A “Verbal Thermometer” for Assessing Neurodegenerative Disease: Automated Measurement of Pronoun and Verb Ratio from Speech

William Jarrold[1], Adrià Rofes[2], Stephen Wilson[3], Peter Pressman[4], Edward Stabler[5], and Marilu Gorno-Tempini[6]

[1]Independent Researcher/Consultant (william.jarrold@gmail.com)
[2]Department of neurolinguistics at the University of Groningen
[3]Department of Hearing and Speech Sciences, Vanderbilt University Medical Center
[4]Department of Neurology, University of Colorado Denver
[5]Department of Linguistics, UCLA
[6]Memory and Aging Center, University of California San Francisco

Abstract— **Clinicians often use speech to characterize neurodegenerative disorders. Such characterizations require clinical judgment, which is subjective and can require extensive training. Quantitative Production Analysis (QPA) can be used to obtain objective quantifiable assessments of patient functioning. However, such human-based analyses of speech are costly and time consuming. Inexpensive off-the-shelf technologies such as speech recognition and part of speech taggers may avoid these problems. This study evaluates the ability of an automatic speech to text transcription system and a part of speech tagger to assist with measuring pronoun and verb ratios, measures based on QPA. Five participant groups provided spontaneous speech samples. One group consisted of healthy controls, while the remaining groups represented four subtypes of frontotemporal dementia. Findings indicated measurement of pronoun and verb ratio was robust despite errors introduced by automatic transcription and the tagger and despite these off-the-shelf products not having been trained on the language obtained from speech of the included population.**

Clinical relevance **Linguistic differences in individuals with neurodegenerative disease can be subtle. Automated measurement of pronoun and verb ratio (PR and VR) showed promise despite the early stage of pipeline development using off-the-shelf tools. Intraclass correlations between automatic versus human measures of PR and VR were in the moderate to excellent range as measured by intra-class correlation.**

I. INTRODUCTION

A. Traditional Speech-Based Assessment

Spontaneous speech tasks are used to assess and characterize language production in people with neurological deficits such as stroke [1], brain tumors [2], and neurodegenerative diseases [3]. Patients are typically asked to describe a picture, respond to an open-ended question, or tell a story [4]. Spontaneous speech tasks are attractive because sufficient data can be collected relatively quickly (about 10 minutes per sample). These tasks are comparable to traditional pen-and-paper testing [5] and may be less prone to test/retest effects [6]. Further, a wealth of recommendations for data acquisition and processing exist [1], [4].

One of the disadvantages of spontaneous speech tasks is the time and expertise required for manual transcription and part-of-speech annotation. In our experience, a typical 10-minute speech sample requires at least 240 minutes of work by a language expert, and verification by another expert is often required.

B. Automation Necessitates Evaluation

A relatively large number of studies have evaluated automated speech-based diagnosis of neurodegenerative disorders using a technology pipeline that includes automated acoustic-prosodic feature extraction, automated transcription, lexical feature extraction, and machine learning [7], [8], [9], [10], [11], [12], [13]. Toward scaling up these end-to-end diagnostic assistance systems, a narrower focus on evaluating sub-components becomes important.

This paper focuses on the performance of two such sub-components: (a) **automated transcription (AT)** and (b) **(automated) part of speech tagging (POST)**. The latter is a specific aspect of lexical feature extraction that involves determining the part of speech (POS) (e.g., noun, verb) for each word in a text. In this study, we apply these components to analyze the spoken language of patients and healthy controls.

Relatively few studies have directly evaluated automated versus human analyses of lexical features (e.g., Fraser [9], Hsu et al. [14]), and existing work has involved relatively few participants, with only two or three diagnostic/control groups. This study builds upon such work by focusing exclusively on POS features, and it involves more than double the number of participants in previous studies. Among the five groups in the current study, one pertains to logopenic primary progressive aphasia, a condition for which computational approaches have only begun to be used.

C. Groups: Four Disease Variants plus Healthy Controls

This study includes **healthy controls (HC)** and individuals with **behavioral variant frontotemporal dementia (bvFTD)** or one of three variants of primary progressive aphasia (PPA): (i) **semantic variant PPA (svPPA)**, (ii) **non-fluent variant PPA (nfvPPA)**, and (iii) **logopenic variant PPA (lvPPA)**. BvFTD involves symptoms such as social disinhibition and impulsivity [15], while PPA is a neurodegenerative syndrome that predominantly damages language

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

processing before the emergence of other symptoms (e.g., attention, memory, executive functioning). Thus, it is especially important to study automated clinical speech analytics in this population.

