AJSLP

Research Article

Auditory-Perceptual Rating of Connected Speech in Aphasia

Marianne Casilio,[a] Kindle Rising,[a] Pélagie M. Beeson,[a,b] Kate Bunton,[a] and Stephen M. Wilson[c]

Purpose: Auditory-perceptual assessment, in which trained listeners rate a large number of perceptual features of speech samples, is the gold standard for the differential diagnosis of motor speech disorders. The goal of this study was to investigate the feasibility of applying a similar, formalized auditory-perceptual approach to the assessment of language deficits in connected speech samples from individuals with aphasia.
Method: Twenty-seven common features of connected speech in aphasia were defined, each of which was rated on a 5-point scale. Three experienced researchers evaluated 24 connected speech samples from the AphasiaBank database, and 12 student clinicians evaluated subsets of 8 speech samples each. We calculated interrater reliability for each group of raters and investigated the validity of the auditory-perceptual approach by comparing feature ratings to related quantitative measures derived from transcripts and clinical measures, and by examining patterns of feature co-occurrence.
Results: Most features were rated with good-to-excellent interrater reliability by researchers and student clinicians. Most features demonstrated strong concurrent validity with respect to quantitative connected speech measures computed from AphasiaBank transcripts and/or clinical aphasia battery subscores. Factor analysis showed that 4 underlying factors, which we labeled Paraphasia, Logopenia, Agrammatism, and Motor Speech, accounted for 79% of the variance in connected speech profiles. Examination of individual patients' factor scores revealed striking diversity among individuals classified with a given aphasia type.
Conclusion: Auditory-perceptual rating of connected speech in aphasia shows potential to be a comprehensive, efficient, reliable, and valid approach for characterizing connected speech in aphasia.

Connected speech in individuals with aphasia reflects underlying impairments in any of the speech/language domains, including lexical retrieval, grammatical construction, phonological encoding, and articulatory agility (Vermeulen, Bastiaanse, & Van Wageningen, 1989). This sensitivity to many different types of disturbances makes analysis of connected speech a valuable tool for assessment, diagnosis, and evaluation of treatment outcomes (Prins & Bastiaanse, 2004). The overall goal of this study was to investigate the feasibility of applying a formalized auditory-perceptual approach to rating features of connected speech in aphasia. We were primarily concerned with structural aspects of language (i.e., semantics, lexicon, syntax, phonology), as opposed to functional and pragmatic aspects of language use, or discourse organization (Linnik, Bastiaanse, & Höhle, 2016).

There are two predominant approaches to the structural analysis of connected speech in aphasia: quantitative linguistic analysis and qualitative rating scales (Prins & Bastiaanse, 2004). Quantitative linguistic analysis entails transcribing speech samples and coding them for relevant features, such as utterance length and complexity; the use of different classes of words and morphemes; and phonological, syntactic, and semantic errors (Bastiaanse, Edwards, & Kiss, 1996; Doyle, Goda, & Spencer, 1995; Haravon, Obler, & Sarno, 1994; Howes, 1967; MacWhinney, Fromm, Forbes, & Holland, 2011; Marini, Andreetta, del Tin, & Carlomagno, 2011; Miller, Andriacchi, & Nockerts, 2015; Nicholas & Brookshire, 1993; Saffran, Berndt, & Schwartz, 1989; Shewan, 1988; Thompson, Ballard, Tait, Weintraub, & Mesulam, 1997; Vermeulen et al., 1989; Wagenaar, Snow, & Prins, 1975; Wilson et al., 2010; Yorkston & Beukelman, 1980). Quantitative approaches are relatively objective,

[a]Department of Speech, Language, and Hearing Sciences, The University of Arizona, Tucson
[b]Department of Neurology, The University of Arizona, Tucson
[c]Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, TN
Correspondence to Stephen M. Wilson: stephen.m.wilson@vanderbilt.edu

Editor-in-Chief: Julie Barkmeier-Kraemer
Editor: Anastasia Raymer
Received August 31, 2018
Revision received October 13, 2018
Accepted October 17, 2018
https://doi.org/10.1044/2018_AJSLP-18-0192

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

and some schemes have been shown to have excellent interrater reliability (Gordon, 2006; Rochon, Saffran, Berndt, & Schwartz, 2000). Moreover, most quantitative approaches yield comprehensive and multidimensional sets of measures that quantify multiple domains of language function. These multidimensional descriptive measures allow individuals to be characterized in terms of profiles of spared and impaired functions (Vermeulen et al., 1989; Wilson et al., 2010). However, quantitative linguistic analysis has several limitations. First, it is extremely time-consuming and requires highly trained transcribers/coders with substantial knowledge of linguistics and aphasia (Prins & Bastiaanse, 2004; Yagata et al., 2017). From a practical point of view, this can preclude the use of quantitative linguistic analysis in clinical practice, large clinical research studies, or large clinical trials. Second, even the most well-specified coding schemes still require subjective decisions to be made, and relevant features may not occur consistently (Bastiaanse, 1995; Boyle, 2014, 2015; Gordon, 2006; Rochon et al., 2000). Third, existing schemes do not effectively capture phenomena arising from apraxia of speech, which frequently co-occurs with aphasia (Duffy, 2013).

