Understanding Language Through TalkBank

Brian MacWhinney🄳
Department of Psychology, Carnegie Mellon University

Abstract

Advances in computer technology have produced a flood of new data sets for understanding human language. However, nearly all of these new data sets are based on written, rather than spoken, language. This means that, despite their importance, open-access data on naturally occurring spoken-language conversations are much more difficult to obtain. The major exception to this is the TalkBank system, which provides online multimedia data for 15 types of spoken-language data: language in aphasia, child language, stuttering, child phonology, autism spectrum disorder, bilingualism, conversation analysis, classroom discourse, dementia, psychosis, right hemisphere damage, Danish conversation, second-language learning, traumatic brain injury, and daylong recordings in the home. This article reviews these resources and describes the ways that they are being used to further our understanding of language learning and usage.

Advances in computer technology have produced a flood of new data sets for understanding human language (Goldstone & Lupyan, 2016). However, nearly all of these new data sets are based on written, rather than spoken, language. The TalkBank system (http://talkbank.org) is the world's largest open-access integrated repository for spoken-language data. It provides language transcripts, media, and resources to support researchers in psychology, linguistics, education, computer science, and speech pathology. There are now more than 12,000 articles based on the use of TalkBank data and programs. The system emphasizes open-science principles by making its data, programs, and tools free and easily accessible. It provides data from 47 languages across 15 content areas. These corpora have been contributed by hundreds of researchers, all using a common transcription format called CHAT. The use of this common format has allowed us to construct an innovative set of resources and tools for analyses across languages, corpora, and speakers. This review describes these resources and some of the many analyses they have supported.

Databases

TalkBank includes 35 GB of transcript data and 12 TB of media data, all served online from the Carnegie Mellon University Campus Cloud facility. Links to each of these 15 TalkBank databases can be found at https://talkbank.org and include:

1. AphasiaBank for the study of language in aphasia
2. The Child Language Data Exchange System (CHILDES), which contains data from 47 languages from language-learning children between infancy and age 6
3. FluencyBank for the study of language fluency and disfluency in stuttering, aphasia, second-language learning, and normal processing
4. HomeBank for the study of daylong recordings in the home and elsewhere
5. PhonBank for the analysis of phonological development in 18 languages
6. ASDBank for child language in autism spectrum disorder
7. BilingBank for the study of adult bilingualism and multilingualism

Corresponding Author:
Brian MacWhinney, Department of Psychology, Carnegie Mellon University
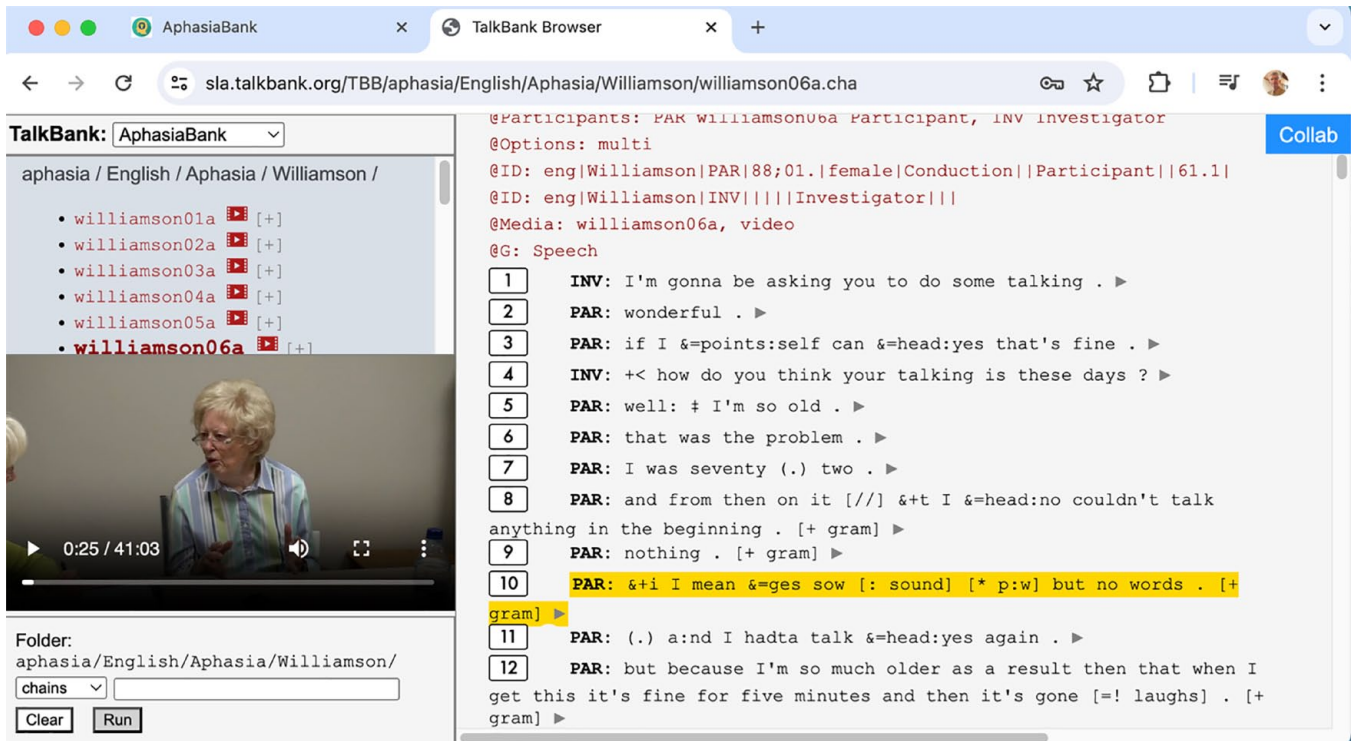Email: macw@cmu.edu

