JSLHR

Research Article

Automation of the Northwestern Narrative Language Analysis System

Davida Fromm,[a] Brian MacWhinney,[a] and Cynthia K. Thompson[b]

Purpose: Analysis of spontaneous speech samples is important for determining patterns of language production in people with aphasia. To accomplish this, researchers and clinicians can use either hand coding or computer-automated methods. In a comparison of the two methods using the hand-coding NNLA (Northwestern Narrative Language Analysis) and automatic transcript analysis by CLAN (Computerized Language Analysis), Hsu and Thompson (2018) found good agreement for 32 of 51 linguistic variables. The comparison showed little difference between the two methods for coding most general (i.e., utterance length, rate of speech production), lexical, and morphological measures. However, the NNLA system coded grammatical measures (i.e., sentence and verb argument structure) that CLAN did not. Because of the importance of quantifying these aspects of language, the current study sought to implement a new, single, composite CLAN command for the full set of 51 NNLA codes and to evaluate its reliability for coding aphasic language samples.
Method: Eighteen manually coded NNLA transcripts from eight people with aphasia and 10 controls were converted into CHAT (Codes for the Human Analysis of Talk) files for compatibility with CLAN commands. Rules from the NNLA manual were translated into programmed rules for CLAN computation of lexical, morphological, utterance-level, sentence-level, and verb argument structure measures.
Results: The new C-NNLA (CLAN command to compute the full set of NNLA measures) program automatically computes 50 of the 51 NNLA measures and generates the results in a summary spreadsheet. The only measure it does not compute is the number of verb particles. Statistical tests revealed no significant difference between C-NNLA results and those generated by manual coding for 44 of the 50 measures. C-NNLA results were not comparable to manual coding for the six verb argument measures.
Conclusion: Clinicians and researchers can use the automatic C-NNLA to analyze important variables required for quantification of grammatical deficits in aphasia in a way that is fast, replicable, and accessible without extensive linguistic knowledge and training.

The spoken language of people with aphasia (PWA) is often impaired. Individuals with nonfluent agrammatic aphasia show a range of grammatical errors, including omissions, additions, and substitutions of grammatical markers and function words, as well as an overall reduction in syntactic structure (Bastiaanse & Thompson, 2012). In contrast, people with fluent aphasia show relatively fewer grammatical errors in production but show difficulty with semantic aspects of production (Edwards, 2005). For both clinical and research purposes, it is important to identify and quantify the grammatical abilities of PWA. Although this may be accomplished by administering structured language tests, production ability also is often based on spontaneous speech samples.

However, determining grammatical production patterns based on spontaneous speech is challenging. Hand coding of speech samples requires a high level of understanding of both lexical and morphosyntactic aspects of the grammar, and it is labor intensive and time-consuming. A solution to these problems is to rely on natural language processing (NLP) methods for automatic morphosyntactic tagging and analysis. However, the question is whether the output derived from automatic coding systems is as accurate as that derived from careful hand coding. Hsu and Thompson (2018) addressed this issue by comparing a state-of-the-art system for hand coding of narrative transcripts—the NNLA (Northwestern Narrative Language Analysis) system (Thompson, Shapiro, Tait, et al., 1995) with a parallel system for automatic analysis—CLAN (Computerized Language Analysis; MacWhinney, 2000).

The NNLA was developed in 1995 (Thompson, Shapiro, Tait, et al., 1995) to quantify production patterns associated with agrammatic aphasia and has since been

[a]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA
[b]Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL
Correspondence to Davida Fromm: fromm@andrew.cmu.edu

Editor-in-Chief: Sean M. Redmond
Editor: Christos Salis
Received October 8, 2019
Revision received January 29, 2020
Accepted February 26, 2020
https://doi.org/10.1044/2020_JSLHR-19-00267

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

used to analyze discourse from individuals with both nonfluent and fluent stroke-induced aphasia, as well as primary progressive aphasia and Alzheimer's disease (Ballard & Thompson, 1999; Barbieri et al., 2019; Faroqi-Shah & Thompson, 2007; Jacobs & Thompson, 2000; Kim & Thompson, 2004; Mack et al., 2015; Mack & Thompson, 2017; Meltzer-Asscher & Thompson, 2014; Thompson et al., 1997, 2012, 2013; Thompson, Shapiro, Li, & Schendel, 1995). NNLA includes rules for transcribing and manually coding language samples. The *transcription* rules specify how to segment the language sample into utterances and how to handle abandoned utterances, interrupted utterances, comments, spelling conventions, word fragments, repetitions, mazes, and other important aspects of language production. NNLA transcription is done in a way that facilitates later analysis through SALT (Systematic Analysis of Language Transcripts; Miller & Chapman, 1983). For *coding,* NNLA includes codes for five separate levels of production: (I) the utterance level, (II) the sentence level, (III) the lexical level, (IV) the bound morpheme level, and (V) the verb argument structure level. The codes on each of the five levels are linearly entered by hand, based on the criteria for assigning each code on each level. Appendix A lists the language variables that are coded at each level. Below is a transcribed and coded utterance produced by a person (S = speaker) with nonfluent aphasia telling the Cinderella story.

```
S (And so but s uh Cinderella uh no) stepmother
(uh) lock/ed up (uh uh)  Cinderella because
(uh) not want to (uh) find (I do/n't know)
<I> [ *s] [ g] [ au]
<II> [ cs] [ as] [ e2] [ ac] [ cc]
<III> [ -dets] [ n] [ v] [ prt] [ n] [ conj] [ -pros] [ -aux]
[ neg] [ v] [ -pros] [ to] [ v]
<IV> [ ed]
<V> [ op3xy] [ xs] [ yo] [ vmi2] [ *cxs'] [ -xs] [ s']
[ *vmi2] [ *cxy] [ -xs] [ -yo] [ vmi2]
```

