Automatic Computation of SUGAR, QPA, and NNLA Profiles using CLAN

Davida Fromm and Brian MacWhinney, Carnegie Mellon University, Pittsburgh, PA

Objective and Rationale

Measurement of spoken language is essential for assessment and management in our field. BUT transcribing and analyzing spoken language require extensive time and expertise.

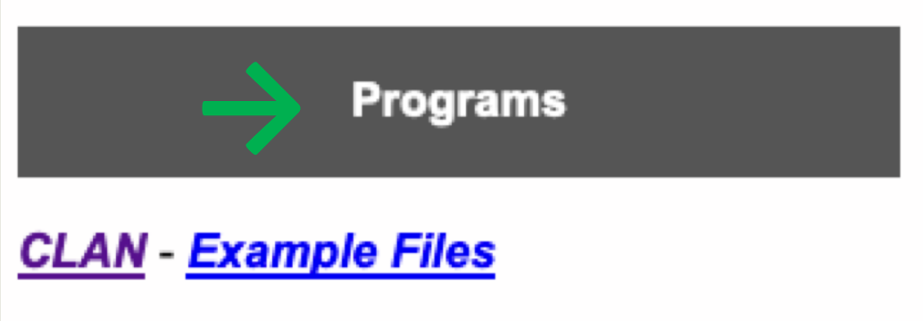
Purpose: to demonstrate new automated CLAN (Computerized Language Analysis) (MacWhinney, 2000) commands for analysis of language transcripts using three profiling systems:

- Sampling Utterances and Grammatical Analysis Revised (SUGAR; Pavelko & Owens, 2017)
- Quantitative Production Analysis (QPA; Berndt et al., 2000)
- Northwestern Narrative Language Analysis (NNLA; Thompson et al., 1995)

CLAN

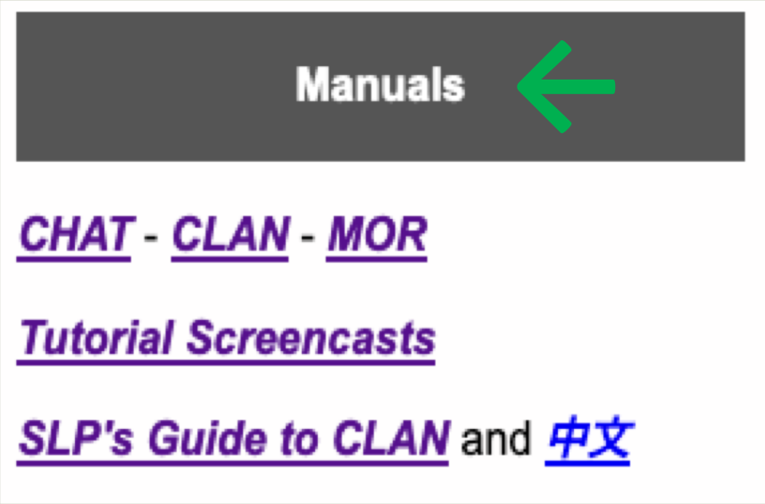
CLAN is a package of automated analysis programs that work on CHAT transcripts. You can transcribe in CHAT or you can convert any text file to a CHAT transcript.

CLAN is freely downloadable for Mac or PC from the TalkBank website: https://talkbank.org/ :


Programs

CLAN - Example Files

CLAN and CHAT manuals and tutorials are at the website too:


Manuals

CHAT - CLAN - MOR

Tutorial Screencasts

SLP's Guide to CLAN and 中文

A CHAT transcript looks this:
(see handout for complete transcript file with headers)

```
*PAR:  the ladder [: stool] [*] is tipping over.
*PAR:  it looks like he might fall.
*PAR:  &-um there's &-um two [/] two [/] two things to drink out_of on the
       table and another plate.
```