Fig. 1. The TalkBank Browser window.

8. CABank for the study of adult conversation often using the methods of conversation analysis
9. ClassBank for the study of language in the classroom
10. DementiaBank for the study of language in dementia
11. PsychosisBank for the study of language in psychosis
12. RHDBank for the study of language in right hemisphere disorder
13. SamtaleBank for the study of conversations in Danish
14. SLABank for the study of second-language acquisition
15. TBIBank for the study of language in traumatic brain injury

TalkBank Browser

The TalkBank Browser (TBB) provides direct online access to realistic multimodal spoken-language interactions across hundreds of corpora in TalkBank. The TBB uses a custom-made browser that allows users to navigate through the corpora and transcripts to play back media continuously while each line is highlighted in the transcript. Figure 1 illustrates how this looks for a transcript from AphasiaBank. The left side of the screen includes the transcript navigator, the video display, and a window for running analysis programs. The video display can be repositioned and enlarged for clearer viewing. The top of the window summarizes information about the currently selected transcript, and the lower part of the window provides utterances for playback.

Collaborative Commentary

In the upper right corner of Figure 1 there is a little blue button labeled "Collab." Clicking on this button opens a facility called Collaborative Commentary (CC) that allows groups to analyze spoken interactions collaboratively, often guided by a coding system. Figure 2 illustrates how students in a senior-level university class in psycholinguistics coded the second utterance in the *biggirl.cha* transcript from the case study by Michael Forrester in CHILDES.

Using one of the 14 codes that had been established by the instructor, the students uniformly coded this utterance as serving to express the child's self-concept ($SELF) in regard to food preferences. The codes were all combined with further comments. Coding and comments such as these are stored in a database separate from the main transcript database. The instructor can also click on any student comment to send an email providing feedback to the student who created the comment.

CC is being used actively by research groups interested in developing new modes of transcript analysis.
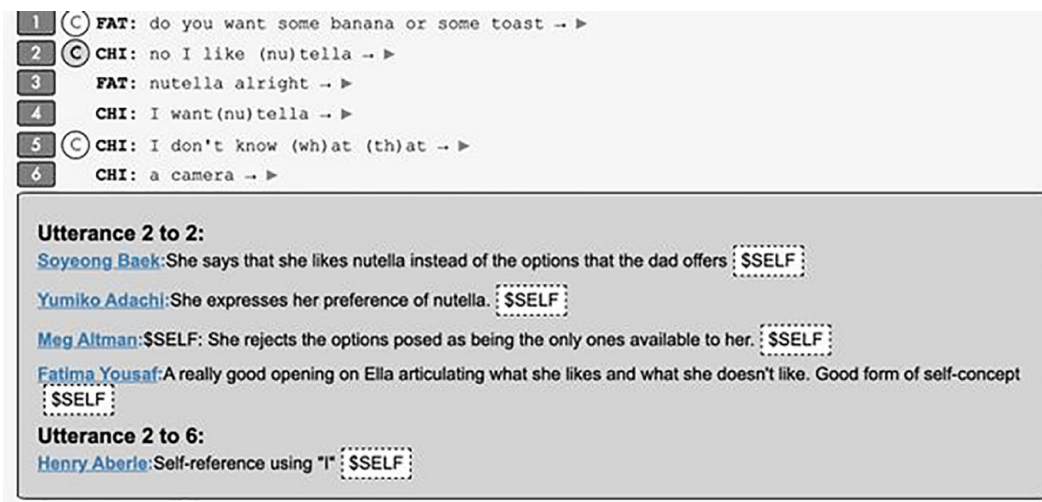
Fig. 2. A single utterance coded in Collaborative Commentary.