The system for automatic analysis, called CLAN (MacWhinney, 2000), is a set of programs that permits automatic analysis of many of the same scoring categories as the NNLA, but implements the computation automatically for language samples transcribed in CHAT (Codes for the Human Analysis of Talk) format. It has been used in a wide variety of disciplines (e.g., child language, first and second language acquisition, conversation analysis) and more than 10 different languages over the last 30 years. Like NNLA and SALT, the CHAT transcription format has conventions for marking abandoned utterances, interrupted utterances, word fragments, repetitions, and other types of behaviors encountered in language transcription. The CLAN program and data in the CHAT format have been used in over 8,000 published papers across a wide variety of disciplines. Software and electronic manuals for CHAT and CLAN are free and downloadable from the TalkBank website: https://talkbank.org/. In CHAT, the sample NNLA utterance given above, produced by the study participant (PAR), would look like this:

```
*PAR: <and so but &+s &-uh Cinderella &-uh no>
      [ //] stepmother &-uh locked up &-uh &-uh
      Cinderella because &-uh not want to &-uh
      find <I don't know> [ e] .[+ gram]
```

Once this level of transcription is achieved, the user runs a single CLAN command (MOR: *mor filename.cha*) for automatic tagging of morphological structures, part-of-speech categories, and grammatical relations, which then appear on %mor and %gra lines under each utterance. The sample utterance in the transcript would now look like this:

```
*PAR: <and so but &+s &-uh Cinderella &-uh no>
      [ //] stepmother &-uh locked up &-uh &-uh
      Cinderella because &-uh not want to &-uh
      find <I don't know> [ e] .[+ gram]
%mor: step#n|mother v|lock-PAST adv|up n:prop|
      Cinderella conj|because neg|not v|want
      inf|to v|find .
%gra: 1|2|SUBJ 2|0|ROOT 3|2|JCT 4|3|POBJ 5|7|
      LINK 6|7|NEG 7|2|CJCT 8|9|INF 9|7|COMP
      10|2| PUNCT
```

In Hsu and Thompson (2018), both systems were used to analyze a common set of transcriptions of narrative productions of the Cinderella story. The results showed that automatic CLAN output was largely consistent with manual NNLA coding, with comparable general, lexical, and morphological outputs (e.g., numbers of utterances, words, adverbs, adjectives, negation markers, infinitival markers, possessive markers, regular and irregular plural markers). However, 15 important measures (e.g., sentence complexity ratio, percent correct inflections, verb argument structure) could not be computed by CLAN and required manual coding. Appendix A lists the NNLA measures that were compared.

Encouraged by these findings, we took a closer look at the areas of agreement and disagreement between manual NNLA and automated CLAN and sought to develop a single, new CLAN command to compute the full set of NNLA measures (the C-NNLA command). We then evaluated the reliability of the C-NNLA command. In other words, we sought to determine the extent to which the rules of a sophisticated language analysis system for aphasia (NNLA) could be reliably implemented by means of an established computer-based analysis system (CLAN) to take advantage of the strengths of both approaches.

Method

Language Sample Transcription

As a preliminary step for comparing NNLA and CLAN, we needed to be sure CHAT transcripts excluded extraneous material so that they would be in line with manual NNLA transcription methods (see Appendix B). Although some conventions were already in place within CHAT transcripts for exclusion of, for example, comments such as "now let's see," "I can't say it" (marked with the

[+ exc] code) and revisions, repetitions, fillers, and sound fragments (marked with [//], [/], &- and &+, respectively), the new C-NNLA command was programmed to also automatically ignore initial conjunctions (e.g., "and," "but," "so," "and so," "well," "then," "and then"), which were otherwise included in CHAT samples for analysis. In addition, we developed a method for marking other elements for exclusion that could not be automated. For this method, [e] is entered following any word or string of words to prevent them from appearing on the %mor and %gra tiers and being counted in the analysis. To summarize, the C-NNLA command handles exclusions through a combination of three methods: exclusion based on normal CHAT codes, automatic exclusion by the program, and hand marking with [e]. In the example utterances below, linguistic analysis will be applied to the bolded words only.

*PAR: <and then> [/] and then &-um [x 5] **the** &+g &+dr **dress**. [+ gram]
*PAR: &-uh then <I think uh &+b but I don't know> [e] something [//] **a** &+m **maid appears**.

