JSLHR

Research Article

A Comparison of Manual Versus Automated Quantitative Production Analysis of Connected Speech

Davida Fromm,[a] Saketh Katta,[b] Mason Paccione,[c] Sophia Hecht,[c] Joel Greenhouse,[c] Brian MacWhinney,[a] and Tatiana T. Schnur[b,d]

Purpose: Analysis of connected speech in the field of adult neurogenic communication disorders is essential for research and clinical purposes, yet time and expertise are often cited as limiting factors. The purpose of this project was to create and evaluate an automated program to score and compute the measures from the Quantitative Production Analysis (QPA), an objective and systematic approach for measuring morphological and structural features of connected speech.
Method: The QPA was used to analyze transcripts of Cinderella stories from 109 individuals with acute–subacute left hemisphere stroke. Regression slopes and residuals were used to compare the results of manual scoring and automated scoring using the newly developed C-QPA command in CLAN, a set of programs for automatic analysis of language samples.

Results: The C-QPA command produced two spreadsheet outputs: an analysis spreadsheet with scores for each utterance in the language sample, and a summary spreadsheet with 18 score totals from the analysis spreadsheet and an additional 15 measures derived from those totals. Linear regression analysis revealed that 32 of the 33 measures had good agreement; *auxiliary complexity index* was the one score that did not have good agreement.
Conclusions: The C-QPA command can be used to perform automated analyses of language transcripts, saving time and training and providing reliable and valid quantification of connected speech. Transcribing in CHAT, the CLAN editor, also streamlined the process of transcript preparation for QPA and allowed for precise linking of media files to language transcripts for temporal analyses.

The challenge of quantifying deficits from connected speech is one that researchers and clinicians continue to address, as connected speech is the most ecologically valid form of language output to measure. The challenge is multifaceted, beginning with the time and expertise necessary to transcribe a language sample and then the time to analyze it accurately and consistently. Those factors are often limiting, as confirmed by results of several recent articles on the topic of linguistic discourse analysis (Bryant et al., 2016, 2017; Cruice et al., 2020; Dietz & Boyle, 2018b). Here, we create and evaluate the automation of the Quantitative Production Analysis (QPA; Rochon et al., 2000; Saffran et al., 1989), an objective and systematic approach for measuring morphological and structural features of connected speech. The QPA is a validated and widely used approach for quantifying connected speech deficits (Bryant et al., 2016). We found that the automated QPA compares favorably to scores tabulated by human experts and thus provides a reliable and valid quantification of connected speech while requiring less time and linguistic expertise.

The QPA is a comprehensive and established approach for measuring the lexical content and sentence structure of connected speech. This tool has been used broadly by researchers to study aphasia and other types of acquired neurological disorders of language, as evidenced by a search on Google Scholar for the Rochon et al. (2000) and Saffran et al. (1989) articles, which indicated a joint total of 853 citations (see also Bryant et al., 2016). The QPA provides a reliable and valid set of measures to quantify relevant aspects of connected speech in the field of adult language disorders (cf. Gordon, 2006; Rochon et al., 2000; Saffran et al., 1989). Specifically, it focuses on the frequency of occurrence of certain grammatical features (e.g., nouns, verbs, determiners,

[a]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA
[b]Department of Neurosurgery, Baylor College of Medicine, Houston, TX
[c]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA
[d]Department of Neuroscience, Baylor College of Medicine, Houston, TX

Correspondence to Davida Fromm: fromm@andrew.cmu.edu

Editor-in-Chief: Stephen M. Camarata
Editor: Julius Fridriksson

Received September 22, 2020
Revision received November 19, 2020
Accepted December 14, 2020
https://doi.org/10.1044/2020_JSLHR-20-00561

Disclosure: The authors have declared that no competing interests existed at the time of publication.

embeddings) and the ways in which utterances are elaborated beyond the basic noun plus verb combination.

The QPA manual (Berndt et al., 2000) includes rules for transcription, extraction of narrative words, segmentation of words into utterances, and scoring for utterances, words, and structural measures. Typically, two or more speech-language pathologists or graduate/undergraduate research assistants with a background in linguistics and/or speech-language pathology are trained in transcription, utterance segmentation, and QPA scoring following the guidelines in the training program outlined in the QPA manual. Training requires meeting interrater reliabilities at each stage generally in excess of 90% (Mirman et al., 2019; Martin & Schnur, 2019; for a detailed approach to assess reliability, see Gordon, 2006). Twenty measures are computed by hand in an utterance-by-utterance analysis worksheet, typically in Excel. Those numbers are tallied and then entered into a summary worksheet that computes additional overall measures of lexical, morphological, and structural aspects of the sample (e.g., *proportion of verbs, proportion of well-formed sentences, mean subject noun phrase [SNP] length*). The time and expertise required for training and using the QPA reliably are substantial, thereby limiting its more widespread use.

The Northwestern Narrative Language Analysis (NNLA) system (Thompson et al., 1995) is another objective and systematic method for detailing connected speech. Like the QPA, the NNLA was originally developed to examine the language deficits in agrammatic Broca's aphasia. In an attempt to take advantage of automated linguistic analysis approaches, Hsu and Thompson (2018) showed how automated coding using Computerized Language ANalysis (CLAN; MacWhinney, 2000) was largely consistent with manual NNLA coding using the NNLA system, concluding that the best approach is a combination of automatic coding and manual coding. That led to the development of C-NNLA, a new CLAN program that allows for the automatic computation of 50 NNLA measures (Fromm et al., 2020). Though some of the NNLA measures are similar to QPA measures (e.g., number of nouns, verbs, pronouns, open and closed class words, embeddings), each system has unique measures and definitions for categories and scoring. In comparison to the QPA, the NNLA involves five levels of hand coding for every utterance, with many more sentence-level codes (e.g., active canonical, passive), lexical codes (preposition, conjunction, adjective, adverb), bound morpheme codes (irregular plural, possessive, superlative), and verb argument codes (obligatory one place, optional two place). However, the QPA has been the most frequently used system for quantifying connected speech (cf. Bryant et al., 2016) in part because it provides a general approach for identification of disordered connected speech useful for research questions which do not require deep syntactic analysis (e.g., identification of sentences produced with or without syntactic movement or utterances with or without verbs entailing one vs. two arguments; e.g., Ding et al., 2020). To date, the QPA remains a manual analysis. Here, we follow the general approach adopted by Fromm et al. to pursue a similar effort for automatically computing QPA measures using a single CLAN command.