For example, the researchers participating in the Academically Productive Talk (APT) Project (https://class .talkbank.org/access/APT.html) are using CC to code methods for producing APT in classroom interactions (Al-Adeimi & O'Connor, 2021) across 51 videotaped classroom sessions in kindergarten through the 11th grade. The resultant coding decisions are then evaluated for reliability to determine strengths and weaknesses in the APT coding system and its underlying theoretical framework.

TalkBank Database

The TalkBank Database (TBDB) allows users to search for strings and patterns within or across all the transcripts in the 15 TalkBank databases. Using search patterns inserted into dialogue boxes, users can retrieve segments of text based on participants, utterances, tokens, types, sequences of types, sequences of tokens, or n-grams. Queries can focus on specific files, specific corpora, or collections of files and corpora. Participants can be selected in terms of their role (e.g., child, investigator, mother), age group, gender, clinical type, social status, or experimental condition. Searches can target specific words, combinations of words, codes, parts of speech, phonological structures, and grammatical dependency relations. The utterances and words that are matched can be downloaded to the user's local machine for further statistical analysis or further analyzed online. An alternative method for interacting with the TBDB is to use one of the application programming interfaces made available for R and Python.

Programs and Tutorials

The CLAN program is used to analyze TalkBank data on the desktop. It can be downloaded from https:// dali.talkbank.org, and the manual is available at https:// talkbank.org/manuals/CLAN.pdf. The best way to learn CLAN is to review the tutorial screencasts provided at https://talkbank.org/screencasts. These tutorials explain how to transcribe in CHAT and how to use CLAN programs, CC, the TBB, and the TBDB.

CLAN includes an editor with commands like those in Microsoft Word along with features for linking to media and for marking spoken-language features. It provides 24 commands for different types of corpus analysis, 10 commands that run packages for language-profiling analyses, 17 format-conversion commands, and 20 format-modification commands.

The CHAT format is also supported by the Phon program at https://phon.ca, which has been used to create the PhonBank database. Both Phon and Phon-Bank were built by Yvan Rose and Greg Hedlund at Memorial University Newfoundland. Phon provides extensive support for the detailed analysis of both child and adult phonological structures and development, including intelligent process automation coding, full integration of the Praat phonological analysis system (Boersma & Weenink, 2001), packages for standard phonological assessments, and methods for examining interactions between segments, syllables, words, and grammatical structures. A transcript created by the CLAN editor can be opened directly in Phon and vice versa.

Batchalign

Batchalign2 (Liu et al., 2023), which can be downloaded from https://github.com/talkbank/batchalign2, is a Python program with three basic functions designed to support the creation of TalkBank corpora:

1. Batchalign2 can use automatic speech recognition (ASR) to generate a CHAT transcript from either audio or video media. This function can either send audio over the Internet to the Rev-AI transcription service or use the OpenAI Whisper ASR model locally. For the languages that it covers, Rev-AI is a bit more accurate, does a partially accurate speaker identification, and does a better job tracking repetitions. However, Rev-AI costs $0.02 per minute, and some institutional review board rules do not permit sending data to cloud services, even though the option of not storing data in the Rev-AI cloud can be selected. For users who need to work locally or for users working with languages not covered by Rev-AI, Whisper is the better choice. The Batchalign2 installation includes a download of the Whisper model. We have found that both systems work well for two-party conversations recorded in good acoustic environments. Using a benchmarking program built into Batchalign, we have found a word error rate for Rev-AI that ranges from 1% for well-recorded TED speeches and two-party interviews to 20% for recordings in poorer audio conditions from people who stutter.