Automatic C-NNLA required two further additions to basic CHAT transcription procedures. In keeping with the use of a [+ gram] code for marking ungrammatical utterances and the [+ exc] code for marking whole utterance exclusion, already a part of basic CHAT transcription, we added [+ sem] to mark semantically flawed utterances. In addition, morphological error coding was expanded to capture irregular versus regular inflectional endings. Examples of these existing codes and the two new manual codes are given below. Also, a complete list of these C-NNLA transcription and coding rules is available in Appendix B as well as at the *Discourse Analysis* link of the AphasiaBank webpage and in the C-NNLA section of the CLAN manual.

- Semantically flawed utterance

PAR: the father_in+law [: father] [s:r] says +"/. [+ sem]

- Grammatically flawed utterance

*PAR: happily ever after . [+ gram]

- Formulaic or unrelated utterance

*PAR: I don't know . [+ exc]

- Regular inflection error

PAR: they yells [: yell] [m:+3s:a] at the little girl all the time . [+ gram]

- Irregular inflection error

PAR: it felled [: fell] [m:+ed] out . [+ gram]

C-NNLA Programming and Rule Modifications

To evaluate the extent to which CLAN could provide automatic computation of NNLA measures, we formulated the C-NNLA command to generate the NNLA outcome measures reported by Hsu and Thompson. This new program, like several other CLAN programs (e.g., EVAL, MORTABLE), bundles separate CLAN commands into a single command and outputs a set of outcome measures. Although Hsu and Thompson succeeded in computing 36 of 51 total NNLA measures with CLAN, a variety of existing CLAN commands were required to do so. This meant that our first goal in building C-NNLA was to bundle the relevant analyses for those 36 measures into a single command. The second goal was to formulate methods to compute the 15 additional measures that could not be computed automatically by the existing version of CLAN (see Appendix A for a list of all 51 measures.)

To do this, we translated the rules from the NNLA manual into rules for CLAN using data from the %mor and %gra tiers in the CHAT files. For example, to count irregular plural forms (a Level IV bound morpheme code in NNLA), C-NNLA searches for *&PL* on the %mor tier. In this simple example, the ampersand indicates irregular affixation and *PL* indicates the plural suffix. In a more complicated example, *wh*-words (a Level III lexical code in NNLA) are identified and tallied from five different part-of-speech codes in CLAN: *pro:rel* (relative pronouns, such as "who" in "the children who want to go"), *pro:int* (interrogative pronouns, such as "where" in "where is she"), *conj* (conjunctions, such as "when" in "when I was a child"), and *det:int* (interrogative determiners, such as "what" in "what fun we had"). The goal was to count and compute the NNLA outcome measures in full compliance with the NNLA manual rules. The complete set of C-NNLA rules is available at the AphasiaBank webpage at the *Discourse Analysis* link and in the C-NNLA section of the CLAN manual at the TalkBank website.

Testing Procedure

The participant transcripts (n = 8 PWA and 10 controls) used to test C-NNLA were the same as those used and described by Hsu and Thompson (with minor modifications described below). Participants were asked to tell the Cinderella story. Mean ages for the aphasia and control participants were 58.2 and 57.4 years, respectively; mean education was 17 years and 17.1 years, respectively. The aphasia group included five men and three women with a mean time poststroke onset of 6.8 years (range: 1.6–18.0). All participants with aphasia were clinically diagnosed with agrammatic aphasia. For additional information on the participants (e.g., language test scores), readers can refer to Table 1 in the original article.

To allow for targeted debugging and testing of the 15 newly implemented outcome measures, two CHAT files were created by randomly selecting 10 consecutive utterances from each of the 18 files and making one composite master test file for controls and one for aphasia. The control master file included 100 utterances (10 participants, 10 utterances each); the aphasia master test file had 80 utterances (eight participants, 10 utterances each). These master test files are available at the AphasiaBank *Discourse*

Analysis link. On a separate coding tier in the CHAT transcripts, we used CLAN's Coder Mode to enter NNLA codes for the 15 new measures computed by the C-NNLA command. The example below shows one of the actual sentences, with a speaker tier (*PAR), a sentence-level coding tier (%slc) for the NNLA codes, and the automatically generated morphological and grammatical relations tiers (%mor, %gra). Reading from left to right, the %slc sentence-level coding tier shows that this is a sentence ($S:s), a simple sentence ($C:ss), an active sentence ($ST:as), with no embeddings ($E:0), with a regular verb that was correctly inflected ($VI:r:c), and was also a two-place obligatory verb used with the correct argument structure ($V:2ob:c).

```
*PAR: and one day they received an invitation
      to a ball.
%slc: $S:s $C:ss $ST:as $E:0 $VI:r:c $V:2ob:c
%mor: coord|and det:num|one n|day pro:sub|
      they v|receive-PAST det:art|a
      n|invite&dv-ATION prep|to det:art|a n|
      ball .
%gra: 1|5|LINK 2|3|QUANT 3|5|JCT 4|5|SUBJ 5|
      0|ROOT 6|7|DET 7|5|OBJ 8|7|NJCT
      9|10|DET 10|8|POBJ 11|5|PUNCT
```