Then you run the MOR command on the transcript file and within seconds it looks like this:

```
*PAR:   the ladder [: stool] [*] is tipping over.
%mor:   det:art|the n|stool aux|be&3S part|tip-PRESP adv|over .
%gra:    1|2|DET 2|4|SUBJ 3|4|AUX 4|0|ROOT 5|4|JCT 6|4|PUNCT
*PAR:   it looks like he might fall.
%mor:   pro:per|it v|look-3S conj|like pro:sub|he mod|might v|fall .
%gra:   1|2|SUBJ 2|0|ROOT 3|6|LINK 4|6|SUBJ 5|6|AUX 6|2|COMP 7|2|PUNCT
*PAR:  &-um there's &-um two [/] two [/] two things to drink out_of on the
       table and another plate.
%mor:  pro:exist|there~cop|be&3S det:num|two n|thing-PL inf|to v|drink prep|out_of
       adv|on det:art|the n|table coord|and qn|another n|plate.
%gra:  1|2|SUBJ 2|0|ROOT 3|4|QUANT 4|2|PRED 5|6|INF 6|4|XMOD 7|6|JCT 8|7|JCT
      9|10|DET 10|8|POBJ 11|10|CONJ 12|13|QUANT 13|11|COORD 14|2|PUNCT
```

Presto -- these morphological and grammatical relations tiers contain all the information necessary to do all of these analyses automatically !

SUGAR

SUGAR was developed to help clinicians do language sample analysis efficiently. SUGAR provides rules and step-by-step methods for:

- eliciting a language sample
- transcribing using a word processing program (e.g., Word)
- computing 4 measures from the first 50 utterances of a 10-minute sample:
  1. mean length of utterance in morphemes (MLU-S)
  2. total number of words (TNW)
  3. words per sentence (WPS)
  4. clauses per sentence (CPS)

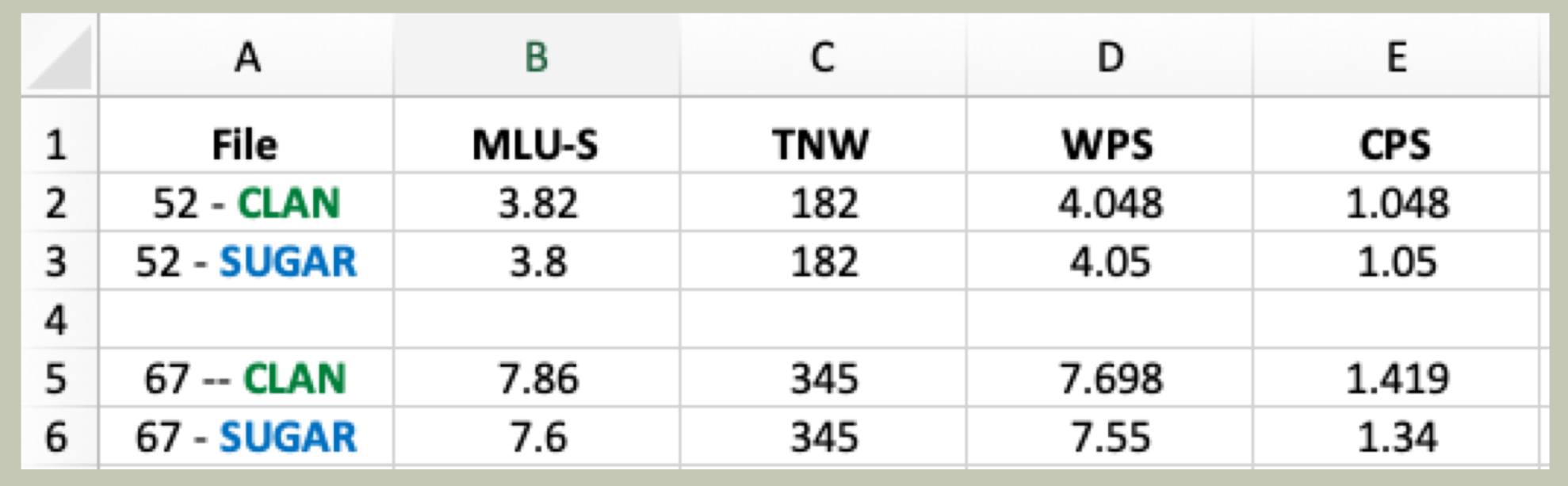
CLAN can reliably compute the 4 outcome measures **automatically** using 3 simple commands and avoiding approximately 10 steps in the SUGAR method (see handout).