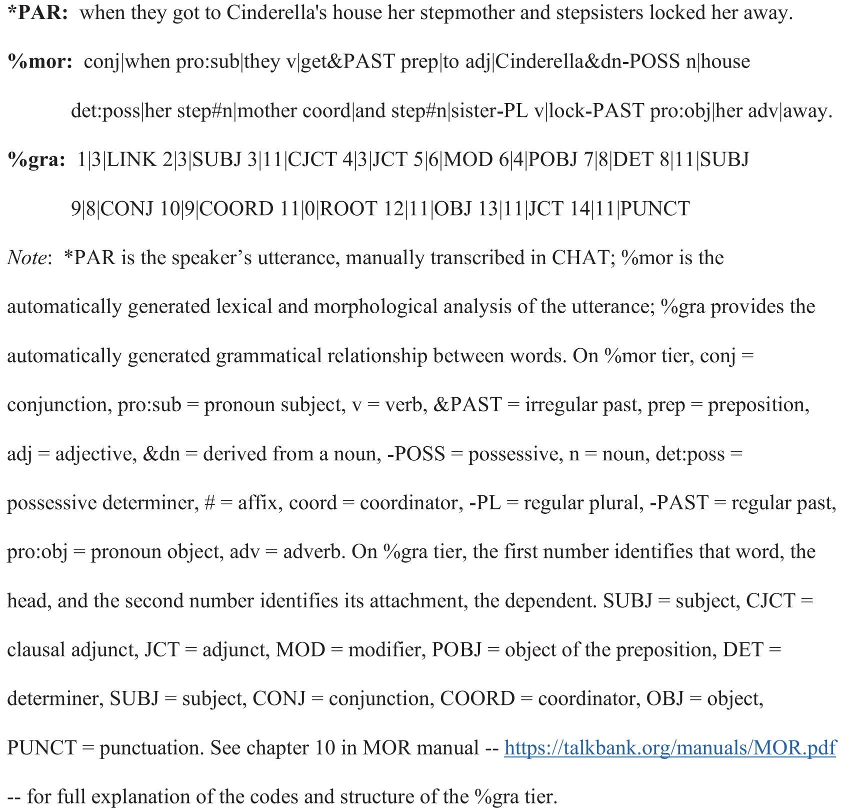
CLAN is a freely downloadable set of programs for automatic analysis of language samples that have been transcribed in the CLAN editor (CHAT). The CHAT transcription system provides a standardized format for capturing spoken communication and is described in detail in an online manual (https://talkbank.org/manuals/CHAT.pdf). CHAT transcripts use normal English orthography as the input (e.g., there is no need to hand-code morphology) and can be linked to the media file for easy replay and accurate transcription of individual utterances. There are also ways to mark repetitions, revisions, fillers, sound fragments, and other types of behaviors encountered in language samples.[1] Careful transcription takes time, and the quality of the transcript affects the quality of the results. Estimates of how long it takes to transcribe a language sample depend on several variables such as the type of sample, the severity of the speaker's impairment, the amount of coding required within the transcript for subsequent analyses of interest (e.g., morphological error coding, paraphasia error coding, pause time, gestures, correct information units), and the method used for transcribing (e.g., Microsoft Word, CLAN). Thus, estimates vary widely in the literature from 6–10 min (Boles, 1998) to "up to an hour" (L. Armstrong et al., 2007) per minute. Transcribing in CLAN is more efficient in many ways than transcribing in Word, Excel, or Systematic Analysis of Language Transcripts (Miller & Iglesias, 2015), but learning a new system for transcription and analysis requires some investment of time.[2] It is important to note that the work of creating a transcript is an essential part of the process, whether one conducts a hand analysis in QPA or an automatic analysis.

Once a language sample is transcribed (see *PAR in Figure 1 for example), the CLAN command, MOR, is used to perform automatic lexical and morphological tagging based on the lexicon for the language and a trained statistical disambiguator for words that have more than one part of speech (e.g., *mean* can be an adjective, verb, or noun). This tagging appears as a %mor tier immediately below each speaker utterance. MOR also creates output that describes the structure of the sentence in terms of pairwise grammatical relations between words. This appears as a %gra tier immediately below the %mor tier. Figure 1 shows examples of these automatically generated tiers (lexical and morphological, grammatical relations). The morphological tagging accuracy of CLAN has consistently been between 95% and 97% (MacWhinney, 2011). Recently, that accuracy has increased to greater than 99%, after additional training and the addition of two files—one that runs before and one that runs after the MOR command to improve

[1] For specifics about marking discourse content within transcripts, see the online CHAT manual (Sections 8–10, https://talkbank.org/manuals/CLAN.pdf) or the more abbreviated SLP's Guide to CLAN (pp. 13–21, https://talkbank.org/manuals/Clin-CLAN.pdf).

[2] Tutorial screencasts are available to help new users learn how to transcribe and analyze transcripts—https://talkbank.org/screencasts/

Figure 1. Example of CHAT transcript with MOR command output.



***PAR:** when they got to Cinderella's house her stepmother and stepsisters locked her away.

%mor: conj|when pro:sub|they v|get&PAST prep|to adj|Cinderella&dn-POSS n|house
det:poss|her step#n|mother coord|and step#n|sister-PL v|lock-PAST pro:obj|her adv|away.

%gra: 1|3|LINK 2|3|SUBJ 3|11|CJCT 4|3|JCT 5|6|MOD 6|4|POBJ 7|8|DET 8|11|SUBJ
9|8|CONJ 10|9|COORD 11|0|ROOT 12|11|OBJ 13|11|JCT 14|11|PUNCT

Note: *PAR is the speaker's utterance, manually transcribed in CHAT; %mor is the automatically generated lexical and morphological analysis of the utterance; %gra provides the automatically generated grammatical relationship between words. On %mor tier, conj = conjunction, pro:sub = pronoun subject, v = verb, &PAST = irregular past, prep = preposition, adj = adjective, &dn = derived from a noun, -POSS = possessive, n = noun, det:poss = possessive determiner, # = affix, coord = coordinator, -PL = regular plural, -PAST = regular past, pro:obj = pronoun object, adv = adverb. On %gra tier, the first number identifies that word, the head, and the second number identifies its attachment, the dependent. SUBJ = subject, CJCT = clausal adjunct, JCT = adjunct, MOD = modifier, POBJ = object of the preposition, DET = determiner, SUBJ = subject, CONJ = conjunction, COORD = coordinator, OBJ = object, PUNCT = punctuation. See chapter 10 in MOR manual -- https://talkbank.org/manuals/MOR.pdf -- for full explanation of the codes and structure of the %gra tier.

CLAN's disambiguation (by filtering possibilities based on the syntactic environment). With a better morphological tier, the grammatical relation's tier accuracy improved to 94%. Given this high level of accuracy, we were able to use the information derived from the MOR analysis command to formulate rules for automated computation of QPA measures.

The goal of the article was to create an automated version of the QPA, assess the degree to which it produced scores similar to those produced by trained manual scorers, and, where disagreements occurred, to understand why.

Method

Participants

One hundred and nine acute left hemisphere stroke patients (65 males, 44 females) were recruited, independent of a clinical diagnosis of aphasia, from three comprehensive stroke centers in the Texas Medical Center in Houston, Texas, as part of an ongoing project. Participants included were native English speakers, diagnosed with an acute ischemic or parenchymal hemorrhagic left hemisphere stroke who were able to produce an intelligible Cinderella story within an average of 4 days after stroke onset (range: 1–17 days). At the time of testing, participants were on average 60.5 years of age (range: 20–85 years) with an average 13.8 years of education (range: 6–33 years). Informed consent was approved by the Baylor College of Medicine Institutional Review Board.

Language Sample Transcription

The rules for CLAN's computations of QPA measures were written to comply with the rules as written in the QPA manual. The new C-QPA program was designed to produce two outputs: an utterance-by-utterance spreadsheet (similar to the QPA worksheet) and a summary spreadsheet (similar to the QPA summary sheet) with outcome measures for lexical content, auxiliary analyses, and structural analysis. When scored manually, totals for each measure (e.g., *number of open class words, number of words in verb phrases*) were tallied from the analysis worksheet and transferred to the summary sheet. These are the nonderived measures, which are used to compute the derived measures (see Table 1). The derived measures most often involve division to compute means (e.g., *mean sentence length* = *number of words in sentences* divided by the *number of sentences*) or proportions (e.g., *proportion of pronouns* = *number of pronouns* divided by *number of nouns* plus *number of pronouns*). Sometimes a derived measure is a simple subtraction, as in *number of closed class words,* which is the total *number of narrative words* minus the *number of open class words*. Comparisons

Table 1. Quantitative Production Analysis outcome measures.

Production analysis	Nonderived measures	Derived measures
Lexical content	# narrative words # open class words # nouns # nouns requiring determiners (NRDs) # NRDs with determiners # pronouns # verbs	# closed class words proportion closed class words determiner (DET) index proportion pronouns proportion verbs
Auxiliary analysis	# matrix verbs total aux score	aux complexity index
Structural analysis	# sentences # words in sentences # well-formed sentences # subject noun phrases (SNPs) # verb phrases (VPs) # words in SNPs # words in VPs # embeddings # utterances	proportion words in sentences proportion well-formed sentences mean SNP length SNP elaboration index mean VP length VP elaboration index sentence elaboration index embedding index mean utterance length[a]

[a]This is mean utterance length (not median utterance length, as given in the Quantitative Production Analysis manual) as this is what was computed in the manually scored data.

between manually generated and automated results were made based on the outcome measures in the summary spreadsheet, as those measures were all transferred from or computed based on the numbers in the utterance-by-utterance analysis worksheet.

