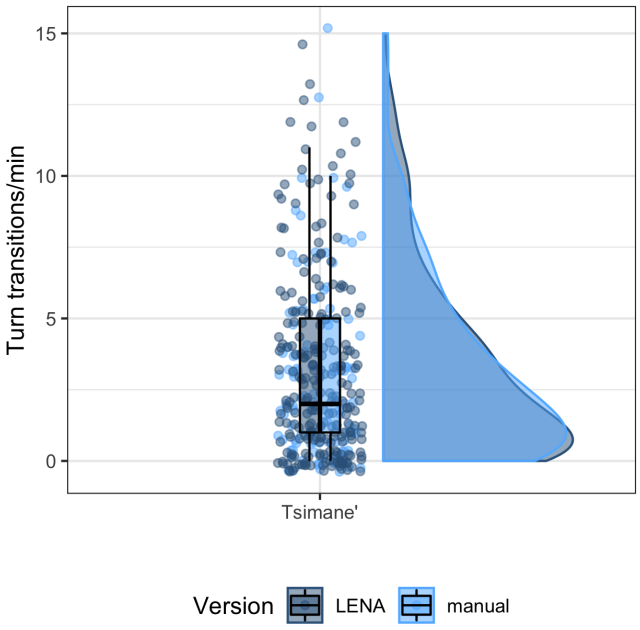

Figure 5: Turn transition rate by annotation type (LENA automated vs. manual) in the same audio clips. Clips are a periodic random sample of the daylong recording.

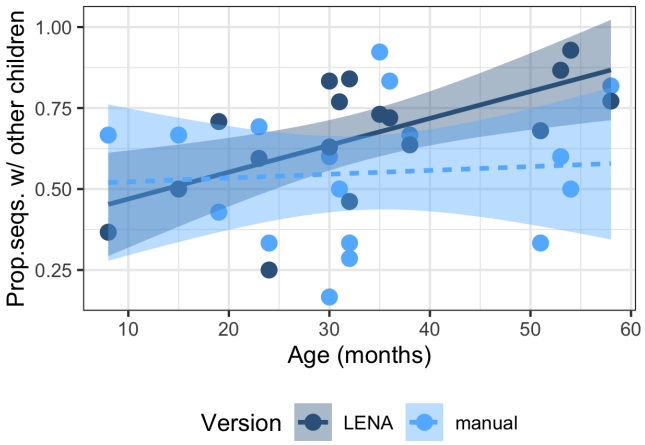

Figure 6: Proportion of interactional sequences involving at least one non-target child across age, by annotation type (LENA automated vs. manual) in the same audio clips.

A second linear mixed effects regression of the proportion of interactional sequences featuring at least one non-target child with predictors of age (in months), annotation type (LENA vs. manual), and their interaction, and a random intercept for child reveals that, as expected, there is a significant increase in other-child interactional sequences with age

(B = 0.01, SE = 0, t = 3.39). There is no evidence for simple effects of annotation type (t = 0.71) or for an age-annotation type interaction (t = -1.39).

Contribution and next steps

The `chattr` package allows users to easily implement theoretically informed contingency analyses on a wide variety of data types, including both automatically and manually annotated data. The package is designed for both straightforward (i.e., basic `fetch_chatter` calls) and customized analysis scenarios and provides detailed outputs that can be merged with other data about the same recordings. By providing a single tool for analyzing the most common input formats used for interactional data in psychology, animal behavior, and speech technology research, `chattr` aims to help build theoretical and methodological connections regarding the nature of contingent behaviors across diverse domains. While `chattr` has now been tested on a variety of child language datasets, new functionality will emerge following user issue-posting and feature requests. Following the beta stage of development, we will make the package available on CRAN for easier distribution. A critical next step will also be the development of tutorial materials to accompany the documentation, enabling new R users to quickly apply the core functions to a sampling of common use cases.

Acknowledgements

We thank Caitlin Fausey, Andrea Imhof, and Kunmi Sobowale for helpful contributions during initial code development. We are also hugely indebted to the participating families and local research assistants who made these three datasets possible. This work is supported by a NWO Veni Innovational Scheme grant (275-89-033) to MC.

References

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer. Retrieved from `http://www.praat.org`

Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., VanDam, M., & Sloetjes, H. (2017). A new workflow for semi-automatized annotations: Tests with long-form naturalistic recordings of children's language environments. In *Proceedings of INTERSPEECH 2017* (pp. 2098–2102). Stockholm, Sweden.

Casillas, M., Brown, P., & Levinson, S. C. (2017). Casillas HomeBank corpus. `http://doi.org/10.21415/T51X12`

Casillas, M., Brown, P., & Levinson, S. C. (2020a). Early language experience in a Papuan community. *Journal of Child Language*, 1–23.

Casillas, M., Brown, P., & Levinson, S. C. (2020b). Early language experience in a Tseltal Mayan village. *Child Development*, *91*(5), 1819–1835.

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the language environment analysis system segmentation and metrics: A systematic review. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1093–1105.

Ferjan Ramírez, N., Hippe, D. S., & Kuhl, P. K. (2021). Comparing automatic and manual measures of parent–infant conversational turns: A word of caution. *Child Development*, 1–10.

Fröhlich, M., Kuchenbuch, P., Müller, G., Fruth, B., Furuichi, T., Wittig, R. M., & Pika, S. (2016). Unpeeling the layers of language: Bonobos and chimpanzees engage in cooperative turn-taking sequences. *Scientific Reports*, *6*(1), 1–14.

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Communication Disorders Quarterly*, *32*(2), 83–92.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568.

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2021). An open-source voice type classifier for child-centered daylong recordings. Retrieved from `http://arxiv.org/abs/2005.12656`

Levinson, S. C. (2019). Interactional foundations of language: The interaction engine hypothesis. In *Human language: From genes and brain to behavior* (pp. 189–200). MIT Press.

Pika, S., Wilkinson, R., Kendrick, K. H., & Vernes, S. C. (2018). Taking turns: Bridging the gap between human and animal communication. *Proceedings of the Royal Society B*, *285*(1880), 20180598.

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology*, *6*, 509.

Rodríguez-Arauz, G., Ramírez-Esparza, N., García-Sierra, A., Ikizer, E. G., & Fernández-Gómez, M. J. (2019). You go before me, please: Behavioral politeness and interdependent self as markers of simpatía in Latinas. *Cultural Diversity and Ethnic Minority Psychology*, *25*(3), 379.

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-million-word gap: Children's conversational exposure is associated with language-related brain function. *Psychological Science*, *29*(5), 700–710.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7–55). Elsevier.

Scaff, C., Stieglitz, J., Casillas, M., & Cristia, A. (n.d.). Daylong audio recordings of young children in a forager-farmer society show low levels of verbal input with minimal age-related change.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I* (Vol. 1). Cambridge university press.

Ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues.

Speech Communication, *47*(1-2), 80–86.
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (pp. 1556–1559).