2. The second function of Batchalign2 is to automatically analyze CHAT transcripts for morphosyntactic structure. To do this, it relies on neural network models from the Stanza system (Qi et al., 2020) trained with data from 70 language treebanks in the Universal Dependencies (UD) project (De Marneffe et al., 2021). Liu and MacWhinney (2024) reported on the use of UD in Batchalign to analyze CHILDES data in 27 languages in terms of parts of speech, lexical features, and grammatical relations. For English, the output of this process is very accurate. An analysis by Liu and MacWhinney of Batchalign UD output for 7,170 words in the Sarah transcripts in the Brown (1973) corpus found only one error type impacting three words. The accuracy of tagging for other languages requires judgments from native-speaker linguists.

3. The third function of Batchalign2 is to automatically align a CHAT transcript with media. Prior to the creation of Batchalign2, the alignment of transcripts to media required the researcher to play through a transcript line by line while listening to the audio and hitting the space bar at the end of each utterance. Batchalign2 automates this process by rerunning ASR and using that output as a backplate for the probabilistic alignment of utterance groups or batches. This method works nearly perfectly if the utterances in the transcript are in the correct order. Moreover, the output includes beginning and end time marks for both utterances and individual words.

Applications

TalkBank data and programs have been used to address a wide range of issues in language learning, processing, disorders, and usage. Summarizing the work in more than 12,000 published articles is beyond the scope of this article, but we can mention a few illustrative projects from some of the 15 components of TalkBank:

- Computer scientists have used data from DementiaBank in three Interspeech challenges to predict the onset of mild cognitive impairment (MCI) on the basis of analyses of acoustic, lexical, and grammatical aspects of recordings and transcripts of picture descriptions (Luz et al., 2021). This work mostly uses the Pitt corpus, which contains a sentence-repetition task and descriptions of the Cookie Theft picture from 270 participants with MCI and 140 control participants. The core task is to classify samples as either MCI/Alzheimer's disease (AD) or control. The systems used to address this task include a variety of machine learning (ML) and artificial intelligence (AI) methods. Earlier approaches to AD recognition through speech used different, often unbalanced and acoustically varied data sets, consequently hindering the reproducibility and comparability of approaches. The three Interspeech challenges have provided a forum for scores of academic and commercial research groups to test their existing methods or to develop novel approaches on a new shared standardized data set, resulting in more than 100 published articles.
- In parallel with this extensive work on AD detection, there have been 87 published articles applying ML and AI methods to the task of improving ASR for the speech of people with language disorders. This work has been based on training data from AphasiaBank, DementiaBank, and FluencyBank (for stuttering), as well as data in CHILDES from children speaking nonmainstream dialects.
- Using videos from AphasiaBank, researchers have shown how gesture substitutes for speech, particularly in nonfluent Broca's aphasia (van Nispen et al., 2017). In these videos, participants with

aphasia describe the story of Cinderella after having reviewed a wordless picture-book version. Because nonfluent aphasia is often accompanied by word-finding difficulties, participants make use of gestures as substitutes for the words they cannot find, thereby showing that they know the object or action, even if they cannot find the word.

- During the first stages of language learning, children fail to mark tense on the finite verb (the main verb of a simple clause). Wexler (1994) proposed that this failure is due to a delayed maturation of an innate universal constraint that requires this marking. However, Freudenthal et al. (2007) showed that the level of use of nonfinite (infinitive or infinitival) forms across four languages in the CHILDES database aligned closely with language variation in the input rather than with a universal maturational constraint.

- Using ClassBank videos from a 14-day program of lessons in a second-grade classroom, Strom et al. (2001) were able create a dynamic flow chart of developing understandings of area and symmetry in geometric patterns. To capture this advancing understanding, they created three concept graphs to illustrate the children's successive formulations of 13 procedures for measuring area, five levels of congruence tests, and four principles derived from earlier work on symmetry and area in quilt design, along with the relations between these knowledge components. Each of these 22 knowledge components was then linked to specific utterances in the ClassBank video transcript. Taken together, this analysis shows how classroom instruction with specific geometric examples can produce a richer network of geometric understandings and how this development can be studied developmentally.