Comparison Between Manual and Automated QPA Output

We conducted a large-scale comparison between the automated C-QPA output and manually generated QPA output analyzed by experienced QPA scorers.[3] The narrative production task as well as the manual transcription, QPA scoring, and reliability procedures used in this data set are described in Martin and Schnur (2019). Utterances from the manually prepared QPA worksheets were copied into text files that were then converted to CHAT files using the TEXT2CHAT command. The CHAT files were then checked and edited to ensure that the morphological parsing and resulting C-QPA would be as accurate as possible. Figure 2 outlines the procedure to prepare transcripts for analysis. It should be noted that these utterances, taken from the manually coded QPA spreadsheets and converted into CHAT files, already had repetitions, revisions, and fillers removed as well as any words or utterances that should be excluded from the QPA (e.g., habitually used starters, comments made by the participant).

Once the CHAT transcripts were completed, we ran the MOR command, which automatically created the morphological and grammatical relations lines in all the CHAT files simultaneously (a < 10-s procedure). Then, we ran the C-QPA command, as written below, to perform the QPA on the participant's utterances in all of the CHAT files:

```
c-qpa +t*txt *.cha
```

This also takes just a few seconds to complete, producing one large spreadsheet with the summary scores (columns) for all of the transcripts (rows) and 109 individual analysis spreadsheets with scores (columns) and utterances (rows) for each CHAT file. Figure 3 shows portions of the analysis and summary spreadsheets, respectively. The analysis spreadsheet (top) is a screenshot of the first few rows and columns of one participant's analysis spreadsheet, showing the utterances in Column A and the QPA measures (*sentence utterance, other utterance, # narrative words, # open class words*, etc.) across the top row. The summary spreadsheet (bottom) shows a screenshot of the first few rows and columns of all participants' transcript analysis results, with the participant's ID in Column A and the QPA summary measures *(# narrative words, # open class words, # closed class words, proportion closed class words, nouns,* etc.) across the top row.

Statistical Analysis

Statistical analyses were performed using R, Version 3.6.3. Each automated and manual score (nonderived and derived) was standardized by subtracting its respective mean and dividing by its respective standard deviation. For each QPA measure, we first generated a scatterplot of the standardized automated scores versus the standardized manual scores to help visualize the degree of linear agreement between the two score types and investigate deviations from linearity. We then formally estimated the slope of the regression of the standardized automated scores on the standardized scores from the manual raters. Specifically, a simple linear regression was fit to each QPA variable V of

[3]Specifics about the C-QPA command are available in Section 8 of the online CLAN manual—https://talkbank.org/manuals/CLAN.pdf

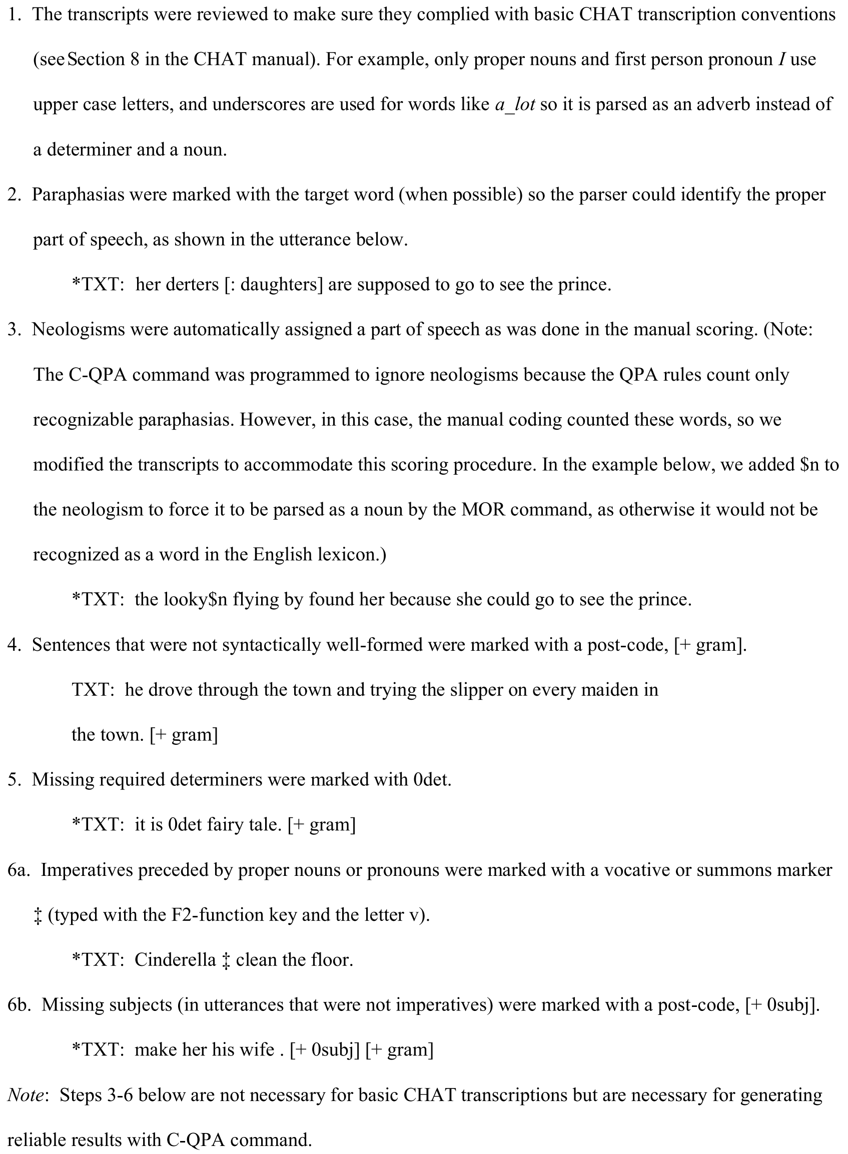


Figure 2. Summarized procedure for transcript preparation.

1. The transcripts were reviewed to make sure they complied with basic CHAT transcription conventions (see Section 8 in the CHAT manual). For example, only proper nouns and first person pronoun *I* use upper case letters, and underscores are used for words like *a_lot* so it is parsed as an adverb instead of a determiner and a noun.
2. Paraphasias were marked with the target word (when possible) so the parser could identify the proper part of speech, as shown in the utterance below.

   *TXT: her derters [: daughters] are supposed to go to see the prince.
3. Neologisms were automatically assigned a part of speech as was done in the manual scoring. (Note: The C-QPA command was programmed to ignore neologisms because the QPA rules count only recognizable paraphasias. However, in this case, the manual coding counted these words, so we modified the transcripts to accommodate this scoring procedure. In the example below, we added $n to the neologism to force it to be parsed as a noun by the MOR command, as otherwise it would not be recognized as a word in the English lexicon.)

   *TXT: the looky$n flying by found her because she could go to see the prince.
4. Sentences that were not syntactically well-formed were marked with a post-code, [+ gram].

   TXT: he drove through the town and trying the slipper on every maiden in the town. [+ gram]
5. Missing required determiners were marked with 0det.

   *TXT: it is 0det fairy tale. [+ gram]

6a. Imperatives preceded by proper nouns or pronouns were marked with a vocative or summons marker ‡ (typed with the F2-function key and the letter v).

   *TXT: Cinderella ‡ clean the floor.

6b. Missing subjects (in utterances that were not imperatives) were marked with a post-code, [+ 0subj].

   *TXT: make her his wife . [+ 0subj] [+ gram]

Note: Steps 3-6 below are not necessary for basic CHAT transcriptions but are necessary for generating reliable results with C-QPA command.