In contrast, qualitative rating scales, such as the profile of speech characteristics on the Boston Diagnostic Aphasia Examination (BDAE; Goodglass, Kaplan, & Barresi, 2001), the fluency scale on the Western Aphasia Battery–Revised (WAB-R; Kertesz, 2007), and the rating scales for spontaneous speech on the Aachen Aphasia Test (Huber, Poeck, & Willmes, 1984), are quick tools intended for use by clinicians. Easy to administer and score, they provide an overall profile of patients' speech, and with experienced raters, the reliability of these scales is excellent (Goodglass & Kaplan, 1972; Kertesz, 1979). The widespread use of these qualitative scales is testament to their practicality for research and clinical applications. However, qualitative rating scales have limitations, too. First, scores are much less consistent between raters when the scales are applied by community clinicians (Gordon, 1998; Trupe, 1984). Second, because few dimensions are rated, these instruments presuppose which connected speech features are important. For example, the Grammatical Form measure on the BDAE is defined in terms of a continuum of agrammatism, precluding quantification of paragrammatism, whereas the Word-Finding measure is defined in terms of empty speech, but not abandoned utterances, word-finding pauses, or other instantiations of word-finding difficulty. Third, the limited dimensionality of existing qualitative scales also entails that distinct phenomena are conflated. For example, the fluency scale of the WAB-R incorporates grammatical, motor speech, and paraphasic features.

In this study, we investigated the feasibility of a new approach to the assessment of connected speech in aphasia based on the auditory-perceptual approach that is used in the assessment and diagnosis of motor speech disorders (Darley, Aronson, & Brown, 1969a, 1969b, 1975). In auditory-perceptual analysis, speech samples are rated on dozens of different perceptual dimensions in order to characterize dysarthria and apraxia of speech (Duffy, 2013; Strand, Duffy, Clark, & Josephs, 2014). The auditory-perceptual approach is reliable in both experienced and inexperienced listeners (Bunton, Kent, Duffy, Rosenbek, & Kent, 2007; Kent, Kent, Duffy, & Weismer, 1998), and different patterns are associated with distinct etiologies (Darley et al., 1969b, 1975). Consequently, this approach remains the gold standard for assessment, diagnosis, and clinical decision making in motor speech disorders (Duffy, 2013).

We describe a system for auditory-perceptual rating of connected speech in aphasia (APROCSA) in which 27 types of disruptions and abnormalities that are commonly observed in connected speech in aphasia are each scored on a 5-point scale. Our aim was to combine the key positive aspects of quantitative linguistic analysis and quantitative rating scales: By rating a large number of dimensions, we would derive multidimensional data-driven outcome measures, yet because our ratings would be qualitative, these measures would be very quick to obtain.

Connected speech samples from 24 individuals with aphasia were retrieved from the AphasiaBank database (MacWhinney et al., 2011) and evaluated by experienced researchers and student clinicians, who represented those with limited prior experience with connected speech analysis. The resulting data were used to quantify the interrater reliability of each feature, to determine to what extent this depends on rater experience, to assess concurrent validity by examining correlations between APROCSA features and measures derived from quantitative linguistic analysis of transcripts and standard clinical measures, and to identify explanatory underlying factors that may account for patterns of co-occurrence among connected speech features.

Method

The APROCSA System

Twenty-seven common features of connected speech in aphasia were defined (see Table 1). These features were selected to cover the range of deficits that commonly occur in aphasias of diverse etiologies, based on our experience of quantitative analysis of connected speech in poststroke aphasia (Yagata et al., 2017), primary progressive aphasia (Wilson et al., 2010), and neurosurgical patients (McCarron et al., 2017). We also considered the features that are captured by existing quantitative (Bastiaanse et al., 1996; Haravon et al., 1994; Howes, 1967; MacWhinney et al., 2011; Marini et al., 2011; Miller et al., 2015; Saffran et al., 1989; Shewan, 1988; Thompson et al., 1997; Vermeulen et al., 1989; Wagenaar et al., 1975) and qualitative (Goodglass et al., 2001; Huber et al., 1984; Kertesz, 2007) schemes. Probably the most widely used and comprehensive system for quantitative analysis of connected speech in aphasia is Codes for the Human Analysis of Transcripts (CHAT; MacWhinney et al., 2011; see bibliography at https://aphasia.talkbank.org/publications). Eighteen of the APROCSA features have close counterparts in CHAT. The remaining nine APROCSA features do not correspond closely to any

Table 1. The 27 features of the auditory-perceptual rating of connected speech in aphasia.