1. **text2chat +c1 +c2 *.txt** -- this converts a text file to a CHAT file
2. **mor *.cha** – this creates the morphological tier
3. **sugar +t*txt *.cha** -- this runs the SUGAR command

SUGAR, cont.

Output -- CLAN spreadsheet results for SUGAR practice samples

CLAN results in green
SUGAR results in blue



	A	B	C	D	E
1	File	MLU-S	TNW	WPS	CPS
2	52 - CLAN	3.82	182	4.048	1.048
3	52 - SUGAR	3.8	182	4.05	1.05
4					
5	67 -- CLAN	7.86	345	7.698	1.419
6	67 - SUGAR	7.6	345	7.55	1.34

C-QPA

The QPA was developed to quantify sentence production in aphasia, emphasizing grammatical structure. It provides rules and step-by-step methods for:

- eliciting a language sample
- transcribing the language sample
- counting words and segmenting utterances in the language sample
- scoring the utterances in a spreadsheet to compute:
  1. 21 measures for an Analysis Worksheet
  2. 38 measures for a Summary Worksheet

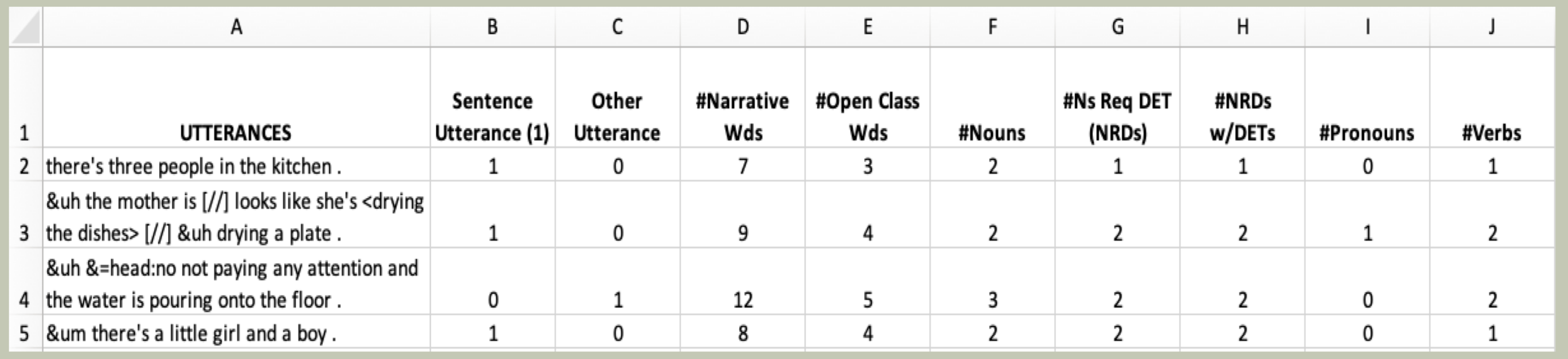
Transcribe the sample in CHAT, following some CHAT conventions to ensure proper coding and scoring (see handout for further info).