D. Diagnostic Multi-Dimensionality

Much prior work has focused on predicting diagnosis. However, the effect of disease on speech is a multi-dimensional phenomenon, and clinicians need more than just a diagnostic category or label for their patients. A given disorder has a wide range of severity levels, as well as variable strength and weakness profiles, and treatment and prognosis can heavily depend on an individual's specific situation. Therefore, an important aspect of this work is not solely to evaluate the performance of POS taggers to assist with classification but also to evaluate their ability to reliably and accurately locate a given patient within a multidimensional symptomotological space.

Clinicians have developed such an objective, quantifiable, and multi-dimensional characterization of speech production in aphasia in the clinical procedure quantitative production analysis (QPA) [1], [16]. QPA involves lexical content measures, syntactic structure and complexity, speech rate and errors, as well as fluency disruptions (e.g., false starts).

In this paper, we focus on the automatic measurement of two important QPA variables: *pronoun ratio* (PR) and *verb ratio* (VR). Based on the frequencies of **pronouns (P)**, **nouns (N)**, and **verbs (V)** in a transcript, PR and VR are defined as follows:

1) **PR : P / (P + N)**
2) **VR : V / (N + V)**

These two measures are chosen because they exhibit several differences as a function of diagnostic group [17] and are easy to compute directly from POS tags.

E. Off-the-Shelf and Baseline Perspective

This initial study is intended to establish a baseline level of performance for future improvement. This "baseline" is defined by all components being freely available today, with no adaptation or customization needed for the task at hand. Typical automatic speech recognition systems, meant for dictation, do not predict punctuation[1] (the speaker needs to say "period" or "comma" to add punctuation). Thus, transcripts fed to the tagger lack periods and commas.[2] They also ignore fillers, such as *um's* and *ah's*. Fillers were transcribed in the current dataset, but we removed them to best approximate current off-the-shelf AT.

F. Focal Questions

We evaluated the ability of AT and POST to automatically reproduce human-annotator based findings of a prior paper [17] regarding the QPA dimensions PR and VR and the ability to detect previously identified differences between groups. We also evaluated POST output when given human transcription (HT) versus AT.

G. Forces Hindering vs. Enhancing Performance

1) Hinderance: POSTs are typically trained on journalistic prose (see II-C). Such training data may not be well-matched to patients' speech due to the language patterns that occur in spontaneous speech, and the lack of neurologically impaired training data likely hinders AT and POST. In addition, the sometimes noisy hospital environment may further undermine AT.

2) Enhancement: QPA measures are hypothesized to be relatively robust for two reasons. First, typical AT and POST measures require assigning the correct sequence of tokens (words, POS tags, respectively) to a speech sample. However, pronoun and verb ratio are frequency based; that is, they are determined by mere *counts* of nouns, verbs, and pronouns. Mis-orderings of POS tags do not hurt frequency counts as long as intruding or dropped tokens do not occur.

A second robustness enhancer involves granularity. Typical POST involves 15 or more POS labels, but QPA requires a relatively coarse-grained classification choice: N, P, V, or other. In addition, AT has to choose the correct word from a vocabulary containing thousands of words. For example, POST suffers if a superlative adverb is misclassified as a comparative, but the QPA measures would not suffer. If AT misrecognizes "cat" as "hat," QPA still correctly counts the word as a noun.

In sum, this paper is aimed at characterizing how these two forces play out in a system that can produce two measures of speech important for neurological diagnoses. In essence, this characterization may serve as a "verbal thermometer."

II. Method

A. Research Participants

Participants were recruited and assessed in the neurology department of a major medical university. All participants gave written informed consent, and the study was approved by the institutional review board. See Wilson [17] for details of participants' comprehensive neurological history and examination, neuropsychological testing, and neuroimaging. Data were shared via data use agreement.

In total, 70 participants were included: 60 with a neurodegenerative disease and 10 HCs. The patient subgroups included individuals with various prominent language impairments: 11 with lvPPA, 25 with svPPA, 14 with nfvPPA, and 10 with bvFTD. Age, handedness, and level of education were not significantly different between groups, but there was a statistically significant difference for sex ($p < 0.01$).

B. Speech Task, Human Transcription, and Annotation

A standardized speech task was administered and recorded as part of previously published research [17]; specifically, the participants completed the picnic picture description component of the Western Aphasia Battery (WAB) [19]. This task involved a clinician prompting participants along the lines of *Take a look at this picture, tell me what you see, and try to talk in sentences.*

Transcription and annotation with POS tags were done by the consensus of two experts: a post-doctoral fellow

[1] See Fraser et al. [18] for research into automated sentence boundary detection in impaired speech.

[2] However, period removal seems inconsequential (see Sec III-A)

TABLE I
QPA Normalization of Expert and POST Tags THE UTTERANCE *sitting down and eating* IS TAGGED AND NORMALIZED TO THE FOUR-WAY DISTINCTION NEEDED TO COMPUTE PR AND VR. V, VERB; A, ADVERB; CONJ, CONJUNCTION; VBG, VERB (PRESENT PARTICIPLE OR GERUND); RP, PARTICLE; JJ, ADJECTIVE OR NUMERAL (ORDINAL).