Connected speech feature	Definition
Anomia	Overall impression of word-finding difficulties, which may be instantiated in many different ways, e.g., word-finding pauses, abandoned utterances, circumlocution, comments on inability to find words. Some of these behaviors are also captured by other specific features. Nonaphasic speakers sometimes have difficulty finding words, so occasional difficulties may be scored *not present* (0).
Abandoned utterances	Utterances are left incomplete. The speaker may move on to another idea, stop talking, attempt to use another modality (e.g., gesture), or give a vague conclusion to the utterance (e.g., shrug shoulders and say "you know").
Empty speech	Speech that conveys little or no meaning. Pronouns and nonspecific words such as *thing, stuff,* and *do* are substituted for content words.
Semantic paraphasias	Substitution of content words for related or unrelated content words, e.g., *cat* for *dog*.
Phonemic paraphasias	Substitution, insertion, deletion, or transposition of clearly articulated phonemes, e.g., *papple* for *apple*.
Neologisms	Word forms that are not real English words. The intended target may or may not be apparent.
Jargon	Mostly fluent and prosodically correct but largely meaningless speech containing paraphasias, neologisms, and unintelligible strings.
Perseverations	Repetition of previously used words or utterances in contexts where they are no longer appropriate.
Stereotypies and automatisms	Commonly used words, phrases, or neologisms produced with relative ease and fluency, e.g., *tan, I know it, dammit*.
Short and simplified utterances	Utterances are reduced in length or complexity. A *mild* rating (1) should reflect utterances that are sometimes shorter than expected based on the context (e.g., simple sentence structures, lack of subordinate clauses). A *severe* rating (4) should be reserved for single-word utterances. Nonsentence responses (e.g., *Did you come with your wife? Yes,* or *Who did you come with? My wife.*) should not be considered.
Omission of bound morphemes	Inflectional or derivational morphemes are not used where they should be, e.g., *I am go to the store*.
Omission of function words	Function words are not used where they should be, e.g., *I going to the store*.
Paragrammatism	Inappropriate juxtaposition of words and phrases and/or misuse of function words and morphemes (e.g., *It's so much wonderful*, *Makes it hard to speech*).
Pauses between utterances	Pauses that occur between utterances may relate to utterance formulation. Pauses between examiner's questions and patient's responses should also be considered. Failure to string together multiple utterances when appropriate can be scored here.
Pauses within utterances	Unfilled or filled (*um, uh*) pauses within utterances. Both prevalence and length of pauses should be taken into account in assessing severity. Because pauses are a feature of unimpaired connected speech, a score of *not present* (0) should be assigned if the number of pauses is within the typical range.
Halting and effortful	Speaking is labored and consequently uneven. Intonation, rhythm, or stress patterns may be reduced, absent, or inappropriately placed. Prosody or melodic line may be disrupted.
Reduced speech rate	The number of words per minute within utterances is reduced. Speaking slowly and pauses within utterances count toward reduced rate. Pauses between utterances, potentially reflecting utterance formulation, do not count.
False starts	Partial words are abandoned after one or two phonemes, e.g., *It's a ca- cat.*
Retracing	Sequences of one or more complete words are made redundant by subsequent repetitions, revisions, amendments or elaborations, e.g., *The kite is (.) the boy is flying the kite*.
Conduite d'approche	Successive approximations at target forms. The target may or may not be achieved. The patient is aware of their errors. These instances also contribute to scores for *Retracing* and *Phonemic paraphasias* or *Neologisms*.
Target unclear	It is not clear what phonemes the speaker is attempting to produce. This is often due to dysarthria, apraxia of speech, muttering, mumbling, or in some cases severe jargon.
Meaning unclear	It is not clear what the speaker is talking about, or the topic may be clear but what is being said about it is not.
Off-topic	It is not clear how what is being said relates to the context.
Expressive aphasia	Language production is disrupted.
Apraxia of speech	Speech contains distortions, substitutions, or omissions that tend to increase with length or complexity of the word or phrase. Groping behaviors or impaired intonation may be present. See Duffy (2013) for more information.
Dysarthria	Speech is difficult to understand and characterized as *slurred, choppy,* or *mumbled*. Errors are consistent and are the result of impaired strength, tone, range of motion, or sequencing. Speech breathing, phonation, resonance, articulation, and prosody may be impaired. See Duffy (2013) for more information.
Overall communication impairment	Overall impression of the extent to which the speaker is impaired in conveying their message. A *mild* rating (1) should reflect an evident speech-language impairment, but no limitation in discussing all topics. A *moderate* rating (2) should be used when the speaker can readily communicate about simple, everyday topics, but is imited in discussion of more complex topics. A *marked* rating (3) should be used when communication about everyday topics is possible with help from the examiner, but the patient shares the burden of communication. A *severe* rating (4) should be used when all communication is fragmentary, and the examiner carries the burden of communication. These guidelines, including some of the specific wording, are based on the Boston Diagnostic Aphasia Examination Aphasia Severity Rating Scale.

CHAT measures, because they are summary measures (*Anomia, Expressive aphasia, Overall communication impairment*), primarily or significantly reflect motor speech deficits (*Apraxia of speech, Dysarthria, Halting and effortful*), or are too subjective to be coded in a quantitative system (*Conduite d'approche, Meaning unclear, Off-topic*). One additional feature, *Circumlocution,* was also defined and rated but was subsequently excluded due to poor interrater reliability.

Features were defined in terms of readily perceptible phenomena, rather than placing the onus on raters to make inferences about underlying mechanisms. For instance, *Short and simplified utterances* can be reflective of either grammatical or motor speech deficits (or both), but raters were not required to adjudicate. It was hoped that underlying factors would emerge from factor analysis of the surface features rated. Note that many APROCSA features overlap with one another to various extents. For instance, a patient with *Anomia* is likely to show *Abandoned utterances* and *Pauses within utterances,* whereas a patient with *Jargon* will exhibit *Semantic paraphasias, Phonemic paraphasias*, and *Neologisms*. The overlap among features is intentional and reflects our view that the many surface features of connected speech in aphasia are instantiations of more fundamental underlying impairments.

The APROCSA was designed primarily to capture language and not motor speech deficits, because there already exist comprehensive auditory-perceptual rating schemes for dysarthria (Darley et al., 1969a, 1969b, 1975) and apraxia of speech (Strand et al., 2014). Just two motor speech features—*Dysarthria* and *Apraxia of speech*—were included in order to capture motor speech deficits in a summary manner. As alluded to above, it should be noted that some other APROCSA features can be impacted by both speech and language deficits, such as *Short and simplified utterances, Halting and effortful,* and *Reduced speech rate*.