After targeted debugging with the two composite test files, we were ready to run the C-NNLA command on the full Cinderella narrative samples from all participants.[1] To enable this direct comparison, all original, manually coded samples were converted into CHAT transcripts for analysis with the new C-NNLA command. The MOR command was run on these transcripts to create the morphological and grammatical relations tiers (%mor and %gra) that contain the information for computing the NNLA outcome measures. The C-NNLA command created a spreadsheet, with data for individual transcripts for all measures, with the exception of *total particles*.[2] Using the Mann–Whitney *U* between-groups test ($p < .05$ significance level, two-tailed, as in the original article), we compared results from the NNLA data summary sheet provided by Hsu and Thompson for each of the 18 individuals (see Hsu & Thompson for group means and standard deviations) with those derived from the automated C-NNLA command. We focused on the 15 newly implemented measures and the four outcome measures that were found to differ significantly in the Hsu and Thompson paper. All other outcome measures had already been shown to be comparable (not significantly different) by Hsu and Thompson.

[1]To run the analysis, we typed C-NNLA +t*par *.cha in the CLAN Commands window. The command performed the C-NNLA analysis on the participant's utterances (+t*par) in all 18 CHAT files (*.cha).
[2]Verb particles do not have a separate part-of-speech tag in CLAN's English lexicon. Because of the difficulty involved in distinguishing particles from prepositions and adverbs, verb particle constructions (e.g., "up," as in "the fairy godmother shows up") are typically tagged as adverbs by CLAN's MOR command.

Results

The C-NNLA command, executed in only a few seconds, computed 50 of the 51 NNLA measures, including the 36 compared by Hsu and Thompson and the 15 newly automated measures minus *verb particles* (see Appendix A for a full list of measures). Results showed no significant differences between the manually generated NNLA results and those derived from the automated C-NNLA for all 36 of the measures studied by Hsu and Thompson. For the PWA group, the Mann–Whitney *U* values were all above 13, and for the Control group, they were all above 23. Although 32 of these 36 measures had been found to be comparable in Hsu and Thompson, four measures had shown significant differences: (a) *percent of verbs* (over all words) for the PWA group, (b) *noun-to-verb ratio* for controls, (c) *total conjunctions* for both groups, and (d) *total modals* for both groups. The reason for this is that the previous comparisons were made between transcripts that were not identical. The transcripts used by Hsu and Thompson for automated analysis did not exclude comments (e.g., "I guess," "I think so"), interjections (e.g., "you know"), or initial conjunctions (e.g., "and," "but"). Once we made a comparison with identical transcripts, the divergences disappeared.

Of the 15 newly implemented measures, eight were comparably computed, with no significant differences found between C-NNLA and manually generated NNLA results (see Table 1). The verb-and-verb argument structure measures, however, showed mixed differences between hand coding and C-NNLA analysis (see Table 2). Notably, for the proportion of one-place and two-place verbs produced, the two methods did not differ for the aphasic participants; however, they did so for the healthy controls. Furthermore, for the aphasic speakers, but not the controls, production of one- and three-place (but not two-place) verbs with correct arguments was tallied similarly using the two methods.

Discussion

This study presents a new CLAN program, C-NNLA, which allows for the automatic computation of 50 NNLA measures analyzed by Hsu and Thompson (2018) in their paper comparing manual and automated analysis of grammatical production in PWA. We compared the results derived from manual NNLA coding to those derived from automatically generated C-NNLA, which included 36 measures (compared by Hsu and Thompson) that previously required execution of multiple CLAN commands, and 14 new measures that previously could not be automatically generated. Of the measures that can now be automatically computed with a single command, the data showed high agreement between the two methods, with the exception of the six verb-and-verb argument structure measures. The new C-NNLA command now makes it possible to compute more NNLA measures automatically and to do so with much more efficiency using one command instead of multiple commands and subsequent spreadsheet formula computations. Because these measures are important for quantifying

Table 1. Newly implemented C-NNLA measures and data (mean and standard deviation) with no statistically significant differences between automated results and manually generated NNLA results.

Outcome measure	C-NNLA, *M* (*SD*)		Hsu & Thompson, *M* (*SD*)	
	PWA	Control	PWA	Control
% sentences produced	79.60 (7.31)	97.51 (3.10)	80.02 (5.86)	98.22 (2.2)
% sentences with flawed syntax	51.68 (19.07)	2.16 (3.24)	49.73 (17.76)	2.18 (3.29)
% sentences with flawed semantics	8.66 (7.25)	0.39 (0.89)	11.9 (8.47)	0.38 (0.89)
sentence complexity ratio	0.27 (0.10)	0.80 (0.34)	0.24 (0.11)	0.82 (0.71)
# of embedded clauses/sentence	0.24 (0.08)	0.73 (0.23)	0.19 (0.09)	0.71 (0.02)
% correct regular inflection	95.61 (7.63)	99.33 (2.11)	86.74 (10.94)	99.23 (2.43)
% correct irregular inflection[a]	90.76 (17.86)	99.74 (0.83)	73.78 (33.63)	86.74 (10.94)
% sentences with correct syntax, semantics[b]	46.28 (19.98)	97.45 (3.33)	44.61 (17.72)	97.21 (3.25)

Note. C-NNLA = CLAN command to compute the full set of NNLA measures; NNLA = Northwestern Narrative Language Analysis; PWA = people with aphasia.