Word	Expert tag (normalized tag)	NLTK Tag (normalized tag)	Same?
sitting	V (verb)	VBG (verb)	Yes
down	A (other)	RP (other)	Yes
and	Conj (other)	CC (other)	Yes
eating	V (verb)	JJ (other)	No

experienced in linguistic fieldwork and aphasic speech and a bachelor's level linguist. The former re-checked the latter's work. If a word was distorted but still intelligible, it was transcribed. Human annotations of fillers (e.g., *umm*, *ah*, *uhh*) and false starts were deleted because most off-the-shelf AT systems automatically remove them. Recordings and transcripts are not publishable to remain within the scope of participants' consent and to avoid violating their privacy. The complete corpus contained 9824 word tokens. Word count did not differ by diagnostic group except for nfPPA, which exhibited fewer words (mean [sd]) than HC and svPPA ($p<0.05$): nfPPA, 85.9 (39.7); svPPA, 145.9 (64.6); lvPPA, 117.7 (58.5); bvFTD, 98.3 (25.3); HC, 156.4 (52.4).

C. Part of Speech Tagger

For the POST, we used the default tagger in the Python-based Natural Language Toolkit (NLTK) [20], invoked by calling the `pos_tag` method. For example, `nltk.pos_tag(['He', 'sat'])` automatically tags the sentence *He sat*. The default tagger is an averaged perceptron tagger that predicts the POS of a word based on the preceding and following words, their tags, and their prefixes and suffixes [21].[3] This NLTK tagger uses a model trained on Wall Street Journal sections of Ontonotes 5 [22].

D. Tag Normalization

The QPA measures (i.e., PR and VR) require a set of tags that are much more coarse grained than typical POSTs provide. We thus mapped the 45 POS tags used by NLTK and the 20 tags used by previous annotators [17] to a set of only four tags: *noun*, *pronoun*, *verb*, and *other*. Thus, adjectives, determiners, and so forth fell into the "other" category. Auxiliaries were also normalized to *other* because Wilson [17] did not consider them part of VR.

In Table I, we demonstrate how the utterance *sitting down and eating* was tagged by a human expert versus a machine and how the tags were respectively normalized. The word *eating* illustrates a case in which the normalized machine-generated tag fails to match the normalized expert annotation.

E. Fully Automated System with Automated Transcription

To evaluate an end-to-end system, participant recordings were fed to Google Cloud AT [4]. Some editing was required to delete extraneous snippets of speech that were not part of the reference transcriptions (e.g., the clinician describing the task to the participant). Smaller pieces of extraneous speech remained due to the meticulous labor required to remove them and the poor performance of automated diarization. AT was fed to the POST. Tags were normalized as described above, resulting in frequencies of the normalized tags (i.e., N, P, V) that enabled computation of PR and VR.

F. Statistical Analysis

Our **first goal** in these analyses was to measure POST performance isolated from AT error (i.e., using HT rather than AT as input). One prong of this POST evaluation on HT input involved traditional POST measures: *per token accuracy*, which quantifies agreement between POST and expert for each word in sequence, and confusion matrices, which quantify error for each type of tag.

A second prong of goal one evaluated the POST's ability to measure PR and VR based on HT. These values were compared with PR and VR obtained from expert annotators via intra-class correlation (ICC) (see III-B for results).

The third prong evaluated how accurately the POST identified already known group-related differences in QPA. In particular, we examined the ability of the tagger to specifically and completely detect all significant differences in PR and VR that human experts [17] found to be dependent on group. The more it is able to detect these known findings, the more potential it shows in ability to identify novel group based differences (for results see III-C).

This third prong involved mirroring the planned statistical analysis done by [17]. Specifically, the independent variable was the participant group and the dependent variables were PR and VR. Following [17], we performed three types of tests: (a) an omnibus test for effect of group, (b) a comparison of each of the four patient groups to HC group, and (c) a comparison of each PPA group (svPP, lvPPA, nfvPPA) to the others. All comparisons are specified in Table IV (for the effect on PR) and V (for the effect on VR).

Given the multiple statistical comparisons, corrections to significance measures were required. In Wilson [17], p-values were corrected with the default single-step procedure used in the R program glht for ANOVAs [23]. In our study, we corrected p-values for multiple comparisons via the Tukey Honest Significant Difference adjustment [24], using the R Statistical Package (version 3.6.2 (2019-12-12).

The **second goal** was to measure end-to-end performance, that is when AT is fed to POST. To this end, analyses analogous to those in the first goal were performed. However, regarding the first prong, tag confusions were not computed due to the lack of word-for-word correspondence between HT and AT. Accuracy of the transcription was measured

[3] Source code for the tagger is published. See https://www.nltk.org/_modules/nltk/tag.html and its dependencies.