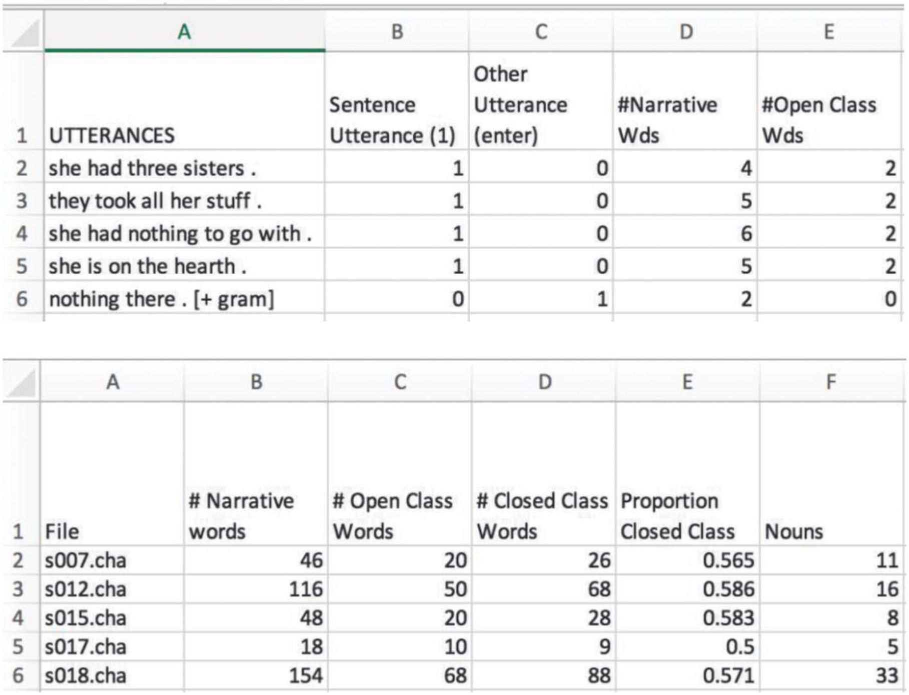
the form: $V_{CLAN} = \beta_V \times V_{Manual} + \varepsilon$. Note the lack of an intercept term in the model as we estimated the hypothesized line of agreement through the origin. When we fit models with intercepts, they resulted in estimated intercepts that were not statistically different from zero for each variable. The adequacy of the fit of all the regression models was assessed using standard regression diagnostics for least squares regression.

We used the regression slopes to assess the degree to which manual and automated QPA outputs across QPA variables agreed. A 95% confidence interval was calculated for each slope to assess whether the value of 1, corresponding to the line of perfect agreement, was contained in the interval. An estimated slope of approximately 1 and/or a relatively tight confidence interval around the estimated slope was considered good agreement between manual scoring and C-QPA.

Results

We estimated dissimilarity between manual and automated QPA output by examining the derived and nonderived variables separately, because derived scores are often ratios of nonderived scores, thereby introducing more variability. Figure 4 provides a visualization via scatterplot of the degree of agreement between manually generated and automated QPA results. Across nearly all nonderived QPA variables shown in Figure 4, there was good agreement between manual and automated CLAN scores. Figure 5 shows scatterplots for the derived QPA variables. Again,

Figure 3. Portions of the C-QPA spreadsheet for one individual's transcript (top) and summary spreadsheet (bottom) across transcripts.



	A	B	C	D	E
1	UTTERANCES	Sentence Utterance (1)	Other Utterance (enter)	#Narrative Wds	#Open Class Wds
2	she had three sisters .	1	0	4	2
3	they took all her stuff .	1	0	5	2
4	she had nothing to go with .	1	0	6	2
5	she is on the hearth .	1	0	5	2
6	nothing there . [+ gram]	0	1	2	0

	A	B	C	D	E	F
1	File	# Narrative words	# Open Class Words	# Closed Class Words	Proportion Closed Class	Nouns
2	s007.cha	46	20	26	0.565	11
3	s012.cha	116	50	68	0.586	16
4	s015.cha	48	20	28	0.583	8
5	s017.cha	18	10	9	0.5	5
6	s018.cha	154	68	88	0.571	33

Figure 4. Joint distributions of Quantitative Production Analysis scores: nonderived variables. NRDs = nouns requiring determiner; SNPs = subject noun phrases; VPs = verb phrases.

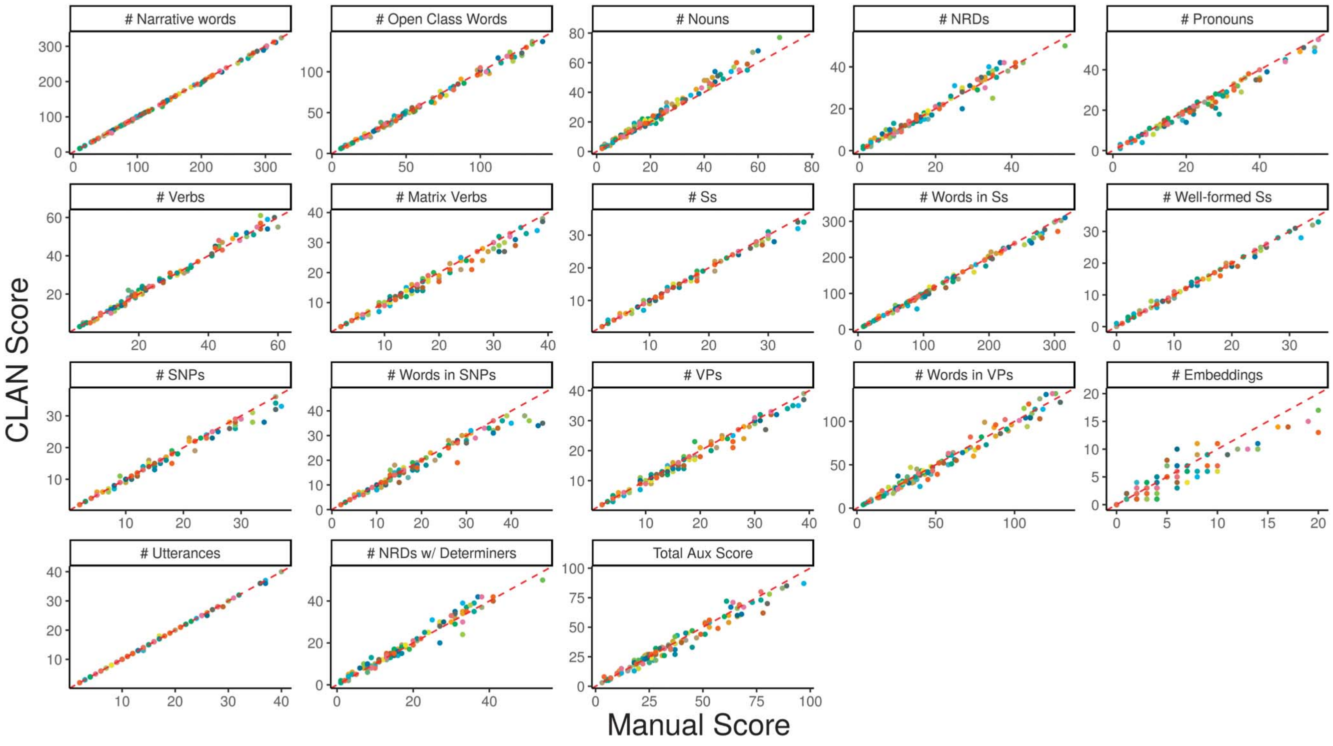
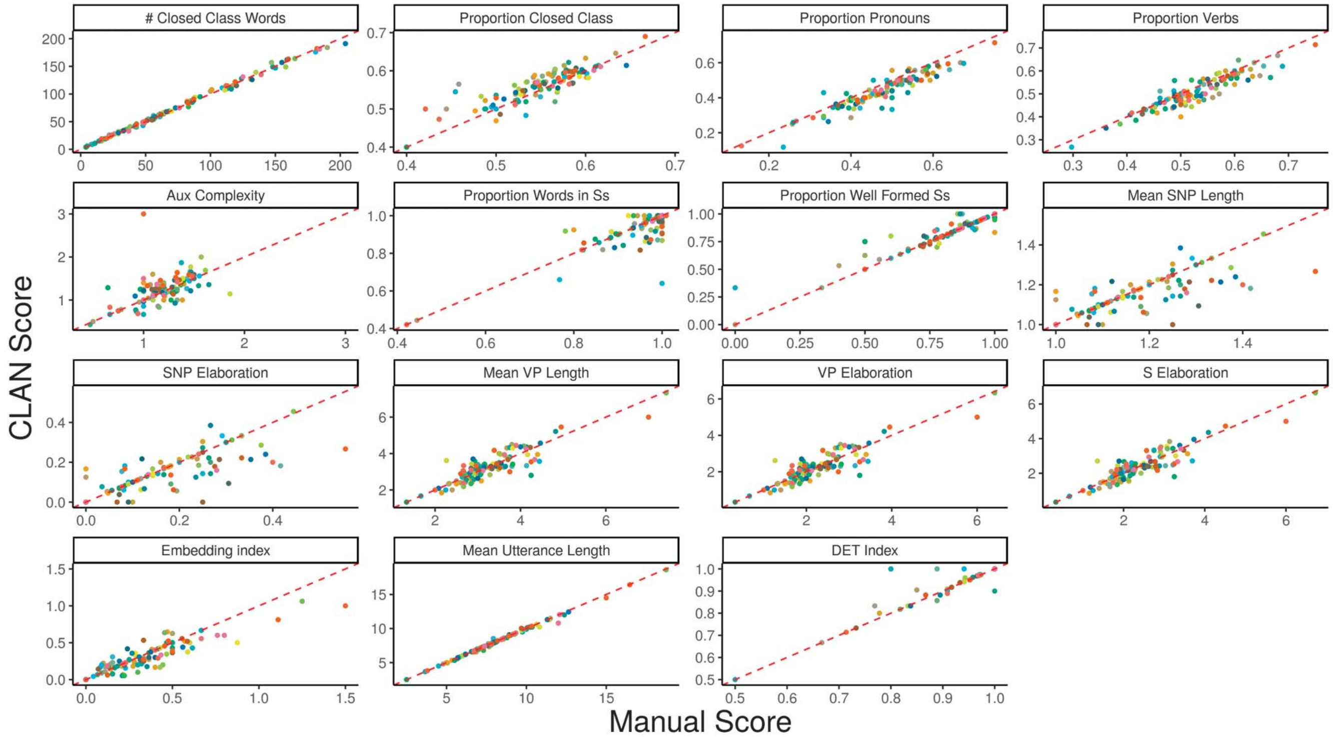