[a]NNLA numbers are lower due to coding errors in one NNLA file. [b]No statistical test was done on this measure due to the lack of individual data from the original source for comparison.

the grammatical abilities of PWA, automated measures are of value to both researchers and clinicians who study and treat individuals with aphasia. Clinically, this information can be used for differential diagnosis of aphasia types, development of treatment targets, and measurement of treatment outcomes in a naturalistic context. In research, the information can be used for careful description of participant language profiles as well as for deeper exploration of syndrome classification and advancing our understanding of the relationships between these measures and neurocognitive variables.

As a case in point, the original purpose of the NNLA, dating back to its development beginning in 1992, was to examine recovery and treatment in individuals with agrammatic Broca's aphasia. However, the current version of the NNLA is intended to be used by clinicians and researchers to document narrative ability in aphasia more generally. The aphasia transcripts used here were from the same eight participants in Hsu and Thompson. Though all had a clinical diagnosis of agrammatic aphasia, their scores from the Western Aphasia Battery–Revised (Kertesz, 2006) Part I subtests classify them as Broca (n = 2), Transcortical Motor (n = 1), and Anomic (n = 5). Automation of the NNLA process will allow for more detailed analyses and comparisons of language samples from larger groups of participants with different types of aphasia as well as other communication disorders (e.g., primary progressive aphasia, dementia), some of which have already been analyzed by hand using the NNLA (Mack et al., 2015; Thompson et al., 1997).

Six of the measures that showed a significant divergence between hand coding in NNLA and automated C-NNLA scoring involved verb argument coding, as listed in Table 2. As Hsu and Thompson noted, it is very difficult for even well-trained human coders to achieve hand-coding consistency in the assignment of thematic roles in narrative discourse. It is difficult to distinguish, for example, between argument-obligatory and optional verbs. For example, the verb "send" is a three-argument (agent–theme–goal) verb as in "Mary sent the letter to her mother," but the Goal argument is optional (e.g., "Mary sent the letter"). That is, the third argument need not be overtly produced for sentences to be grammatical. This is true for optional two-argument (agent–theme) verbs as well (e.g., "Mary ate the cake" vs. "Mary ate during her lunch break"). Both instantiations of the verb "eat" are grammatical. In contrast,

Table 2. C-NNLA compared to manual NNLA results for verb-and-verb argument structure measures.

Outcome measure	C-NNLA, *M* (*SD*)		Hsu & Thompson, *M* (*SD*)	
	PWA	Control	PWA	Control
% 1-place verbs/all verbs	19.75 (16.93)	20.58 (3.96)*	29.39 (14.01)	31.16 (5.39)
% 2-place verbs/all verbs	36.08 (17.04)	28.62 (5.75)*	54.78 (16.70)	62.12 (7.12)
% 3-place verbs/all verbs	12.64 (7.62)*	11.88 (4.37)*	4.57 (6.12)	5.97 (4.06)
% 1-place verbs, correct arguments	75.11 (18.99)	52.83 (10.17)*	89.31 (11.27)	100.00 (0.00)
% 2-place verbs, correct arguments	60.31 (9.84)*	56.83 (14.41)*	89.78 (5.19)	100.00 (0.00)
% 3-place verbs, correct arguments	60.00 (30.09)	46.63 (17.60)*	88.28 (7.42)	99.00 (3.16)

Note. C-NNLA = CLAN command to compute the full set of NNLA measures; NNLA = Northwestern Narrative Language Analysis; PWA = people with aphasia.

*$p < .05$, statistically significant difference between C-NNLA and manual coding.

obligatory verbs require the second or third argument, as in "The nurse weighed the patient" (two-argument) or "My friend lent me his car" (three-argument). Without the required elements, the sentences would be grammatically flawed. In addition, alternative meanings may be associated with variant argument structures. For example, the verb "sail" is an unaccusative one-place intransitive as in "The boat sailed" (with a Theme thematic role in the subject position), but it also may be used transitively as in the sentence "Tim sailed his boat."

Within the field of NLP, the gold standard for thematic role assignment has been PropBank (Palmer et al., 2005). This resource presents a large number of alternative frames for many common verbs, along with methods for extending the analysis to additional similar verbs. However, there is no effective algorithm for selecting the correct alternative frame in a given case (Gildea & Jurafsky, 2002). More recently, Zettlemoyer and colleagues (FitzGerald et al., 2018) have used crowdsourcing to create a large database of human semantic role assignments in verb argument structures for a specified set of verbs. In some cases, these judgments are fairly consistent, but in others, there remains a large amount of disagreement between human coders. Using this database, they have trained classifiers to make judgments that generally correspond to those made by human coders for this limited set of verbs. Eventually, NLP work of this type will permit increasingly accurate automatic characterization of correct and incorrect use of verb argument structure in transcripts from agrammatic aphasia. However, given the fact that human coders disagree on some assignments, there can never be complete accuracy for judgments regarding these structures.