[4] https://cloud.google.com/speech-to-text/docs/async-recognize

TABLE II

Confusions Between Normalized Tags: Expert versus POST applied to HT Tags. PRO, PRONOUNS; ACC, BALANCED ACCURACY.

		Expert				
		other	noun	pro	verb	acc
POST	other	4427	122	131	13	0.90
	noun	126	1699	97	78	0.94
	pro	58	0	634	0	0.86
	verb	473	39	4	1383	0.94

TABLE III

Per Tag Accuracy (ACC) AVERAGED ACROSS ALL PARTICIPANTS WAS IN THE LOW 90% RANGE. PARTICIPANT GROUPS ARE DIVIDED INTO TWO LEVELS: ∼HC OR <HC DEPENDING ON WHETHER THAT VARIANT'S CONFIDENCE INTERVAL OVERLAPPED WITH OR WAS BELOW THAT OF HC.

	Participant Group	Mean Acc	95% Conf Interval	
			Lower	Upper
	All	0.92	0.91	0.92
∼HC	HC	0.94	0.93	0.95
	bvFTD	0.94	0.92	0.95
	nfvPPA	0.94	0.92	0.95
< HC	lvPPA	0.91	0.89	0.92
	svPPA	0.90	0.89	0.91

as part of the first prong. Corresponding to prong two, the performance of the end-to-end system was assessed by computing (a) the ICC between **automated pronoun ratio (APR)** and **human pronoun ratio (HPR)** as well as (b) ICC between **automated noun ratio (ANR)** and **human noun ratio (HNR)**.

III. RESULTS

A. Per-Token Accuracy (Comparing Expert vs POST on HT)

Addressing the first prong of the first goal (see II-F; see Table II for tagger confusions), we were only interested in confusions between N, V, P (post-normalization tags) and everything else ("other") because the two QPA dimensions of interest in this paper, PR and VR, only involve pronoun, noun, and verb frequencies. P is the most frequently confused tag because its balanced accuracy is the lowest.

Overall accuracy was 92% (See Table III). Per-token accuracy was significantly affected by group (see Table III), and non-overlapping 95% confidence intervals indicate that the accuracy for svPPA and lvPPA were below that of the other groups.

B. Pronoun and Verb Ratios Based on HT

P, N, and V frequencies were used to compute PR and VR. A scatterplot showing the correlation between POST and Expert measurement of PR is shown in Figure 3. It addresses the second prong of the first goal as defined in II-F). The ICC between POST- and Expert-based measurement of PR was in the excellent range (r=0.88 with a 95% confidence interval from 0.81 to 0.92). ICC for VR (see Table 4) was also in the excellent range (r=0.90 with a 95% confidence from 0.84 to 0.94). (See Cicchetti [25] for establishment of ICC qualitative ranges of poor, excellent, etc.) The form of ICC used is the most strict - the function 'ICC1' in R statistical package, that is, a one-way random effects, absolute agreement, single rater/measurement form of ICC, form A-1 according to the nomenclature of McGraw and Wong [26].

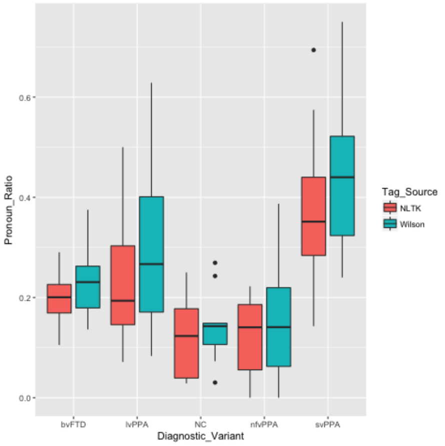

Fig. 1. Box Plots of Pronoun Ratio as a function of Group and Tagger.

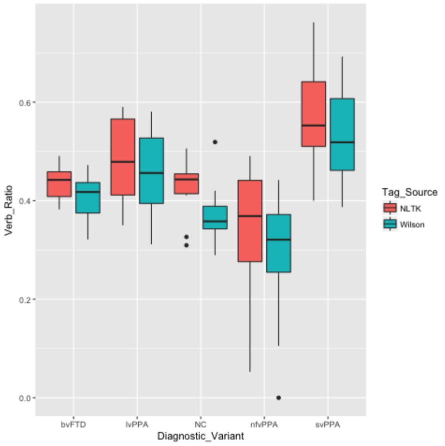

Fig. 2. Box Plots of Verb Ratio as a function of Group and Tagger.

C. Significant Difference Perspective Based on HT

The previous two sections focused on a *quantitative* evaluation of the tagger. In this section, addressing the third prong of goal one (as defined in II-F), we *qualitatively* assess the tagger by examining its ability to detect known group-based differences in QPA outcomes.