A 5-point, equal-appearing interval scale was used to rate each feature (see Table 2), modified from a similar scale used for auditory-perceptual rating of apraxia of speech (Strand et al., 2014). Each point on the scale was explicitly defined, taking into account both severity and frequency. The score of 0 (*not present*) was defined to include the range of healthy older speakers; it is not uncommon for individuals with normal language function to exhibit some APROCSA features, such as retracing a phrase or pausing to find a word. The rating scale and a list of the 27 features were included on a one-page scoresheet used by the raters (see the Appendix).

Table 2. The 5-point rating scale used in the auditory-perceptual rating of connected speech in aphasia.

Score	Severity	Description
0	Not present	Not present or within the range of healthy older speakers
1	Mild	Detectable but infrequent
2	Moderate	Frequently evident but not pervasive
3	Marked	Moderately severe, pervasive
4	Severe	Nearly always evident

Note. The scale is based on Strand et al. (2014).

Individuals With Aphasia

Twenty-four videotaped connected speech samples of speakers with chronic poststroke aphasia (aged 49–76 years, 12 men and 12 women) were selected from the AphasiaBank database (MacWhinney et al., 2011). All speakers were right-handed, monolingual English speakers with vision and hearing (aided or unaided) adequate for testing. Demographic information and standardized test scores are presented in Table 3. Speech samples were collected at participating universities and outpatient clinics across the United States.

The speech samples were selected such that patients were diverse in aphasia severity (WAB-R Aphasia Quotient [AQ] range 20.3 to 92.7) and aphasia type according to the WAB-R (seven Anomic, five Conduction, four Broca's, four Wernicke's, two Global, one Transcortical Motor, one Transcortical Sensory). This distribution of aphasia types was selected to approximately reflect prevalence of aphasia types within typical outpatient populations of individuals with chronic poststroke aphasia (Kertesz, 1979). Furthermore, within each aphasia type, patients were chosen to reflect a range of degrees of impairment.

Connected Speech Samples

Excerpts containing approximately 5 min of patient speech were clipped from the first speech sample available for each patient. Previous research identified this time frame as adequate to evaluate communicative efficacy in aphasia, assuming that all diagnostic behaviors occur at least three times per minute (Boles & Bombard, 1998). All excerpts were taken from the Free Speech Samples portion of the AphasiaBank protocol, during which patients talked about their speaking abilities, stroke, and recovery and, in some cases, recounted a memorable life event. All ratings were carried out based on audiovisual samples (i.e., raters listened to and watched the patients talking).

Raters

Two groups of raters participated in the study. Raters in both groups passed a hearing screening (25 dB HL at 1, 2, 4 kHz) and spoke English with native proficiency.

The first group comprised three researchers (authors of this article: S. M. W., K. R., M. C.) with experience in connected speech analysis. S. M. W. was an aphasia researcher with 14 years of experience in aphasia research and expertise in quantitative linguistic analysis of connected speech in aphasia. K. R. was a licensed speech-language pathologist with more than 10 years of experience as a research clinician in an aphasia research laboratory. M. C.

Table 3. Patient characteristics.

Patient	Age (years)	Sex	Race	Education (years)	Time postonset (months)	WAB-R AQ (/100)	Aphasia type (WAB-R)	Apraxia of speech	Clip
Fridriksson05a	58	F	W	12	149	92.7	Anomic	Y	0:00–5:39
TAP18a	53	F	W	16	23	90.3	Anomic	Y	0:00–5:37
Whiteside06a	62	M	W	12	91	88.8	Anomic	Y	0:00–5:50
Adler01a	58	M	W	13	16	86.8	Anomic	Y	2:55–8:44
Kurland07a	70	F	W	16	13	83.0	Anomic	N	0:21–6:26
Kurland28a	62	M	W	16	6	78.7	Anomic	N	4:26–9:42
Scale30a	48	M	W	18	46	68.5	Anomic	N	0:04–5:48
ACWT09a	56	F	W	13	94	80.1	Conduction	Y	0:00–5:25
Wright203a	66	M	W	18	80	76.3	Conduction	N	0:00–5:33
Williamson04a	60	M	W	14	296	70.6	Conduction	Y	0:00–6:18
Kurland20a	50	F	AA	12	6	67.0	Conduction	N	0:15–6:50
TCU07a	49	F	W	16	15	52.0	Conduction	Y	0:00–6:32
Williamson16a	63	F	W	16	58	66.4	Trans Sensory	N	0:05–5:48
ACWT02a	53	F	W	14	39	74.6	Trans Motor	Y	0:02–6:07
Elman12a	57	M	W	20	54	74.4	Wernicke	N	0:00–6:15
Elman14a	76	F	AA	17	55	65.7	Wernicke	N	0:00–5:27
Thompson05a	63	F	W	16	155	58.5	Wernicke	—	0:15–5:35
Kurland18a	74	M	AA	16	9	44.0	Wernicke	N	0:25–6:03
Scale33a	57	F	W	—	104	71.1	Broca	N	0:00–5:39
TCU08a	57	M	AA	14	95	63.9	Broca	Y	0:00–6:21
TAP11a	62	F	W	14	44	58.1	Broca	Y	0:00–5:56
BU08a	64	M	W	12	110	39.7	Broca	N	0:12–6:22
TAP09a	71	M	W	16	36	20.5	Global	Y	0:00–6:24
Scale09a	66	M	W	12	240	20.3	Global	Y	0:00–6:10

Note. Em dashes indicate data not available. WAB-R = Western Aphasia Battery–Revised; AQ = Aphasia Quotient; Clip = portion of audiovisual file excerpted for this study; M = male; F = female; W = White; AA = African American; Trans = transcortical; Y = yes; N = no.

was a master's student in speech-language pathology at The University of Arizona, with 3 years of transcription experience and specific training in the transcription and coding of connected speech in aphasia.