Based on the present findings, we can rely with increasing confidence on automatic computation through C-NNLA for the majority of grammatical variables that are important for quantifying the production patterns of PWA. Reliance on automatic C-NNLA scoring can provide eight benefits for clinicians and researchers:

1. Smoother transcription. Because CLAN uses normal English orthography as the input to automatic analysis, there is no need for hand coding of morphology in forms such as *waste/ed, stepmother/z,* or *can/'t* as required for input to NNLA and subsequent entry into SALT. Instead, the transcriber simply enters *wasted, stepmother's,* and *can't* in standard English orthography, and the morphological structure of these forms is analyzed automatically.
2. Sound playback. Because the CLAN editor provides direct playback from media, it is easy to replay individual utterances to maximize transcription accuracy.
3. Faster analysis. Automatic scoring of the individual measures in NNLA through C-NNLA greatly reduces the time required to analyze a transcript. This benefit becomes even more important when there are large numbers of transcripts to score because C-NNLA can analyze hundreds of transcripts in a few seconds.
4. Less demand for expertise. Many of the concepts needed for accurate NNLA scoring require a high level of understanding of concepts in linguistic analysis. To get this right, each researcher or clinician who might use NNLA must be trained thoroughly and have extensive linguistic knowledge in order to achieve good reliability. For analysis through C-NNLA, adherence to the requirements of linguistic training and knowledge is achieved during the act of programming without any further reliance on training of individual human analysts.
5. Spreadsheet output. The scores computed by C-NNLA can be output to spreadsheet formats suitable for further statistical analysis. This is an automatic process, whereas similar coding from NNLA to SALT and then from SALT to statistical analysis requires manual transfer of files.
6. Replicability. Automatic computation is thoroughly reliable and replicable because repeated runs of a computer program always produce the same result. This means that it is possible for researchers to replicate a published computerized analysis, as long as the input data and the analysis program are also made public. This is now easy to do through the version and data control methods in the AphasiaBank system. Ability to conduct such replications allows researchers to address the replication crisis facing the biomedical and social sciences (Munafò et al., 2017).
7. Database comparison. Results computed by C-NNLA for a given transcript can be automatically compared with similar transcripts for over 300 PWAs and over 200 control participants in AphasiaBank. This comparison will show how the PWA compares in terms of each of the component measures in NNLA, as well as dozens of other automatic CLAN analyses.
8. Facilitation of debugging and improvement. Errors and gaps in computerized analysis can be systematically diagnosed and corrected. Although no human or automatic system for spoken language can ever achieve 100% accuracy on all inputs, it is possible to continually improve the accuracy of automatic methods through the construction of larger training sets and the structuring of more accurate grammatical pattern detectors. Because CLAN is updated continually and because the updates are freely downloadable, researchers and clinicians are always able to take advantage of the most recent improvements.

Conclusions

The digital revolution has led to increasing automation in all aspects of our lives from self-driving cars to the Internet of Things. However, this revolution has barely begun to impact clinical practice for aphasia or research studies of spoken narratives. In this article, we show how computer analysis based on a series of NLP tools can begin

to improve our ability to characterize language in aphasia. Crucially, this work benefits from the groundwork of detailed linguistic analysis provided by the NNLA system. The research literature on agrammatic aphasia provided the motivation for each of the measures included in NNLA. Based on this groundwork, we have now successfully automated 50 of the 51 NNLA measures, with good agreement between manual and C-NNLA analysis methods studied earlier by Hsu and Thompson. This automation provides the eight benefits of automated analyses mentioned earlier: smoother transcription, sound playback, faster analysis, less demand for expertise, provision of spreadsheet output, replicability, facilitation of database comparison, and support for debugging and improvement.

Ongoing and future work will focus on improving, adding, and testing outcome measures in the new C-NNLA program and automating programs for other analysis systems, such as the Quantitative Production Analysis (Rochon et al., 2000; Saffran et al., 1989) and correct information units (Nicholas & Brookshire, 1993). We encourage the aphasia clinical research community's use, input, and feedback on these efforts in the interest of compiling data on psychometric properties of valid and reliable measures to be used for clinical and research purposes.

The transcripts used in this project, along with the output from C-NNLA (with participant demographics) are available at https://aphasia.talkbank.org/discourse/C-NNLA/. We continually add to the *Discourse Analysis* collection for the demonstration and testing of not only C-NNLA analyses but also Quantitative Production Analysis, correct information units, and other analyses. This collection constitutes a shared workspace of publicly available and fully analyzed data on agrammatic and fluent aphasia that could play a central role in addressing the need for core outcome sets for discourse (Armstrong, 2018; de Riesthal & Diehl, 2018; Dietz & Boyle, 2018; Kintz & Wright, 2018; Kurland & Stokes, 2018; Pritchard et al., 2017; Wallace et al., 2018; Whitworth, 2018), ways to test the psychometric properties of these measures (Dietz & Boyle, 2018; Pritchard et al., 2017), and ways to address the replicability crisis (Munafò et al., 2017).