Column 1 of Table IV shows human experts found six significant differences in PR from comparisons between participant groups. Column 2 shows that all but two of those differences were detected using normalized POST applied to HT. Specifically this automated approach missed two significant differences, i.e. HC-vs-lvPPA and lvPPA-vs-nfPP and no false-positive significant differences were detected.

Column 1 of Table V shows four significant differences in VR based on normalized expert annotations. Column 2 shows that all and only these four differences were detected via the automated approach.

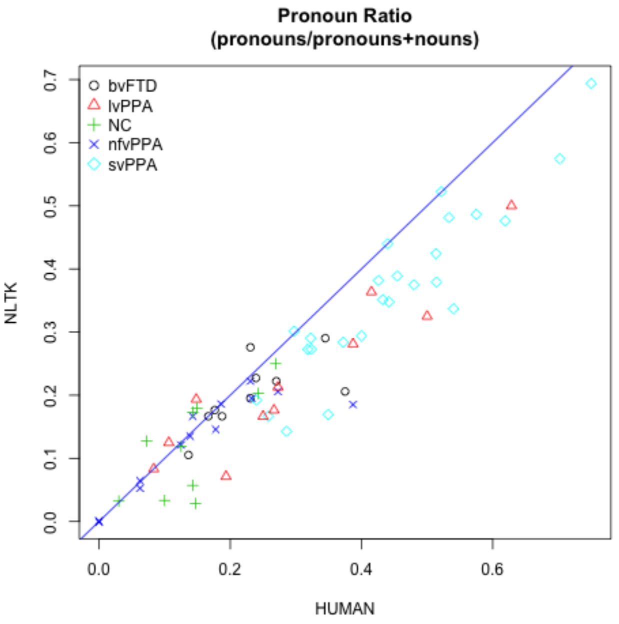

Fig. 3. **Pronoun Ratio (PR) for Human Transcripts (HT): POST versus Expert**. Each point represents a participant, y-axis corresponds to POST-based PR, x-axis to Expert or manually PR. Participant group is encoded via color/shape of a point specified (see legend). Diagonal line is y = x representing where points should lie for perfect tagging. ICC = 0.88, or excellent (see text)

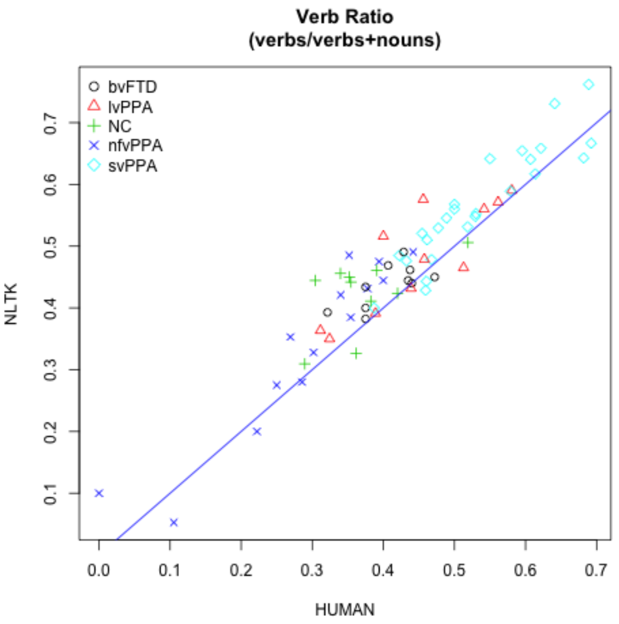

Fig. 4. **Verb Ratio (VR) for HT: POST versus Expert**. Verb ratio analogue to Figure 3, ICC = 0.90, or excellent (see text).

TABLE IV

Tagger Ability to Replicate Pronoun Ratio Findings Based on Human Expert Annotations WILSON [17] EXAMINED HOW DISEASE VARIANT AFFECTED PRONOUN RATION AND FOUND OF EIGHT TESTS PERFORMED, SIX WERE SIGNIFICANT. IN OUR AUTOMATED ANALYSIS (OBTAINED BY NORMALIZED POST TAGS INTO N, P, V, AND OTHER) FOUR OF THOSE SIX DIFFERENCES WERE FOUND WHETHER HT OR AT WAS FED TO POST. AUTOMATED ANALYSES MISSED THE SIGNIFICANT DIFFERENCE BETWEEN HC-VS-LVPPA AND AND LVPPA-VS-NFVPPA (ALTHOUGH THE LATER APPROACHED SIGNIFICANCE AT P = .065) * <= .05, ** <= .01, *** <= .001

Statistical Test	HT+Expert Finding	HT+POST Finding	AT+POST Finding
1. Omnibus	***	***	***
2. HC-vs-bvFTD	NS	NS	NS
3. HC-vs-lvPPA	*	NS	NS
4. HC-vs-nfvPPA	NS	NS	NS
5. HC-vs-svPPA	***	***	***
6. lvPPA-vs-svPPA	*	***	***
7. lvPPA-vs-nfvPPA	*	NS	NS
8. nfvPPA-vs-svPPA	*	***	***