Run one CLAN command to compute all of the outcome measures and create both worksheet outputs as spreadsheets.

c-qpa +t*par filename.cha

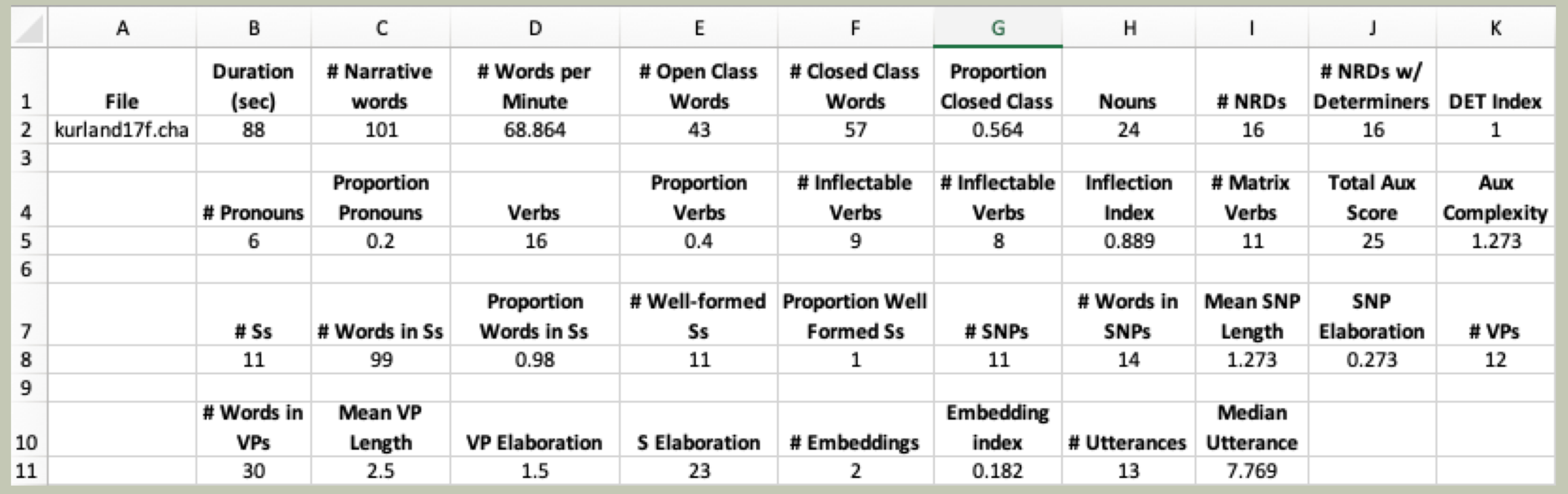
Output -- Analysis spreadsheet (first 4 lines of transcript) and
Summary spreadsheet for cookie theft picture description, adult with aphasia.

Analysis spreadsheet: 21 measures



	A	B	C	D	E	F	G	H	I	J
1	UTTERANCES	Sentence Utterance (1)	Other Utterance	#Narrative Wds	#Open Class Wds	#Nouns	#Ns Req DET (NRDs)	#NRDs w/DETs	#Pronouns	#Verbs
2	there's three people in the kitchen .	1	0	7	3	2	1	1	0	1
3	&uh the mother is [//] looks like she's <drying the dishes> [//] &uh drying a plate .	1	0	9	4	2	2	2	1	2
4	&uh &=head:no not paying any attention and the water is pouring onto the floor .	0	1	12	5	3	2	2	0	2
5	&um there's a little girl and a boy .	1	0	8	4	2	2	2	0	1

Summary Spreadsheet: 38 measures



	A	B	C	D	E	F	G	H	I	J	K
1	File	Duration (sec)	# Narrative words	# Words per Minute	# Open Class Words	# Closed Class Words	Proportion Closed Class	Nouns	# NRDs	# NRDs w/ Determiners	DET Index
2	kurland17f.cha	88	101	68.864	43	57	0.564	24	16	16	1
3											
4		# Pronouns	Proportion Pronouns	Verbs	Proportion Verbs	# Inflectable Verbs	# Inflectable Verbs	Inflection Index	# Matrix Verbs	Total Aux Score	Aux Complexity
5		6	0.2	16	0.4	9	8	0.889	11	25	1.273
6											
7		# Ss	# Words in Ss	Proportion Words in Ss	# Well-formed Ss	Proportion Well Formed Ss	# SNPs	# Words in SNPs	Mean SNP Length	SNP Elaboration	# VPs
8		11	99	0.98	11	1	11	14	1.273	0.273	12
9											
10		# Words in VPs	Mean VP Length	VP Elaboration	S Elaboration	# Embeddings	Embedding index	# Utterances	Median Utterance		
11		30	2.5	1.5	23	2	0.182	13	7.769		