- Anderson and Schooler (1991) used data from a subsection of CHILDES to test the claim that "the probability that a memory will be needed shows reliable relationships to frequency, recency, and pattern of prior exposures," (p. 396) as evidenced in parental input to language-learning children. They modeled this relation in terms of the retention function, the practice function, and the spacing effect.

- Using transcript data from CHILDES, Karniol (2010) developed a book-length analysis of social development as preference management, which she rephrased as "how infants, children, and parents get what they want from one another." She examined issues such as temporizing preferences, disciplining noncompliance, coping and self-regulating, and the parental channeling of child preferences. The development of each of these aspects of social development, as well as several

others, is illustrated through specific verbal interactions from CHILDES and similar sources.

- Eppler et al. (2017) used data from code-switched interactions in the Eppler corpus in BilingBank to test the predictions of three syntactic theories—the minimalist program, word grammar, and the matrix language frame model—in terms of their ability to predict grammatical feature agreement between determiners and nouns. Examples of combinations in which features do not agree between the languages stand as failures for each of the theories. For example, the phrase "alle bus" (all buses) instead of "alle buses" has a plural German determiner with a singular English noun, and this combination is not predicted by these models.

- Using corpora from 56 university learners majoring in French in SLABank, McManus et al. (2021) followed the progression of the language skills of complexity, accuracy, lexicon, and fluency across 3 years to assess patterns of second-language learning, maintenance, and attrition. Results showed ongoing improvements on most measures, including accuracy. There were long-term relationships between fluency and vocabulary, but relationships involving accuracy and complexity emerged only in instructed contexts.

- Using data from daylong recordings in the home in the Cougar corpus in HomeBank, Kondaurova et al. (2023) found that mothers increased their vocal pitch when speaking with toddlers as opposed to when they were speaking with other adults. This effect has been interpreted as a way for mothers to provide emotional support for their children. However, fathers did not show the same increase in pitch when speaking with toddlers. The absence of this pitch adjustment in fathers supports the view formulated in Gleason's (1975) bridge hypothesis, according to which fathers provide young children with challenges that allow them to interact with an increasingly wider range of adults.

- LaSalle and Wolk (2023) used transcripts from 43 children across five corpora in FluencyBank to examine the effects of parental responses to utterances in which the child stutters. If the parent's response involved recasting, then the following child utterance contained significantly less further stuttering than if the parent's response did not involve recasting. This was interpreted as showing that parental recasting served to facilitate children's fluency, perhaps by providing clear lexical and grammatical forms that children could use in further productions. However, this difference between recasting and nonrecasting responses did not obtain for nine of the 43 children who had persistent stuttering. The lack of a

positive effect for this group could be due to comorbidity with phonological disorders or perhaps their heightened emotional reaction against recasting and correction.

Future Directions

The construction and dissemination of TalkBank data and methods have benefited from the generous contributions of data from hundreds of researchers who have recorded interactions and transcribed them in CHAT format, ongoing support from the National Institutes of Health and the National Science Foundation, the usage of TalkBank data in more than 12,000 published articles, and the participation of more than 10,000 registered users of TalkBank data. Going forward, we are particularly interested in furthering five efforts:

1. We want to make maximal use of the many new tools offered by new developments in AI and ML to improve language data collection, transcription, and analysis.
2. We want to extend the cross-linguistic coverage of TalkBank. The CHILDES database has data from 47 different languages, CABank has data from six languages, SLABank has data from CABank, and SLABank has data on the learning of nine languages. However, the other databases have data primarily from English, and we need to extend their cross-linguistic coverage.
3. A great strength of the CHILDES database is the longitudinal nature of many of the corpora that track language development in children from birth to the age of 5 years. We need data with a similar longitudinal design for the study of areas such as dementia, recovery from strokes in aphasia, and second-language learning.
4. Many current analyses examine single dimensions of language learning and production, such as lexicon, grammar, conversation, phonology, accuracy, fluency, or complexity. We are developing new methods in the TBDB that will allow us to track patterns and interactions across all of these systems.
5. In the first 10 months of its availability, instructors, clinicians, and researchers have generated 12,652 comments using CC. We want to work with specific projects and classes to develop and test new methods of using the commentary database to test hypotheses and coding systems.