Acknowledgments

This work was supported by National Institute on Deafness and Other Communication Disorders Grant R01-DC008524 (2007-2022, awarded to MacWhinney) and National Institute on Deafness and Other Communication Disorders Grant R01-DC01948 (1992–2022, awarded to Thompson).

References

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

Ballard, K. J., & Thompson, C. K. (1999). Treatment and generalization of complex sentence production in agrammatism. *Journal of Speech, Language, and Hearing Research, 42*(3), 690–707. https://doi.org/10.1044/jslhr.4203.690

Barbieri, E., Mack, J., Chiappetta, B., Europa, E., & Thompson, C. K. (2019). Recovery of offline and online sentence processing in aphasia: Language and domain-general network neuroplasticity. *Cortex, 120,* 394–418. https://doi.org/10.1016/j.cortex.2019.06.015

Bastiaanse, R., & Thompson, C. K. (2012). *Perspectives on agrammatism*. Psychology Press. https://doi.org/10.4324/9780203120378

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. (2018). Discourse measurement in aphasia research: have we reached the tipping point. *Aphasiology, 32*(4), 459–464. https://doi.org/10.1080/02687038.2017.1398803

Edwards, S. (2005). *Fluent aphasia*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486548

Faroqi-Shah, Y., & Thompson, C. K. (2007). Verb inflections in agrammatic aphasia: Encoding of tense features. *Journal of Memory and Language, 56*(1), 129–151. https://doi.org/10.1016/j.jml.2006.09.005

FitzGerald, N., Michael, J., He, L., & Zettlemoyer, L. (2018). *Large-scale qa-srl parsing*. arXiv preprint arXiv:1805.05377. https://doi.org/10.18653/v1/P18-1191

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, 28*(3), 245–288. https://doi.org/10.1162/089120102760275983

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

Jacobs, B. J., & Thompson, C. K. (2000). Cross-modal generalization effects of training noncanonical sentence comprehension and production in agrammatic aphasia. *Journal of Speech, Language, and Hearing Research, 43*(1), 5–20. https://doi.org/10.1044/jslhr.4301.05

Kertesz, A. (2006). *Western Aphasia Battery-Revised* (WAB-R). Pro-Ed.

Kim, M., & Thompson, C. K. (2004). Verb deficits in Alzheimer's disease and agrammatism: Implications for lexical organization. *Brain and Language, 88*(1), 1–20. https://doi.org/10.1016/S0093-934X(03)00147-0

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

Mack, J. E., Chandler, S. D., Meltzer-Asscher, A., Rogalski, E., Weintraub, S., Mesulam, M.-M., & Thompson, C. K. (2015). What do pauses in narrative production reveal about the nature of word retrieval deficits in PPA. *Neuropsychologia, 77,* 211–222. https://doi.org/10.1016/j.neuropsychologia.2015.08.019

Mack, J. E., & Thompson, C. K. (2017). Recovery of online sentence processing in aphasia: Eye movement changes resulting from treatment of underlying forms. *Journal of Speech, Language, and Hearing Research, 60*(5), 1299–1315. https://doi.org/10.1044/2016_JSLHR-L-16-0108

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

Meltzer-Asscher, A., & Thompson, C. K. (2014). The forgotten grammatical category: Adjective use in agrammatic aphasia. *Journal of Neurolinguistics, 30,* 48–68. https://doi.org/10.1016/j.jneuroling.2014.04.001

Miller, J., & Chapman, R. (1983). *SALT: Systematic Analysis of Language Transcripts, user's manual*. University of Wisconsin Press.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1,* Article number 0021. https://doi.org/10.1038/s41562-016-0021

Nicholas, L., & Brookshire, R. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36*(2), 338–350. https://doi.org/10.1044/jshr.3602.338

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71–105. https://doi.org/10.1162/0891201053630264

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., Berndt, R., & Schwartz, M. (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., Berndt, R., & Schwartz, M. (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

Thompson, C. K., Ballard, K. J., Tait, M. E., Weintraub, S., & Mesulam, M. (1997). Patterns of language decline in non-fluent primary progressive aphasia. *Aphasiology, 11*(4–5), 297–321. https://doi.org/10.1080/02687039708248473

Thompson, C. K., Cho, S., Hsu, C.-J., Wieneke, C., Rademaker, A., Weitner, B. B., Mesulam, M. M., & Weintraub, S. (2012). Dissociations between fluency and agrammatism in primary progressive aphasia. *Aphasiology, 26*(1), 20–43. https://doi.org/10.1080/02687038.2011.584691

Thompson, C. K., Meltzer-Asscher, A., Cho, S., Lee, J., Wieneke, C., Weintraub, S., & Mesulam, M. (2013). Syntactic and morphosyntactic processing in stroke-induced and primary progressive aphasia. *Behavioural Neurology, 26*(1–2), 35–54. https://doi.org/10.1155/2013/749412

Thompson, C. K., Shapiro, L. P., Li, L., & Schendel, L. (1995). Analysis of verbs and verb-argument structure: A method for quantification of aphasic language production. *Clinical Aphasiology, 23,* 121–140.