TABLE V

Ability to Replicate Verb Ratio Findings Based on Human Expert Annotations ANALOGOUS TO TABLE IV EXCEPT WITH REGARD TO VERB RATIO. WHEN HT WAS FED TO POST, ALL FOUR SIGNIFICANT DIFFERENCES WERE FOUND BASED ON THE EXPERT ANNOTATIONS. WHEN AT WAS FED TO POST, ONLY THE SIGNIFICANT DIFFERENCE INVOLVING LVPPA-VS-NFVPPA WAS MISSED. * <= .05, ** <= .01, *** <= .001

Statistical Test	HT+Expert Finding	HT+POST Finding	AT+POST Finding
1. Omnibus	***	***	***
2. HC-vs-bvFTD	NS	NS	NS
3. HC-vs-lvPPA	NS	NS	NS
4. HC-vs-nfvPPA	NS	NS	NS
5. HC-vs-svPPA	***	**	**
6. lvPPA-vs-svPPA	NS	NS	NS
7. lvPPA-vs-nfvPPA	*	**	NS
8. nfvPPA-vs-svPPA	*	***	***

D. Evaluations Based on AT

Consider goal two: evaluating the end-to-end-system in which AT is fed to POST. First, AT performance was poor: mean(sd) word error rate (WER) was 39.1(23.4) and word recognition rate (WRR) was 68.9(19.1).

Second, we compare ICC between (i) fully automated vs. fully human PR (APR/HPR) and (ii) fully automated vs. human VR (AVR/HVR). APR/HPR ICC was in the moderate range (0.59 with a 95% CI from 0.45 to 0.71), and AVR/HVR ICC was in the good range (0.82, 95% CI from 0.75 to 0.88) using the same conservative forms of ICC as above. Pronoun/Verb-based ICC's rise to 0.75 / 0.86 if one relaxes evaluation criteria to accept consistency rather than absolute agreement (i.e. ICC3, a two-way mixed effects, consistency, single rater/measure rater/measurement form of ICC).

Evaluation of POST applied to AT in terms of the third prong or qualitative angle can be found in the third column for PR (Table IV) and VR (Table V). In terms of PR, the fully automated approach detected the same pattern of differences as did POST applied to HT (i.e. column 2). In terms of VR, the fully automated approach missed only one significant difference (i.e. lvPPA-vs-nfvPPA) compared with POST applied to HT.

IV. Discussion

Assessment of spontaneous speech using QPA is an important aspect of evaluating individuals with neurological disorders, but the clinical potential of this task is often offset by the time and expertise required for analysis. To address the problem, we aimed to assess the utility of an off-the-shelf POS tagger (i.e., the default tagger from NLTK) for automatic measurement of two QPA outcomes: PR and VR.

A. Main Findings

Despite high POST and AT error, full automation (in which AT fed to POST) performed surprisingly well – AT-based PR and VR exhibited moderate to good correlations with those derived manually by human experts (APR/HPR ICC was 0.59, and AVR/HVR ICC was 0.82). Second, POST, whether AT- or HT-based, was able to detect a super-majority of the group-based PR/VR differences identified manually by Wilson [17]. When HTs were fed to POST, 8 of 10 group-based differences were detected. With full automation, i.e. feeding ATs to POST, 7 were detected.

B. Tagger Error

In section I-F, we anticipated two opposing forces affecting accuracy: mismatch between tagging training and task language samples would hinder performance. The coarse-grained nature of PR and VR would help.

We found support for hindered tagger accuracy because mean per-token tagger accuracy on HT was 88%. This result was mediocre given that POS taggers perform in the high 90% range in typical evaluations involving journalistic prose [27]. Such poor performance may be surprising considering the task as described (choosing the four normalized tags *noun*, *verb*, *pronoun*, and *other*) has a baseline chance (guessing) accuracy of 25% whereas in typical tagger evaluations there are dozens of tags to choose from. NLTK with 45 tags would have a guessing baseline of 2%.

Why was performance so low despite the "easier" more coarse-grained tagging task? Lack of similarity to the tagger training data was anticipated in section I-F as the culprit. But was the lack of similarity rooted in (a) speech-based (as opposed to written) language patterns of tagger training data or (b) unusual speech patterns found with neurological impairment—or both? Table III data suggest that the answer depends on the disease variant. bvFTD and nfvPPA were in the same confidence interval as HC. Thus, it seems likely that (a) is the bigger driver of tagger error for these disorders. lvPPA and svPPA confidence bands' were lower than the HC confidence band, Thus, By contrast, both (a) and (b) are both likely drivers of tagger error for lvPPA and svPPA because confidence bands were below the band for HC accuracy.