The second group consisted of 12 second-year master's students in speech-language pathology at The University of Arizona who had completed graduate coursework in aphasia and had at least 25 hr of clinical experience in aphasia (see Table 4). Aside from these minimum requirements, the students varied considerably in terms of their relevant experience; this variability was considered representative of relatively inexperienced raters whose rating abilities we aimed to assess.

The study was approved by The University of Arizona Institutional Review Board. Student clinician raters provided written informed consent to participate and were compensated for their time.

Rating Procedures

Researcher Calibration

Prior to rating the speech samples, one separate sample from AphasiaBank, Elman03a, was selected for rating calibration and discussion among the three researcher raters. Elman03a was a 52-year-old man who was 11 years poststroke. His AQ was 66.2, and he met WAB-R criteria for

Table 4. Student clinician rater characteristics.

Age	22–33 years (M = 25.5 ± 3.3)
Sex	11 female, 1 male
First language	10 English, 1 Shanghainese and English, 1 Korean
Highest degree earned	10 bachelor's, 2 master's
Clinical experience in adult language	25–200 hr (M = 86 ± 50 hr)
Clinical settings in adult language	University aphasia clinic (all), acute care (5), inpatient rehabilitation (5), private clinic (2)
Research experience	0–4,220 hr (M = 1,099 ± 1,211 hr)
Transcription experience	0–1,920 hr (M = 376 ± 677 hr)
Auditory-perceptual evaluation of motor speech disorders experience	0–50 hr (M = 12.5 ± 16.0 hr)
Confidence in knowledge of aphasia	4–5 on a 5-point Likert scale (M = 4.4 ± 0.5)
Confidence in knowledge of motor speech disorders	2–4 on a 5-point Likert scale (M = 3.2 ± 0.7)
Graduate coursework in aphasia	All completed
Graduate coursework in motor speech disorders	1 Completed, 11 in progress

Broca's aphasia. He also had a clinical diagnosis of apraxia of speech. He differed from the patients included in the study in that he was bilingual in English and Mandarin. This particular speech sample was selected because Elman03a was one of the few speakers with relatively moderate aphasia who presented with almost all of the APROCSA features. The three researchers rated Elman03a independently and then met to discuss their ratings. For each feature that did not demonstrate exact agreement across all three researchers, a consensus score was reached through discussion and review of the videotaped speech sample. The Elman03a sample and the consensus scores were then used as part of a training session developed for student raters, described below.

Researcher Rating Procedures

Each researcher then independently rated all 24 patient samples. In general, we listened to each sample once, rated most of the features, and then listened to about half of the sample again while making decisions about the remaining features. Ratings were completed over a 1-month time frame, using a variety of personal computers and headphones.

Student Clinician Training

Prior to rating speech samples, the student clinicians participated in a 2.5-hr training session that reviewed the purpose of the APROCSA, scoring procedures, and an in-depth explanation of the 27 connected speech features. Trainings took place on two different dates to accommodate raters' schedules. M. C. delivered the training presentation with the assistance of a doctoral candidate with expertise in motor speech disorders, who led a 20-min session on how to distinguish phonological impairments from apraxia of speech. The training session included a practice exercise in which each student clinician rated the Elman03a sample. The consensus scores were then presented, student clinicians compared the scores they had assigned to the consensus scores, and discrepancies were discussed. Student clinicians were encouraged to ask questions throughout the training presentation. The most prominent topics of discussion were how to differentiate paragrammatism from agrammatism and phonemic paraphasias from apraxia of speech (see definitions in Table 1).

Student Clinician Rating Procedures

Within 2 weeks of the training session, each student clinician rated a quasirandomized selection of eight of the 24 speech samples. The samples were assigned such that each sample was rated by four student clinicians. To limit listener fatigue, student clinicians performed their ratings in two 1-hr sessions, rating four patients in each session. Student clinicians listened to samples in a quiet room on a ThinkPad T60 laptop with Audio-Technica QuietPoint ATH-ANC7b headphones. They were instructed to listen to each sample twice and to spend no more than 15 min per sample. A four-page manual was provided, consisting of a version of Table 1, along with instructions based on the methods described above. Student clinicians were encouraged to take notes, including potentially transcribing utterances they found particularly informative, and many did so.

Interrater Reliability

The reliability of each feature was assessed in terms of intraclass correlation coefficients (ICCs; McGraw & Wong, 1996). For the researchers, we calculated ICCs for two-way models, because each of the 24 patients was rated by all three of the researchers. Both the patients that were rated and the researchers were considered to be random factors (i.e., researchers were in principle drawn from a pool of similar researchers). ICCs reflecting absolute agreement were calculated. As such, the appropriate ICCs for the researchers were ICC(A,1), which estimates the absolute agreement of any two measurements, and ICC(A,k), which estimates the absolute agreement of measurements that are averages of k independent measurements, where $k = 3$ (because three researchers rated each patient).