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.

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

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

Appendix A

Summary of NNLA Measures and Agreement Between Hand-Coded and Automated Coding (CLAN and C-NNLA)

No.	NNLA measures	Good agreement between CLAN and NNLA in Hsu & Thompson	Good agreement between C-NNLA and NNLA	Some disagreement between C-NNLA and NNLA
	General language measures (*n* = 3)			
1	MLU	√		
2	Number of utterances	√		
3	Number of words	√		
	Utterance and sentence-level measures (*n* = 6)			
4	Proportion of sentences		√	
5	Proportion of sentences with correct syntax and semantics		√	
6	Proportion of sentences with flawed syntax		√	
7	Proportion of sentences with flawed semantics		√	
8	Sentence complexity ratio		√	
9	Number of embedded clauses per sentence		√	
	Lexical level measures (*n* = 23)			
10	Total number of open-class words	√		
11	Proportion of open-class words over all words	√		
12	Total number of closed-class words	√		
13	Proportion of closed-class words over all words	√		
14	Open-to-closed word ratio	√		
15	Total nouns	√		
16	Proportion of nouns over all words	√		
17	Total verbs	√		
18	Proportion of verbs over all words		√	
19	Noun-to-verb ratio		√	
20	Total adjectives	√		
21	Total adverbs	√		
22	Total determiners	√		
23	Total pronouns	√		
24	Total auxiliaries	√		
25	Total conjunctions		√	
26	Total modals		√	
27	Total prepositions	√		
28	Total negation markers	√		
29	Total infinitival markers	√		
30	Total quantifiers	√		
31	Total *wh*-words	√		
32	Total particles			√
	Bound morpheme-level measures (*n* = 13)			
33	Total comparative suffixes	√		
34	Total superlative suffixes	√		
35	Total possessive markers	√		
36	Total regular plural markers	√		
37	Total irregular plural forms	√		
38	Total regular past tense markers	√		
39	Total third-person present tense markers	√		
40	Total irregular past tense markers	√		
41	Total regular perfect aspect markers	√		
42	Total irregular perfect participles	√		
43	Total progressive aspect markers	√		
44	Proportion of correct regular inflection		√	
45	Proportion of correct irregular inflection		√	
	Verb argument–level measures (*n* = 6)			
46	1-place verbs over all verbs			√
47	Proportion 1-place verbs with correct argument structure			√
48	2-place verbs over all verbs			√
49	Proportion 2-place verbs with correct argument structure			√
50	3-place verbs overall verbs			√
51	Proportion 3-place verbs with correct argument structure			√

Note. NNLA = Northwestern Narrative Language Analysis; CLAN = Computerized Language Analysis; C-NNLA = CLAN command to compute the full set of NNLA measures; MLU = mean length of utterance.

Appendix B

Transcription and Coding Rules

For accurate computation of outcome measures according to NNLA rules, the following CHAT conventions must be followed.

1. Exclusions. NNLA rules call for a number of exclusions in computing outcome measures. Many of these exclusions are automatic. CLAN already excludes repetitions marked with [/], revisions marked with [//], fillers transcribed with &-, and fragments transcribed with &+. The C-NNLA command also automatically excludes the following conjunctions when they are used in the beginning of an utterance: *and, but, or, then, so, well, and then, but then, and so*. To exclude other words from C-NNLA analysis, as per the NNLA manual (e.g., interjections, comments), transcribers must manually insert the [e] code after the word(s) to be excluded. For entire utterances to be excluded, use the [+ exc] code after the final punctuation. Here are three examples:

```
*PAR: the prince says oh [e] it is you.
*PAR: <I think> [e] her name was Cinderella.
*PAR: I can't do this. [+ exc]
```

2. Utterance level coding. Two codes are needed to mark grammatically flawed [+ gram] and semantically flawed [+ sem] utterances, as per the NNLA manual.

```
*PAR: she was a really nice dress. [+ sem]
*PAR: he has a wonderful time. [+ sem] (talking about Cinderella)
*PAR: looking at the clock. [+ gram]
*PAR: it is just stepmother and three stepsisters. [+ gram]
```

3. Morphological error coding. The error-coding chapter in the CHAT manual (https://talkbank.org/manuals/CHAT.pdf), Chapter 18, provides word-level error codes that can be used in CHAT files for phonological, semantic, neologistic, and morphological errors. For accurate computation of several morphological outcome measures in C-NNLA, it is important to mark the morphological errors, as in the following examples. The Morphological Errors section in the CHAT manual lists the full set of error codes for the range of morphological errors (e.g., missing morphemes, superfluous morphemes, substituted morphemes). Note, that the target word is entered in square brackets with a single colon, followed by the word-level error code.

```
*PAR: both was [: were] [* m:vsg:a] very mean. [+ gram]
*PAR: it felled [: fell] [* m:+ed] out. [+ gram]
*PAR: she was push [: pushing] [* m:0ing] pins and needles. [+ gram]
```