Figure 5. Joint distributions of Quantitative Production Analysis scores: derived variables. SNP = subject noun phrase; VP = verb phrase; DET = determiner.



agreement between manual and automated scores is good but with a little more variability around the line with the slope 1. As can be seen by inspection of the scatterplots in both Figures 4 and 5, the model fits were very good. Formal regression diagnostics did not provide evidence for significant deviations from the usual linear regression model assumptions.

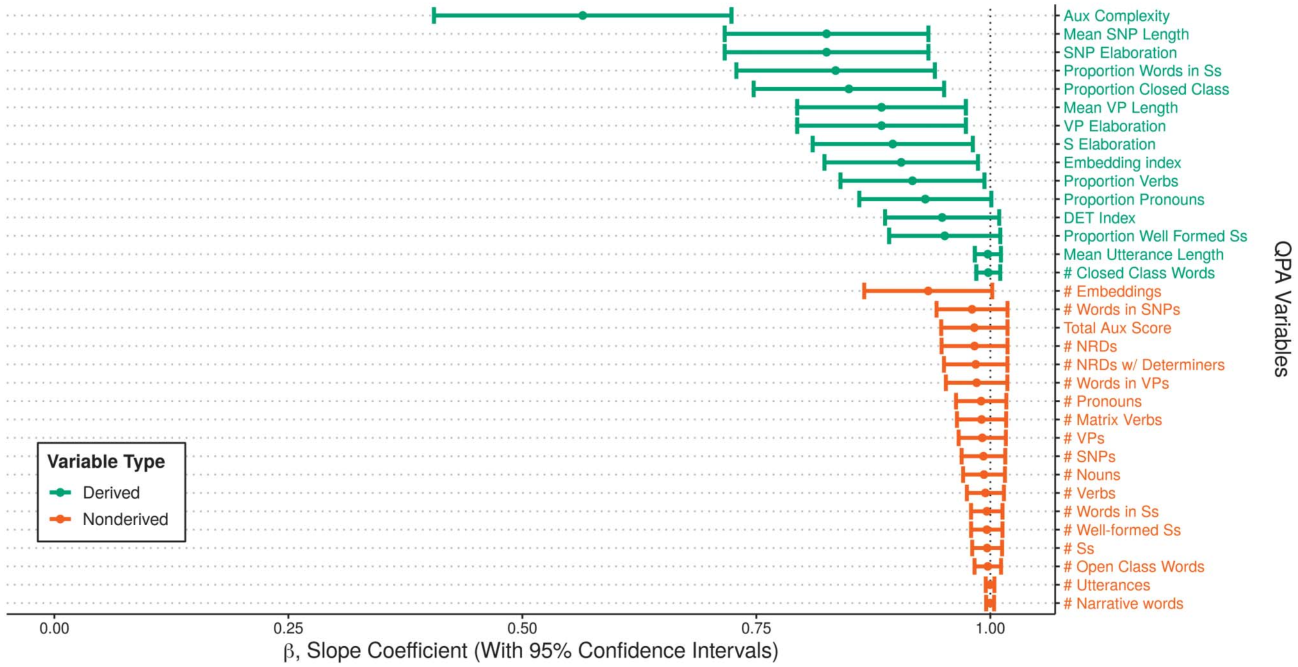
Figure 6 displays the 95% confidence intervals for the slope coefficients for the linear regressions for all QPA measures presented in Figures 4 and 5. The point estimates for the slopes for all but two of the nonderived measures are less than 1, indicating a tendency for the automated CLAN scores to be consistently lower than the manual scores. Nevertheless, for the nonderived measures, all confidence intervals for the slope coefficient for agreement contained 1, indicating statistically strong agreement between automated and manual scoring. Of the 15 derived scores, the point estimates for all but two again showed systematic underestimations of automated scores relative to manual scores. Five of the 15 derived scores that had confidence intervals including 1 (*number of closed class words, mean utterance length, proportion well-formed sentences, determiner index, proportion pronouns*) and 9 (*proportion verbs, embedding index, sentence elaboration, mean verb phrase length, verb phrase elaboration, proportion closed class words, proportion words in sentences, mean SNP length, SNP elaboration*) had confidence intervals between .8 and .9. *Auxiliary complexity* was the only score with a point estimate for the slope significantly lower than the rest and with a confidence interval for the slope nearly twice the width of the next largest.

In summary, there is good agreement between the automated C-QPA command results and manual QPA scoring for 32 of the 33 measures in the summary sheet. The score that showed the most variability and least close alignment was *auxiliary complexity*.

Discussion

Analysis of spontaneous speech using the QPA (Rochon et al., 2000; Saffran et al., 1989) provides systematic quantification of lexical, morphological, and structural errors, which can help identify those who have disordered speech following stroke or neurodegenerative disease and measure change following treatment (e.g., Ding et al., 2020; Gordon, 2006; Linebarger et al., 2007; Maher et al., 2006; Medina et al., 2012; Mirman et al., 2019; Thothathiri et al., 2010; Wilson et al., 2010). However, the QPA scoring process, detailed in an 18-page manual (Berndt et al., 2000), is challenging and time consuming for manual implementation, and hours of training are required for scorers to score reliably. Typically, scorers are students who rotate through programs every couple years, so the time that gets devoted to training, retraining, and reliability continues on a regular basis. To overcome these limitations, we developed an automated QPA using the established CLAN program (MacWhinney, 2000). We demonstrated good agreement between

Figure 6. Comparison of automated versus manual raters' standardized Quantitative Production Analysis scores.