NNLA

The NNLA was developed to analyze discourse in individuals with agrammatic aphasia and has been used with individuals with all types of aphasia as well as PPA and dementia. It provides rules for:

- transcription
- 5 levels of manual coding for each utterance -- utterance, sentence, lexical, bound morpheme, and verb argument structure.

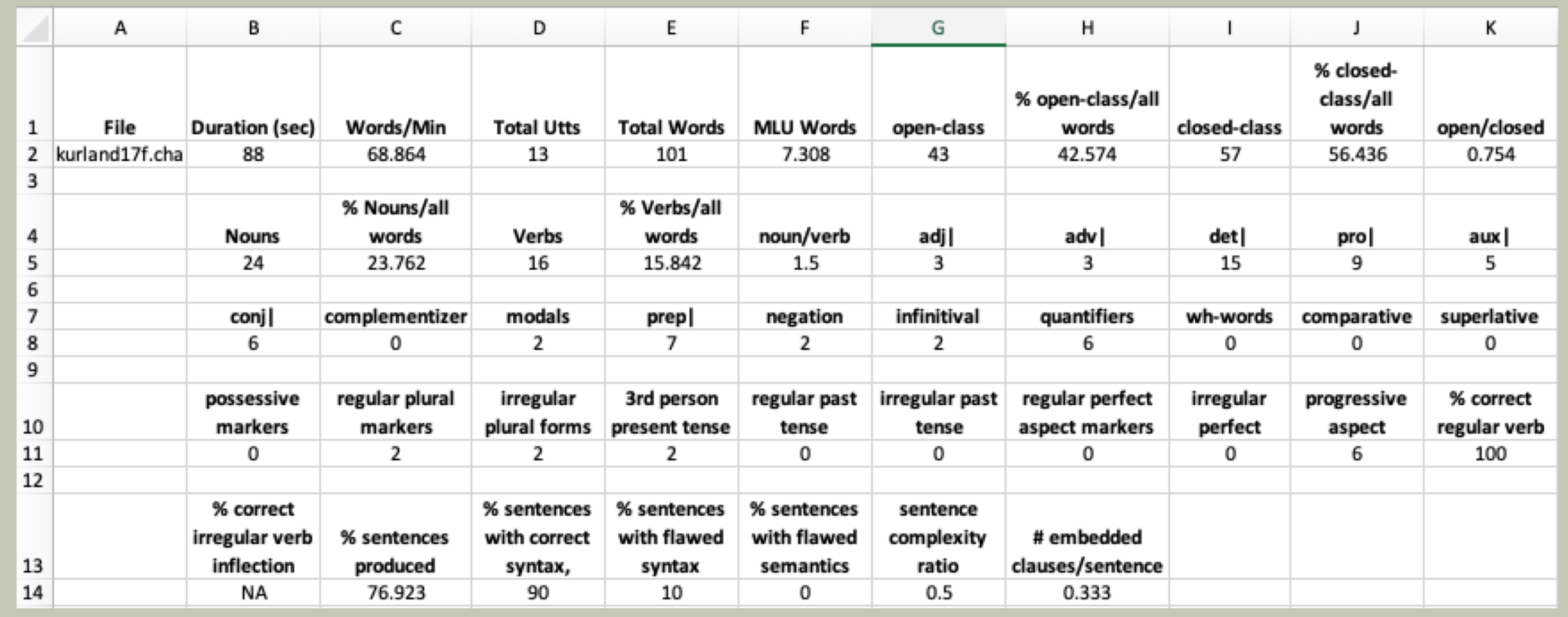
Transcribe the sample in CHAT, following some CHAT conventions to ensure proper coding and scoring (see handout for further info).