Conclusion

TalkBank has made major empirical contributions to research in developmental psychology, cognitive psychology, psycholinguistics, linguistics, conversation analysis, second-language acquisition, education, clinical linguistics, speech and hearing, natural-language processing, and computer science. These contributions arise from the commitment to the principles of data sharing, open science, replicability, and the use of a consistent data format. The success of these methods can serve as a guide to other data-sharing projects in the behavioral sciences.

Recommended Reading

Goldstone, R., & Lupyan, G. (2016). (See References). Explains how searches and comparisons across digital databases can be used to study psychological patterns.

Liu, H., & MacWhinney, B. (2024). (See References). Describes a system for automatic creation and analysis of spoken language interactions.

MacWhinney, B., & Fromm, D. (2023). Collaborative commentary for understanding communication disorders. *American Journal of Speech-Language Pathology*, *32*(5S), 2580–2588. Explains how to use a system for microanalysis, coding, and commentary of spoken language multimedia materials accessible through a web browser.

Transparency

ORCID iD

Brian MacWhinney https://orcid.org/0000-0002-4988-1342

References

Al-Adeimi, S., & O'Connor, C. (2021). Exploring the relationship between dialogic teacher talk and students' persuasive writing. *Learning and Instruction*, *71*, Article 101388. https://doi.org/10.1016/j.learninstruc.2020.101388

Anderson, J., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Boersma, P., & Weenink, D. (2001). *Praat: Doing phonetics by computer*. https://www.praat.org

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, *47*(2), 255–308.

Eppler, E. D., Luescher, A., & Deuchar, M. (2017). Evaluating the predictions of three syntactic frameworks for mixed determiner–noun constructions. *Corpus Linguistics and Linguistic Theory*, *13*(1), 27–63.

Freudenthal, D., Pine, J. M., Aguado-Orea, J., & Gobet, F. (2007). Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science*, *31*(2), 311–341.

Gleason, J. B. (1975). Fathers and other strangers: Men's speech to young children. *Developmental Psycholinguistics: Theory and Applications*, *1*, 289–297.

Goldstone, R., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, *8*, 548–568. https://doi.org/10.1111/tops.12212

Karniol, R. (2010). *Social development as preference management: How infants, children, and parents get what they want from one another.* Cambridge University Press.

Kondaurova, M. V., VanDam, M., Zheng, Q., & Welikson, B. (2023). Fathers' unmodulated prosody in child-directed speech. *The Journal of the Acoustical Society of America*, *154*(6), 3556–3567.

LaSalle, L., & Wolk, L. (2023). Adult recasts as fluency-facilitators in preschoolers who stutter: Evidence from FluencyBank. *Journal of Fluency Disorders*, *76*, Article 105971. https://doi.org/10.1016/j.jfludis.2023.105971

Liu, H., & MacWhinney, B. (2024). Morphosyntactic analysis for CHILDES. *Language Development Research*, *4*(1), 233–258. https://doi.org/10.34842/j97r-n823

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research*, *66*, 2421–2433. https://doi.org/10.1044/2023_JSLHR-22-00642

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., & MacWhinney, B. (2021). Editorial: Alzheimer's dementia recognition through spontaneous speech. *Frontiers in Aging Neuroscience*, *3*, Article 780169. https://doi.org/10.3389/fcomp.2021.780169

McManus, K., Mitchell, R., & Tracy-Ventura, N. (2021). A longitudinal study of advanced learners' linguistic development before, during, and after study abroad. *Applied Linguistics*, *42*(1), 136–163.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System demonstrations* (pp. 101–108). Association for Computational Linguistics.

Strom, D., Kemeny, V., Lehrer, R., & Forman, E. (2001). Visualizing the emergent structure of children's mathematical argument. *Cognitive Science*, *25*(5), 733–773.

van Nispen, K., van de Sandt-Koenderman, M., Sekine, K., Krahmer, E., & Rose, M. L. (2017). Part of the message comes in gesture: How people with aphasia convey information in different gesture types as compared with information in their speech. *Aphasiology*, *31*(9), 1078–1103. https://doi.org/10.1080/02687038.2017.1301368

Wexler, K. (1994). Optional infinitives, head movement and the economy of derivations. In N. Hornstein & D. Lightfoot (Eds.), *Verb movement* (pp. 305–350). Cambridge University Press.