Abbreviations: Aux = Auxiliary; NRD = # Nouns Requiring Determiners; S = Sentence; Ss = Sentences; SNP = Subject Noun Phrase; VP = Verb Phrase.

the automated C-QPA command results and manual QPA scoring for 32 of the 33 QPA measures. By automating a structured analysis of spontaneous speech, we provide speech-language pathologists an avenue for easier systematic identification of language deficits, which will facilitate opportunities for therapeutic interventions (cf. Linebarger et al., 2007; Medina et al., 2012). Below, we discuss advantages of the automated approach and explanations for discrepancies between manual and automated scores, highlighting lessons learned through this process of programming and testing the C-QPA command.

Advantages of the Automated C-QPA Approach

Based on the findings presented here, the automated QPA can provide significant advantages to clinicians and researchers who rely on these measures to quantify aspects of language production by increasing analysis speed, reducing demand for linguistic expertise, and increasing replicability. These transcription efficiencies coupled with automated analysis programs will help address the major obstacles to discourse analysis in clinical settings: time and expertise (E. Armstrong, 2000; Bryant et al., 2016).

1. Faster analysis. Once a transcript is prepared, it takes a matter of seconds to run the C-QPA command on that transcript or any number of transcripts, yielding both the individual's analysis worksheet and the summary sheet for all transcripts included in the command. In comparison, manual QPA scoring requires a hand entry for 21 QPA measures for every sentence in the analysis worksheet. On average, this manual scoring took 30–60 min per Cinderella story for trained and experienced coders.
2. Less demand for expertise. Accurate scoring requires a strong grasp of linguistic knowledge, specifically syntax. Even a simple tally of *number of pronouns,* for example, requires the ability to distinguish consistently among demonstrative pronouns, personal pronouns, pronouns *that* and *all* used in place of nouns, and pronouns that serve a syntactic function (e.g., introducing a complement clause).
3. Smoother transcription. When transcribing in CHAT, one transcript can replace the multiple versions of transcript preparation for manual QPA scoring. The manual procedure begins with an original transcription of everything in the language sample. The second step is to copy and paste that original transcription into another file (or section of the same document), where it is necessary to remove all of the fillers (e.g., *um*) and uncontract contractions to count total number of words uttered. The third step is to copy and paste this revised original transcription into another file (or section), where you remove any non–storytelling speech (e.g., comments on the task, habitually used starters, direct discourse markers, frozen elements of the story such as *once upon a time*) and track those changes. The fourth step is to copy and paste the content from the third step into another file (or section), to separate the content into utterances. These steps involve following QPA rules for what to remove and how to

create utterances. Manual transcribers and coders are not always consistent with these determinations. Using CHAT, several of these steps are not necessary as they are built into the transcription or analysis program. For example, fillers are not counted as words, contractions are automatically counted as two words, and the C-QPA analysis automatically excludes words like *once upon a time* and habitually used starters such as *and then, well,* and *so*. Also, it is easy to make changes (e.g., noticing that something should have been excluded, adding the exclusion code [e] next to the word or phrase) and then rerun the MOR and C-QPA commands.

4. Temporal alignment. Using the automated approach, it is simple to time-link utterances in a CHAT file to the audio or video file and then compute words per minute for the participant's utterances only. When done manually, this computation is an estimate based on measuring the entire length of the sample, measuring the amount of time that was examiner speaking time, and then subtracting that from the total. (*Words per minute* was not done in the current study, as the manually coded samples did not compute words per minute or include audio files for time linking to CHAT files.)
5. Replicability. Automatic computation is reliable and replicable because repeated runs will always produce the same results, and scoring will not vary from person to person or study to study.

Improving the psychometric properties of discourse measurement is an important endeavor in our field. Much of the recent literature in aphasia and adult neurogenic communication has focused on discourse from the perspective of both assessment and treatment, making the case that the various forms of spontaneous speech (e.g., narrative, conversation, exposition) are more relevant and ecologically valid than, for example, measures of repetition or naming ability. A review of linguistic analysis of discourse in 165 studies in the aphasia literature reported a variety of stimuli used to elicit discourse samples and a large variety of measures used (n = 536) to analyze samples (Bryant et al., 2016). The calls for standardization and core outcome sets are many (E. Armstrong, 2018; de Riesthal & Diehl, 2018; Dietz & Boyle, 2018a; Kintz & Wright, 2018; Kurland & Stokes, 2018; Pritchard et al., 2017; Wallace et al., 2018; Whitworth, 2018), yet discourse is one area that has eluded any official recommendations in an international effort to establish core outcome sets for measurement in aphasia treatment research (Wallace et al., 2019). A working group, *FOQUSAphasia,* has been established to address spoken discourse collection, analysis, and reporting in aphasia to ensure that psychometrically sound outcome measures and norms are available for clinicians and researchers in the field (Stark et al., 2020). To support these efforts, a set of AphasiaBank CHAT files analyzed by the C-QPA command are available on the AphasiaBank *Discourse Analysis* webpage at the C-QPA link (https://aphasia.talbank.org/discourse/). Rules for running the C-QPA command are provided there and in the CLAN manual. This collection adds to a shared workspace of publicly available and fully analyzed data that facilitates the work toward creating a core outcome set of psychometrically tested measures for discourse in aphasia.

Sources of Discrepancies Between Automated and Manual QPA Scoring

To identify the potential causes of discrepancies between automated and manual scoring, we consulted two trained research assistants who were part of the team that did the manual scoring. In general, when discrepancies were determined to be the result of manual scoring errors, those errors were small and not systematic. Discrepancies arose primarily from inconsistent application of QPA rules, errors during transcription, or manual data input errors. Manual data input errors were often simple, random miscounts (e.g., miscounting the number of verbs in a sentence). However, if lexical category memberships were incorrect, the miscount then led to miscounts for other categories (e.g., *number of open class words*) and the computations based on those category scores (e.g., *number of closed class words, proportion of verbs*). Thus, analysis spreadsheet scoring errors get compounded in subsequent computations, thereby contributing to agreement issues in the derived scores for proportions, means, and indexes.

The one score that did not have good agreement between manual scorers and the automated program was *auxiliary complexity index*. This important score is considered to be an index of the morphological complexity of the matrix verb, the tensed verb in the main clause of the sentence. The score is computed by dividing the *total auxiliary score* by the *number of matrix verbs* and then subtracting 1, because the *total auxiliary score* always contains, at minimum, one uninflected matrix verb. The *total auxiliary score* assigns one additional point for each auxiliary element of the matrix verb (e.g., inflection for tense or agreement on the main verb, auxiliaries, inflection on auxiliaries, modals, semi-auxiliaries). In the QPA manual, instructions for this measure include many details and special instructions. At least half of the discrepancies identified for review were determined to be errors in the manual scoring, some of which were simply miscounts, such as: "Cinderella is working with the stepsisters" received a score of 2 instead of 3 (*is* + *work* + inflection); and "she makes it home before she got eaten by a pumpkin" received a score of 4 instead of 2 (*make* + inflection) because *got eaten* was erroneously scored as the matrix verb. Some errors involved not consistently applying rules, such as how to count conjoined verbs and when to score a point for *not* as part of the verb phrase. Other errors were debatable based on how the instructions were interpreted. For example, one tricky issue involved semi-auxiliaries *is going to* and *have to* that function as modals *will* and *must* and, therefore, receive only 1 point. Modals are given 1 point regardless of mood (e.g., will, would). However, in "they were going to have this ball," the question arises as to whether the past tense of the semi-auxiliary should be ignored such that the matrix verb *have* gets 1 point, and *were going to* gets only 1 point because it is like *would.* In a similar case, *did not get to go* was scored as 3 points, again ignoring the past tense: 1 point for the matrix verb *go,* 1 for *not* as part of the verb phrase, and 1 for *did get to* because it was considered to be similar to the semi-auxiliary *could.* The advantage of a computer-based analysis is that the scoring will not vary from person to person or study to study; the results will be consistent. If C-QPA results are determined to be inaccurate for certain constructions, the program rules can be modified to address the issue. The fact that the nonderived scores used to compute the *auxiliary complexity index* (*number of matrix verbs, total auxiliary score*) showed strong agreement indicates that small disagreements were amplified through the subsequent computations for the derived score. While further work will help hone the accuracy of the *auxiliary complexity index,* users for whom this is an important outcome measure may consider reviewing the component scores in the C-QPA spreadsheet to confirm the numbers and modify any subsequent computations accordingly.