Another possibility was that high tagger error rate was caused by the lack of periods or other punctuation used in our main analysis to emulate typical speech to text output (see Section I-E, first paragraph). Per-token accuracy was measured again, taking advantage of human annotations of utterance boundaries and placing a period at the end of each utterance. Under this condition, accuracy was found not to differ significantly, thus ruling out the lack of punctuation as the cause of the high error rate.

So far, the speech-based nature of the task seems to be the leading explanation for tagger error. However, we identified another possibly substantial source of error while inspecting the confusions, finding that some of the confusions arose not from tagger error but rather "normalization error." In particular, there were frequent confusions in which demonstratives such as *this* and *that* were counted by POST as determiners and were thus normalized as "other" rather than "pronoun." By contrast, the expert had tagged them not as determiners but as pronouns when they were functioning in certain contexts (e.g., as referring expressions). For this reason, POST will tend to undercount normalized pronouns compared to expert annotators. In figure 1 POST-based PR for each group is often less than and never greater than the corresponding value from the expert-derined tags [17]. Analyzing normalization error is an important next step.

C. Automatic Transcription Error

Mean WER was high (39.1%). In the future more meticulous removal of extraneous speech may help. That said, WER may be less of an issue because many believe that as treatments for these diseases become available, they must be administered early in the disease process [31]. Thus, this tool will be most valuable when used in mildly impaired individuals whose speech has not deteriorated to the extent that it substantially hurts WER.

D. Future Work

Future work may focus on a detailed error analysis. For example, is there any pattern to the errors made on the svPPA and lvPPA cases that would explain their lower accuracy? Likewise, is there a pattern that explains an apparent tendency to underestimate PR and overestimate VR?

Exploring alternative taggers is another important next step. While normal English text can be accurately tagged with the NLTK perceptron tagger using features of the previous and next word, additional contextual features and a more abstract analysis may be desirable for the language analyzed here. For example, in journalistic English, a determiner like *the* is a good predictor of a following noun, and the NLTK tagger uses that fact. But using a tagger trained with normal text to evaluate sometimes ungrammatical patient text can yield suboptimal results. More contextual features and different tagger architectures have been found to be much better for tagging languages that lack articles like *the* and languages with freer word order than English, e.g. Sanskrit [28] or Turkish [29]. For language of people with neurological deficits, similar such alternative taggers these might classify parts of speech more like a human expert than the present system does.

Instead of using an off-the-shelf tagger and AT, one could also explore *training* the tagger for this paper's specific purpose. Tagger training data based on conversational speech could help. Adapting both automated transcription and tagger models based on language data obtained from other administrations of the picture description task could also help.

E. Importance

First, this study establishes an off-the-shelf baseline against which more advanced systems can be compared. Second, there is clinical importance because PR and VR are part of QPA, a set of well-established clinical outcome measures. Although novel natural language processing (NLP) features have shown usefulness in other work, evaluating them against quantitative outcomes established by the clinical community can help increase the explainability and trust of automated diagnostics to non-NLP medical practitioners. Further, as described in Sec I-B this paper fills a gap: there are many automatic diagnosis studies while fewer have focused on quantitative multidimensional assessment.

Acknowledgments

Funding: R01: NINDS R01 NS050915, FTD PPG: NIA P01 AG019724, ADRC: NIA P50 AG023501, K24: NIDCD K24 DC015544 (M.GT). Global Brain Health Institute (A.R.) Thanks to Henk Harkema, Prateek Jain, Emily Prud'hommeaux, Wendy Shwe and Ariane Welch, David & Naomi Zubin.

References

[1] Saffran, E. M., R. S. Berndt, and M. F. Schwartz, "The quantitative analysis of agrammatic production: Procedure and data," Brain and Language, vol. 37, no. 3, pp. 440–479, 1989.

[2] Rofes, A., E. Mandonnet, J. Godden, M. H. Baron, H. Colle, A. Darlix, and M. Wager, "Survey on current cognitive practices within the European Low-Grde Glioma Network: towards a European assessment protocol," Acta Neurochirurgica, vol. 159, no. 7, pp. 1167–1178, 2017.

[3] Boschi, V., E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: a review," Front. Psychol., vol. 8, p. 269, 2017.

[4] Prins, R., and R. Bastiaanse, "Review analysing the spontaneous speech of aphasic speakers," Aphasiology, vol. 18, no. 12, pp. 1075–1091, 2004.

[5] Rofes, A., A. Talacchi, B. Santini, G. Pinna, L. Nickels, R. Bastiaanse, and G. Miceli, "Language in individuals with left hemisphere tumors: Is spontaneous speech analysis comparable to formal testing?" J. Clin. Exp. Neuropsychol., vol. 40, no. 7, pp. 722–732, 2018.