Run one CLAN command to automatically compute 44 outcome measures and create a spreadsheet with the results. (Additional measures are being programmed and tested.)

c-nnla +t*par filename.cha

C-NNLA, cont.

Output – CLAN spreadsheet results for utterance, sentence, lexical, and bound morpheme level outcome measures.



	A	B	C	D	E	F	G	H	I	J	K
1	File	Duration (sec)	Words/Min	Total Utts	Total Words	MLU Words	open-class	% open-class/all words	closed-class	% closed-class/all words	open/closed
2	kurland17f.cha	88	68.864	13	101	7.308	43	42.574	57	56.436	0.754
3											
4		Nouns	% Nouns/all words	Verbs	% Verbs/all words	noun/verb	adj|	adv|	det|	pro|	aux|
5		24	23.762	16	15.842	1.5	3	3	15	9	5
6											
7		conj|	complementizer	modals	prep|	negation	infinitival	quantifiers	wh-words	comparative	superlative
8		6	0	2	7	2	2	6	0	0	0
9											
10		possessive markers	regular plural markers	irregular plural forms	3rd person present tense	regular past tense	irregular past tense	regular perfect aspect markers	irregular perfect	progressive aspect	% correct regular verb
11		0	2	2	2	0	0	0	0	6	100
12											
13		% correct irregular verb inflection	% sentences produced	% sentences with correct syntax,	% sentences with flawed syntax	% sentences with flawed semantics	sentence complexity ratio	# embedded clauses/sentence			
14		NA	76.923	90	10	0	0.5	0.333			

Advantages of Automated Analysis with CLAN

1. Smoother transcription

   CLAN uses normal English, so no hand-coding of morphology is needed.

   Examples:
   CLAN – wasted **SUGAR** – wast ed **NNLA and SALT** – waste/ed
   CLAN – don't **SUGAR** – do n't **NNLA and SALT** – do/n't
2. Faster analysis

   All measures are automatically counted and computed immediately.

   CLAN programs can analyze hundreds of transcripts in a few seconds.
3. Less demand for expertise

   Researchers and clinicians require thorough training in linguistic theory and analysis rules to reliably use QPA and NNLA. With CLAN, the rules are written into the programs and achieved without relying on human training.
4. Spreadsheet output

   Scores computed by CLAN are automatically output to spreadsheet format for further statistical analysis or display purposes. No reformatting or reentry of numbers is necessary.
5. Replicability

   Repeated runs of the CLAN program will always produce the same result.

   No need to do inter-rater and intra-rater reliability for coding and computations.
6. Database comparison

   Results can be automatically compared with hundreds of transcripts from relevant comparison groups – e.g., children, adults, people with aphasia
7. Facilitation of debugging and improvement

   CLAN is updated continually, improving the accuracy of automatic methods and adding new features.

References

Berndt, R. S., Wayland, S., Rochon, E., Saffran, E. and Schwartz, M. 2000. *Quantitative production analysis: A training manual for the analysis of aphasic sentence production*, Philadelphia: Psychology Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Pavelko, S. L., & Owens Jr, R. E. (2017). Sampling Utterances and Grammatical Analysis Revised (SUGAR): New normative values for language sample analysis measures. *Language, Speech, and Hearing Services in Schools*, *48*(3), 197-215.

Thompson, C. K. (2015). *Northwestern Narrative Language Analysis (NNLA) Theory and Methodology.* Retrieved from http://anr.northwestern.edu/wp-content/uploads/2016/09/NNLA_REVISED.Fall2015.pdf

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

Acknowledgments

This work was funded by NIH-NIDCD grant R01-DC008524 (20017-2022).