A critical element in the calculation of many QPA scores depends on whether an utterance qualifies as a sentence or not. The QPA manual provides instructions (pp. 13–14, Berndt et al., 2000) and examples to explain the criteria for when an utterance is a sentence. The principal criterion is that an utterance must include a noun and a main verb. Inevitably, the rules will not be clear for every possible utterance, and individual research groups will create operational definitions and conventions to follow. For example, what if the main verb is missing its auxiliary, as in “other women trying to tell her something”? What about a case like “young lady who was working the house with a bunch of mean sisters,” where perhaps something like “there was a” is implied? If these are scored as sentences (as they were in the manual scoring), then all remaining QPA measures in the analysis spreadsheet are computed; if not (as occurred with the automatic scoring), then nine remaining QPA measures in the analysis spreadsheet do not get scored (e.g*., number of matrix verbs, number of SNPs, number of embeddings*). If those measures do not get scored, then the subsequent scores derived from those measures for the summary spreadsheet, such as *proportion of words in sentences, S elaboration index,* and *auxiliary complexity index*, will be affected. Thus, occasional discrepancies in some of the measures may not represent any issues with the computation of those measures, but instead may reflect the fact that a particular utterance did or did not get those measures scored. Though these discrepancies occurred, they were infrequent enough not to interfere with overall good agreement between the manual and automated QPA scoring.

Some other scoring rules that showed some inconsistency involved which words to include in counting the *number of words in SNPs* and which words to exclude from the original transcript in counting *number of narrative words.* A common mistake involved determining if an adverb or adverbial phrase modified the whole sentence or just the SNP. If it modifies the whole sentence, the words are not counted as part of the SNP. The mistake was usually in the direction of counting the adverb or adverbial phrase when it should not be counted (e.g., “at midnight the pumpkins would happen,” “her three stepsisters constantly abused her”). Similarly, determining if clauses are embedded in the SNP and should be counted can be challenging depending on how the sentence is worded and if it includes paraphasias or agrammatic elements. Words to be counted in SNPs are limited to open class words plus pronouns (only those that are used in place of nouns). Sentences that start with *there* (e.g., “there was three sisters and one,” “there was going to be a ball”) were always counted (in manual scoring) as having one SNP and one word in the SNP even though *there* is not counted as an open class word or pronoun. Also, the word *all* as in “all the sisters were kind of being snubby” was counted as a pronoun though it was not in place of a noun. Again, these occasional discrepancies caused some variability in the agreement on nonderived and derived SNP measures, but overall agreement was still good. We will continue to drill down on these types of issues to ensure that the automated program is as accurate as possible and faithful to the QPA scoring rules.

In the transcription process, certain narrative words are supposed to be excluded. These include items such as frozen elements (e.g., *once upon a time, happily ever after*) neologisms, direct responses to specific questions, comments, habitually used starters, coordinating conjunctions that join two otherwise independent sentences, and direct discourse markers. The manual scoring did not consistently exclude habitual starters (e.g., *all of a sudden, alright, well*) and direct discourse markers (e.g., *he said*). Although we matched the transcripts used for automatic scoring as closely as possible to the manually scored transcripts for purposes of comparing similar inputs (e.g., if the manually scored transcripts included discourse markers, we did not exclude them from being counted in the CHAT transcripts), the C-QPA program is written to automatically ignore several habitual starters and frozen elements. This may have contributed to some of the variability seen in the *proportion of words in sentences* measure (*number of words in sentences* divided by *number of narrative words*), but again, the discrepancies were minor enough so that agreement between manual and automated scoring was still good. The ease of being able to insert a marker, [e], in the CHAT transcript next to words that should be excluded and then rerun the MOR and C-QPA programs is an added advantage of the automated approach.

Finally, one set of QPA scores, those involving inflection (*number of inflectable verbs, number of inflectable verbs inflected,* and *inflection index*), was not amenable to automated analysis and was not included in the C-QPA command. To count *number of inflectable verbs,* the QPA manual states the following.

> *Count all verbs that could be grammatically inflected with the addition of a suffix or a stem-change, including those occurring outside of a sentential or phrasal context. Include as inflectable: inflectable tokens of any regular verb, whether inflected or not; regularly inflectable*

tokens of irregular main verbs (e.g., "I go" –> "am going").

So, the word *walk* in "I walk to the party" would be counted as inflectable but not inflected, because it **could be** grammatically inflected according to the rules. It would be scored the same (inflectable but not inflected) in "I walk to the party yesterday" or "she walk to the party," when in both cases it **should have been** inflected. The verb *going* used correctly in a sentence such as "I am going to the party" or incorrectly as in "I going party" would be counted as an inflectable verb that was inflected. Basically, all present progressive verb productions count as inflectable verbs that are inflected. Thus, it is not entirely clear what the *inflection index* (*number of inflected verbs that were inflected* divided by the *number of inflectable verbs*) reveals. Interestingly, in the manually coded score sheets, the *inflection index* was exactly 1.0 for 80% of the participants, so the range of scores was limited. These measures involving inflection can be pursued further as research or clinical needs warrant.

Conclusions and Future Directions

In their review of the literature on linguistic analysis of discourse in aphasia, Bryant et al. (2016) found that the QPA was the most frequently used system by the group of studies (*n* = 17) that used multiple measures of linguistic structures. It is a system that has been around for many decades and is familiar and useful to many in the field. Automating the process dramatically reduces the time and expertise necessary for its application in clinical settings as a tool for assessment, treatment planning, and treatment outcome measurement. These same advantages are amplified in the research setting where large numbers of transcripts can be reliably and consistently analyzed within minutes. As is inevitable with research, multiple passes through the data sets are necessary for any number of reasons including correcting errors, changing criteria, or adding codes for further analyses. In these cases, the program can be rerun in minutes, generating the full set of summary data in a spreadsheet format. In addition, the transcribed language samples can be analyzed more broadly using other CLAN programs, for example, to examine type and frequency of word errors, gestures, self-corrections, and comparisons within and across shared clinical databases in the TalkBank system. Finally, although outside the scope of this current project, future analyses of the C-QPA results from this data set may tease out the individuals with language problems in this group of left-hemisphere stroke participants.

Acknowledgments

This work was supported by National Institute on Deafness and Other Communication Disorders Grants R01-DC008524 (2007-2022, awarded to MacWhinney) and R01-DC014976 (awarded to Schnur). The authors wish to thank Erica Johns, Bowie Lin, Hao Yan, Rachel Zahn, and Riya Mehta for transcription and analysis of narrative speech samples. In particular, the authors thank Rachel Zahn and Riya Mehta for their expertise coupled with their willingness to answer a seemingly infinite number of questions, both of which were invaluable to the project. The authors gratefully acknowledge and thank their research subjects and their caregivers for their willingness to participate in this research. This work was presented at the Clinical Aphasiology Conference (2019) and the Academy of Aphasia (2020).