For student clinicians, we calculated ICCs for a one-way model in which patients rated were a random factor. Each patient was rated by four students, but because a different subset of students rated each patient, there was no inherent order to the four ratings obtained for each patient. Accordingly, the appropriate ICCs were ICC(1), which estimates the absolute agreement of any two measurements, and ICC(k), which estimates the absolute agreement of measurements that are averages of k independent measurements, where $k = 4$ (because four students rated each patient). ICCs were interpreted as poor ($r < .40$), fair ($.40 \leq r < .60$), good ($.60 \leq r < .75$), or excellent ($r \geq .75$), following Cicchetti (1994).

The reliability of each individual researcher and each individual student on each APROCSA feature was assessed by calculating an ICC (Type A,1) between the individual and the mean of the other two researchers (in the case of researchers) or the mean of the three researchers (in the case of students) on the relevant set of rated patients (24 for researchers, eight for students). For each individual, the 27 ICCs (one per feature) were converted to z scores (McGraw & Wong, 1996, Appendix B), averaged together, and converted back to r. The mean ICCs of the researchers and students were then compared with the two-sample Kolmogorov–Smirnov test.

Concurrent Validity

The concurrent validity of the APROCSA connected speech features was investigated by calculating Pearson correlations between APROCSA scores (averaged across the three researchers) and 23 measures derived from AphasiaBank. Seventeen of these were quantitative linguistic measures calculated from the available CHAT-coded transcriptions, all of which have been reviewed for accuracy by two transcribers, one of whom is a licensed speech-language pathologist (MacWhinney et al., 2011). Calculations were performed using the CLAN (Computerized Language Analysis) programs FREQ and EVAL (MacWhinney, 2000).

The quantitative measures are shown in Table 5, along with details of how each was calculated. The other six measures were the WAB-R AQ and subscores for information content, fluency, comprehension, repetition, and naming.

For 23 of the 27 APROCSA features, one or more AphasiaBank measures were identified a priori as representing similar or related constructs. For example, the APROCSA feature *Abandoned utterances* was related to the AphasiaBank transcript-based measure *Abandoned utterances (per hundred words)*, which was calculated by counting CHAT postcodes for abandoned utterances.

Patterns of Feature Co-occurrence

To examine patterns of co-occurrence among the APROCSA features, pairwise Pearson correlations were first computed between all APROCSA feature scores (again, averaged across the three researchers). Then, factor analysis with varimax rotation was performed using *factoran* in MATLAB (MathWorks). Four connected speech features—*Conduite d'approche, Off-topic, Dysarthria,* and *Overall communication impairment*—were excluded from this analysis, as the algorithm required fewer features than patients. Three of these four features—*Conduite d'approche, Off-topic,* and *Dysarthria*—were excluded due to their relatively low reliability and relatively restricted distribution in the patient sample. *Overall communication impairment* was excluded because it was similar to and highly correlated with the *Expressive aphasia* feature. A model with four factors yielded the most explanatory dimensionality reduction of the data, as described in the Results section.

Results

Most APROCSA features showed broad distributions across the 24 patients for both the researcher and student clinician raters (see Figure 1, Columns 1 and 4), showing that the selected patient sample varied in terms of presenting features and the severity of those features.

Interrater Reliability

Researchers

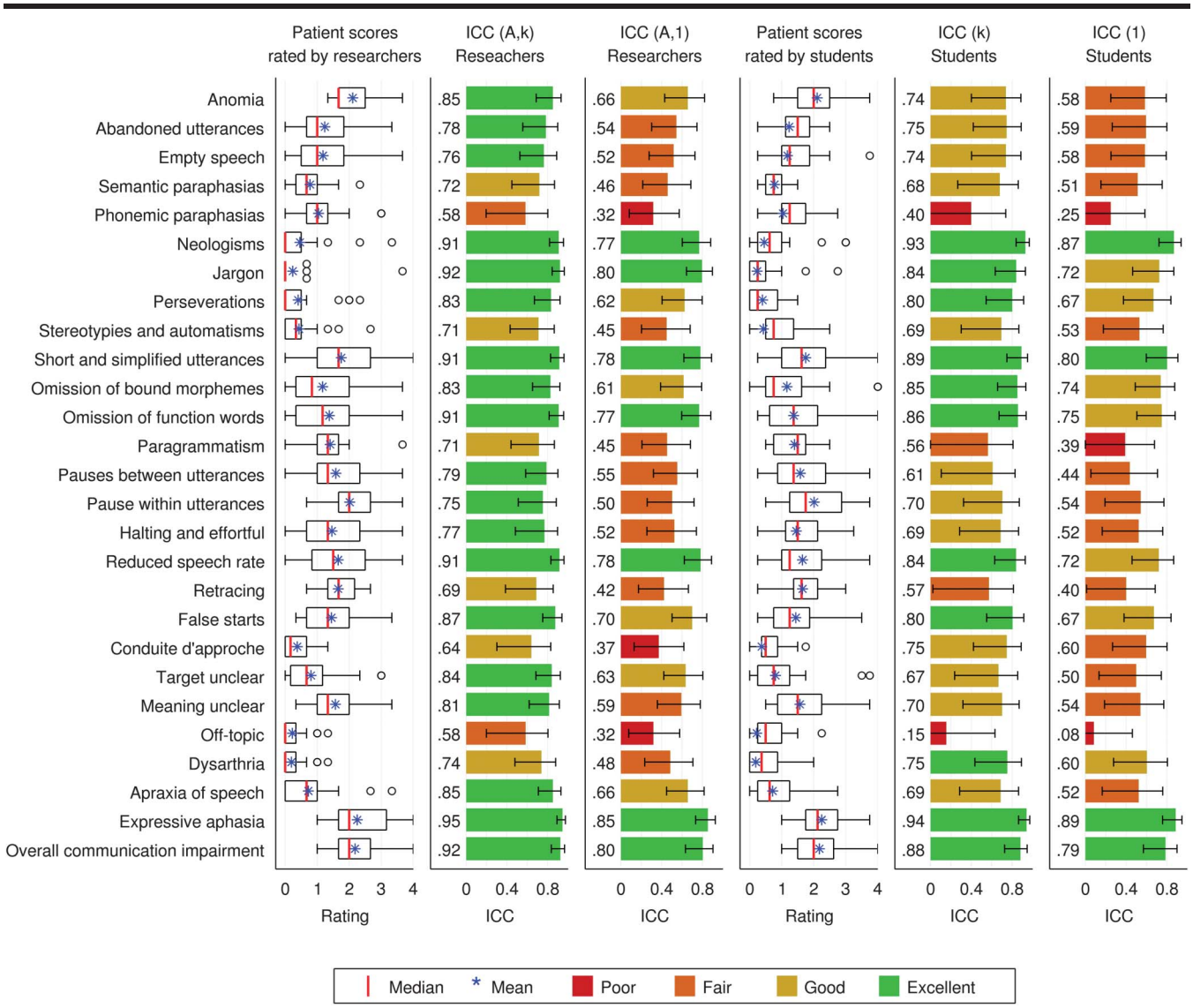
ICC(A, *k*), an estimate of reliability when ratings were averaged across the three researchers, was excellent ($r \geq .75$) for 19 features, good ($.60 \leq r < .75$) for six features, and fair ($.40 \leq r < .60$) for two features (see Figure 1, Column 2). ICC(A, 1), an estimate of reliability in a situation