[6] Brookshire, R. H., and L. E. Nicholas, "Speech sample size and test-retest stability of connected speech measures for adults with aphasia," J. Speech Lang. Hear. Res., vol. 37, no. 2, pp. 399–407, 1994.

[7] Guinn, C. I., and A. Habash, "Language analysis of speakers with dementia of the Alzheimer's type," in AAAI Fall Symposium: Artificial Intelligence for Gerontechnology, Menlo Park, CA, 2012, pp. 8–13, 2012.

[8] Fraser, K., F. Rudzicz, N. Graham, and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," in Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies, 2013, pp. 47–54.

[9] Fraser, K. C., J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon E, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts, " Cortex, vol. 55, pp. 43–60, Jun. 2014.

[10] Garrard, P., V. Rentoumi, B. Gesierich, B. Miller, and M. L. Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse," Cortex, vol. 55, pp. 122–129, Jun. 2014.

[11] Jarrold, W., B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in Proc. ACL Workshop on Computational Linguistics and Clinical Psychology, Jun. 2014, pp. 27–36.

[12] Rentoumi, V., L. Raoufian, S. Ahmed, C. A. de Jager, and P. Garrard, "Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology," J. Alzheimers Dis., vol. 42, suppl. 3, pp. S3–S17, 2014.

[13] Fraser, K. C., J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," J. Alzheimers Dis., vol. 49, no. 2, pp. 407–422, 2015.

[14] Hsu, C. J., and C. K. Thompson, "Manual versus automated narrative analysis of agrammatic production patterns: The Northwestern Narrative Language Analysis and Computerized Language Analysis," J. Speech Lang. Hear. Res., vol. 61, no. 2, pp. 373–385, 2018.

[15] Pressman, P. S., and B. L. Miller, "Diagnosis and management of behavioral variant frontotemporal dementia," Biol. Psychiatry, vol. 75, no. 7, pp. 574–581, 2014.

[16] Berndt, R. S., S. Wayland, E. Rochon, E. M. Saffran, and M. Schwartz, Quantitative Production Analysis: A Training Manual for the Analysis of Aphasic Sentence Production. Hove, UK: Psychology Press, 2000.

[17] Wilson, S. M., M. L. Henry, M. Besbris, J. M. Ogar, N. F. Dronkers, W. Jarrold, and M. L. Gorno-Tempini, "Connected speech production in three variants of primary progressive aphasia," Brain, vol. 133, no. 7, pp. 2069–2088, 2010.

[18] Fraser, K. C., N. Ben-David, G. Hirst, N. Graham, and E. Rochon, "Sentence segmentation of aphasic speech," in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 862–871.

[19] Kertesz, A, Western Aphasia Battery. New York: Grune and Stratton, 1982.

[20] Bird, S., and E. Loper, "NLTK: the natural language toolkit," in Proc. of the ACL 2004 on Interactive poster and demonstration sessions, Association for Computational Linguistics, 2004, p. 31.

[21] Honnibal, M., "A good part-of-speech tagger in about 200 lines of python," 2013. `https://explosion.ai/blog/part-of-speech-pos-tagger-in-python`.

[22] Weischedel, R., S. Pradhan, L. Ramshaw, J. Kaufman, M. Franchini, and M. El-Bachouti, "OntoNotes Release 5.0," 2012, `https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf`.

[23] Hothorn, T., F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," Biom. J., vol. 50, no. 3, pp. 346–363, 2008.

[24] Tukey, J., "Comparing individual means in the analysis of variance," Biometrics, vol. 5, no. 2, pp. 99–114, 1949.

[25] Cicchetti, D. V., "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," Psychological Assessment, vol. 6, no. 4, pp. 284–290, 1994.

[26] McGraw, K. O., and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," Psychological Methods, vol. 1, no. 1, pp. 30–46, 1996.

[27] Manning, C. D., "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?" in International Conference on Intelligent Text Processing and Computational Linguistics," Berlin, Heildelberg: Springer, Feb. 2011, pp. 171–189.

[28] Krishna, A., B. Santra, S. P. Bandaru, G. Sahu, V. D. Sharma, P. Satuluri, and P. Goyal, "Free as in free word order: an energy based model for word segmentation and morphological tagging in Sanskrit," Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2550–2561.

[29] Ehsani, R., M. E. Alper, G. Eryiğit, and E. Adalı, "Disambiguating Main POS tags for Turkish," in Proc. of the Twenty-Fourth Conference on Computational Linguistics and Speech Processing, 2012, pp. 202–213.

[30] Koo, T. K., and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," J. Chiropr. Med., vol. 15, no. 2, pp. 155–163, 2016.

[31] Kumar A, Singh A, Ekavali, "A review on Alzheimer's disease pathophysiology and its management: an update," Pharmacol. Rep., vol. 67, no. 2, pp. 195–203, 2015.