References

Armstrong, E. (2020). Aphasic discourse analysis: The story so far. *Aphasiology, 14*(9), 875–892. https://doi.org/10.1080/02687030050127685

Armstrong, E. (2018). The challenges of consensus and validity in establishing core outcome sets. *Aphasiology, 32*(4), 465–468. https://doi.org/10.1080/02687038.2017.1398804

Armstrong, L., Brady, M., Mackenzie, C., & Norrie, J. (2007). Transcription-less analysis of aphasic discourse: A clinician's dream or a possibility? *Aphasiology, 21*(3–4, 374), 355. https://doi.org/10.1080/02687030600911310

Berndt, R. S., Wayland, S., Rochon, E., Saffran, E., & Schwartz, M. (2000). *Quantitative Production Analysis: A training manual for the analysis of aphasic sentence production*. Psychology Press.

Boles, L. (1998). Conversational discourse analysis as a method for evaluating progress in aphasia: A case report. *Journal of Communication Disorders, 31*(3), 261–274. https://doi.org/10.1016/S0021-9924(98)00005-7

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics, 30*(7), 489–518. https://doi.org/10.3109/02699206.2016.1145740

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31*(10), 1105–1126. https://doi.org/10.1080/02687038.2016.1239013

Cruice, M., Botting, N., Marshall, J., Boyle, M., Hersh, D., Pritchard, M., & Dipper, L. (2020). UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International Journal of Language & Communication Disorders, 55*(3), 417–442. https://doi.org/10.1111/1460-6984.12528

de Riesthal, M., & Diehl, S. K. (2018). Conceptual, methodological, and clinical considerations for a core outcome set for discourse. *Aphasiology, 32*(4), 469–471. https://doi.org/10.1080/02687038.2017.1398805

Dietz, A., & Boyle, M. (2018a). Discourse measurement in aphasia research: Have we reached the tipping point? *Aphasiology, 32*(4), 459–464. https://doi.org/10.1080/02687038.2017.1398803

Dietz, A., & Boyle, M. (2018b). Discourse measurement in aphasia: Consensus and caveats. *Aphasiology, 32*(4), 487–492. https://doi.org/10.1080/02687038.2017.1398814

Ding, J., Martin, R. C., Hamilton, A. C., & Schnur, T. T. (2020). Dissociation between frontal and temporal-parietal contributions to connected speech in acute stroke. *Brain, 143*(3), 862–876. https://doi.org/10.1093/brain/awaa027

Fromm, D., MacWhinney, B., & Thompson, C. K. (2020). Automation of the Northwestern Narrative Language Analysis System. *Journal of Speech, Language, and Hearing Research, 63*(6), 1835–1844. https://doi.org/10.1044/2020_JSLHR-19-00267

Gordon, J. K. (2006). A Quantitative Production Analysis of picture description. *Aphasiology, 20*(02–04), 188–204. https://doi.org/10.1080/02687030500472777

Hsu, C.-J., & Thompson, C. K. (2018). Manual versus automated narrative analysis of agrammatic production patterns: The

Northwestern Narrative Language Analysis and Computerized Language Analysis. *Journal of Speech, Language, and Hearing Research, 61*(2), 373–385. https://doi.org/10.1044/2017_JSLHR-L-17-0185

Kintz, S., & Wright, H. H. (2018). Discourse measurement in aphasia research. *Aphasiology, 32*(4), 472–474. https://doi.org/10.1080/02687038.2017.1398807

Kurland, J., & Stokes, P. (2018). Let's talk real talk: An argument to include conversation in a D-COS for aphasia research with an acknowledgment of the challenges ahead. *Aphasiology, 32*(4), 475–478. https://doi.org/10.1080/02687038.2017.1398808

Linebarger, M., McCall, D., Virata, T., & Berndt, R. S. (2007). Widening the temporal window: Processing support in the treatment of aphasic language production. *Brain and Language, 100*(1), 53–68. https://doi.org/10.1016/j.bandl.2006.09.001

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Erlbaum.

MacWhinney, B. (2011). The expanding horizons of corpus linguistics. In J. Newman, H. Baayen, & S. Rice (Eds.), *Corpus-based studies in language use, language learning, and language documentation* (pp. 177–212). Rodopi.

Maher, L. M., Kendall, D., Swearengin, J. A., Rodriguez, A., Leon, S. A., Pingel, K., Holland, A., & Rothi, L. J. G. (2006). A pilot study of use-dependent learning in the context of constraint induced language therapy. *Journal of the International Neuropsychological Society, 12*(6), 843. https://doi.org/10.1017/S1355617706061029

Martin, R. C., & Schnur, T. T. (2019). Independent contributions of semantic and phonological working memory to spontaneous speech in acute stroke. *Cortex, 112,* 58–68. https://doi.org/10.1016/j.cortex.2018.11.017

Medina, J., Norise, C., Faseyitan, O., Coslett, H. B., Turkeltaub, P. E., & Hamilton, R. H. (2012). Finding the right words: transcranial magnetic stimulation improves discourse productivity in non-fluent aphasia after stroke. *Aphasiology, 26*(9), 1153–1168. https://doi.org/10.1080/02687038.2012.710316

Miller, J. F., & Iglesias, A. (2015). *Systematic Analysis of Language Transcripts* [Computer software]. SALT Software.

Mirman, D., Kraft, A. E., Harvey, D. Y., Brecher, A. R., & Schwartz, M. F. (2019). Mapping articulatory and grammatical subcomponents of fluency deficits in post-stroke aphasia. *Cognitive, Affective, & Behavioral Neuroscience, 19*(5), 1286–1298. https://doi.org/10.3758/s13415-019-00729-9

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2017). Reviewing the quality of discourse information measures in aphasia. *International Journal of Language & Communication Disorders, 52*(6), 689–732. https://doi.org/10.1111/1460-6984.12318

Rochon, E., Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language, 72*(3), 193–218. https://doi.org/10.1006/brln.1999.2285

Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language, 37*(3), 440–479. https://doi.org/10.1016/0093-934X(89)90030-8

Stark, B. C., Dutta, M., Murray, L., Bryant, L., Fromm, D., MacWhinney, B., Ramage, A. E., Roberts, A., den Ouden, D. B., Brock, K., McKinney-Bock, K., Paek, E. J., Harmon, T. G., Yoon, S. O., Themistocleous, C., Yoo, H., Aveni, K., Gutierrez, S., & Sharma, S. (2020). Standardizing assessment of spoken discourse in aphasia: A working group with deliverables. *American Journal of Speech-Language Pathology*. https://doi.org/10.1044/2020_AJSLP-19-00093

Thompson, C. K., Shapiro, L. P., Tait, M. E., Jacobs, B. J., Schneider, S. L., & Ballard, K. J. (1995). A system for the linguistic analysis of agrammatic language production. *Brain and Language, 51,* 124–129.

Thothathiri, M., Schwartz, M. F., & Thompson-Schill, S. L. (2010). Selection for position: The role of left ventrolateral prefrontal cortex in sequencing language. *Brain and Language, 113*(1), 28–38. https://doi.org/10.1016/j.bandl.2010.01.002

Wallace, S. J., Worrall, L. E., Rose, T., & Le Dorze, G. (2018). Discourse measurement in aphasia research: have we reached the tipping point? A core outcome set… or greater standardisation of discourse measures? *Aphasiology, 32*(4), 479–482. https://doi.org/10.1080/02687038.2017.1398811

Wallace, S. J., Worrall, L., Rose, T., Le Dorze, G., Breitenstein, C., Hilari, K., Babbitt, E., Bose, A., Brady, M., Cherney, L. R., Kelly, H., Kiran, S., Laska, A. C., Marshall, J., & Webster, J. (2019). A core outcome set for aphasia treatment research: The ROMA consensus statement. *International Journal of Stroke, 14*(2), 180–185. https://doi.org/10.1177/1747493018806200

Whitworth, A. (2018). The tipping point: Are we nearly there yet? *Aphasiology, 32*(4), 483–486. https://doi.org/10.1080/02687038.2017.1398812

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., Miller, B. L., & Gorno-Tempini, M. L. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain, 133*(7), 2069–2088. https://doi.org/10.1093/brain/awq129