Table 5. Quantitative linguistic measures derived with CLAN from CHAT transcriptions.

Quantitative linguistic measure	How the measure was calculated
Anomia (phw)	Utterance terminators +… and +..? for abandoned utterances, postcode [+es] for empty speech, codes (.), (..) and (…) for unfilled pauses, and &ah, &eh, &ew, &hm, &mm, &uh, &uhm, and &um for filled pauses were summed using FREQ, divided by the number of words, and multiplied by 100.
Abandoned utterances (phw)	Utterance terminators +… and +..? for abandoned utterances were summed using FREQ and expressed per hundred words as above.
Empty speech (phw)	Postcode [+es] for empty speech.
Semantic errors (phw)	Word-level error codes [*s:r], [*s:ur], [*s:uk], and [*s:per].
Phonological errors (phw)	Word-level error codes [*p:w], [*p:m], and [*p:n].
Neologisms (phw)	Word-level error codes [*n:k] and [*n:uk].
Jargon (phw)	Word-level error codes [*s] for semantic errors, [*p] for phonological errors, and [*n] for neologistic errors, and postcode [+jar] for jargon.
Mean length of utterance (morphemes)	Calculated using EVAL. Revisions, fillers, and unintelligible utterances were excluded.
Bound morphemes (proportion)	%mor tier codes for bound morphemes (plurals, 3S, 1S/3S, PAST, PASTP, PRESP) and free morphemes (nouns, verbs, auxiliaries, prepositions, adjectives, adverbs, conjunctions, determiners/articles, pronouns) were summed using EVAL, then the number of bound morphemes was divided by the total number of bound and free morphemes.
Closed class words (proportion)	%mor tier codes for closed class words (auxiliaries, prepositions, conjunctions, determiners/articles, pronouns) and open class words (nouns, verbs, adjectives, adverbs) were summed using EVAL, then the number of closed class words was divided by the total number of open and closed class words.
Pronouns (proportion)	%mor line codes for nouns and pronouns were summed using EVAL, then the number of pronouns was divided by the total number of nouns and pronouns.
Agrammatic utterances (phw)	Postcode [+gram] was summed using FREQ and expressed per hundred words.
Pauses (phw)	Codes (.), (..) and (…) for unfilled pauses, and &ah, &eh, &ew, &hm, &mm, &uh, &uhm, and &um for filled pauses.
Words per minute	Calculated using EVAL based on time-stamped codes embedded in the transcript files.
Retraced sequences (phw)	Codes [/] [//] for retracings.
False starts (phw)	Words beginning with &, except for those denoting gestures (identified manually) or filled pauses (enumerated above).
Unintelligible sequences (phw)	Words coded as xxx.

Note. phw = per hundred words.

Figure 1. Distribution and interrater reliability of the 27 connected speech features. Each row shows one connected speech feature. The first column shows the distribution of the 24 patients' scores, where each patient's score is the mean of the three researchers' ratings. Boxes: interquartile ranges; whiskers: ranges excluding outliers; circles: outliers; red lines: medians; blue asterisks: means. The second column shows the intraclass correlation coefficient (ICC), type A,*k* for the three researchers. This is the expected correlation between scores averaged across the three researchers, and scores averaged across three different hypothetical researchers from the same population of researchers. Error bars indicate 95% confidence intervals. The third column shows the ICC, type A,1, for the three researchers. This is the expected correlation between pairs of researchers from the population of researchers. The fourth column shows the distribution of the 24 patients' scores, where each patient's score is the mean of four student ratings (only four of the 12 students rated each patient). Red lines: medians; blue asterisks: means; black circles: outliers. The fifth column shows the ICC, type 1,*k*, for the students. This is the expected correlation between scores averaged across four students, and scores averaged across a different set of four students, with all students drawn at random from the population of students. The sixth column shows the ICC, type 1, for the students. This is the expected correlation between pairs of students from the population of students.



where patients were rated by a single researcher, was excellent for seven features, good for six features, fair for 11 features (.40 ≤ *r* < .60), and poor for three features (*r* ≤ .40; see Figure 1, Column 3).

Student Clinicians

ICC(*k*), an estimate of reliability where ratings were averaged across four students drawn from the population of students described, was excellent for 11 features, good for 12 features, fair for two features, and poor for two features (see Figure 1, Column 5). ICC(1), an estimate of reliability where patients were rated by single random students drawn from the population of students described, was excellent for four features, good for seven features, fair for 12 features, and poor for three features (see Figure 1, Column 6).

Comparison Between Researchers and Student Clinicians

The mean ICCs (across features) of the three researchers were very similar (S. M. W.: $r = .68$; K. R.: $r = .69$; M. C.: $r = .69$), whereas the student clinicians were much more variable (mean $r = .56$, $SD = .11$, range: .42–.70). The distributions of the two groups were significantly different (two-sample Kolmogorov–Smirnov test = 0.75, $p = .033$, one tailed). However, it is noteworthy that at least three of the student clinicians were as reliable as the researchers (means of $r = .68$, .70, and .70) and another three were in the vicinity ($r = .60$, .62, and .63), suggesting that a subset of student clinicians who will perform comparably to experienced researchers can be identified. Given that the researchers were more reliable than the students as a group and the fact that each researcher rated all 24 samples, our subsequent investigations of concurrent validity and factor analysis were carried out based on the means of the three researchers' scores.

Concurrent Validity

Concurrent validity was assessed by examining correlations between APROCSA features and the measures derived from AphasiaBank (see Figure 2). As described above, for 23 of the 27 APROCSA features, one or more AphasiaBank measures were identified a priori as representing similar or related constructs; these are outlined in yellow in Figure 2. Of these 23 APROCSA features, 19 showed strong ($|r| \geq .5$) and statistically significant correlation(s) with one or more of the relevant measure(s), strongly supporting the validity of APROCSA. For example, correlations between the APROCSA feature *Omission of function words* and the related CHAT transcript measures *Closed class words (proportion)* and *Agrammatic utterances (per hundred words)* were $r = -.70$ and $r = .90$ respectively.

One of the 23 features, *Off-topic* was significantly but not strongly ($r = .47$) correlated with its corresponding measure of *Comprehension,* whereas three features—*Semantic paraphasias, Phonemic paraphasias,* and *Conduite d'approche*—did not exhibit significant correlations with their corresponding AphasiaBank measure(s).

Patterns of Feature Co-occurrence

Correlations Among APROCSA Features

Pearson correlations between each pair of APROCSA features were computed (see Figure 3). Not surprisingly, there were many instances in which pairs of APROCSA features correlated strongly ($|r| \geq .5$) with one another. For example, the correlation between *Omission of bound morphemes* and *Omission of function words,* two features associated with agrammatism, was $r = .92$.

Factor Analysis

Patterns of co-occurrence among the APROCSA features were identified using factor analysis (see Figure 4). A model with four factors provided the most explanatory dimensionality reduction of the data, accounting for 79.5% of the variance. We labeled the factors Paraphasia, Logopenia (paucity of speech), Agrammatism, and Motor Speech, based on the features that loaded on them (see Figure 4 and the Discussion section). The eigenvalues of these factors were 5.31, 5.21, 4.38, and 3.39, and the percentage of variance explained was 23.1%, 22.6%, 19.1%, and 14.7%, respectively. Communality values of the APROCSA features ranged from 0.56 to 0.97, indicating that a high proportion of the variance for each feature was explained by the four factors.

Models with fewer than four factors conflated one or more of these four factors and explained substantially less of the variance in the data. In particular, a two-factor model conflated the Logopenia, Agrammatism, and Motor Speech factors and explained 58.8% of the variance, whereas a three-factor model conflated the Logopenia and Motor Speech factors and explained 70.2% of the variance. In contrast, a five-factor model yielded four factors similar to those identified in the four-factor model, as well as an additional factor with an eigenvalue of 0.73 (i.e., < 1) that explained only 3.2% of the variance, and the factor loadings of which had no evident interpretation.

Factor Loadings by Patient

The factor loadings for individual patients were plotted and showed considerable diversity among patients of any given aphasia type (see Figure 5). For example, of the four patients with Broca's aphasia, Scale33a loaded most heavily on Agrammatism and Motor Speech, TCU08a loaded on Agrammatism and to a lesser extent Logopenia, TAP11a loaded on Logopenia, and BU08a loaded most heavily on Motor Speech with lesser loadings on the other three factors. Similarly, the patients with Wernicke's aphasia were highly diverse: Only one of the four (Elman14a) showed the expected loading on Paraphasia, whereas Thompson05a loaded most heavily on Agrammatism, Kurland18a loaded on Logopenia, and Elman12a showed no positive loadings.

Conversely, in many cases, patients with similar connected speech profiles met WAB-R criteria for different aphasia types. For instance, Kurland07a and Elman14a both loaded most heavily on Paraphasia, with near-zero loading on Logopenia and negative loadings on Agrammatism and Motor Speech, but Kurland07a met criteria for Anomic aphasia whereas Elman14a met criteria for Wernicke's aphasia.

Discussion

Our results showed that most of the features of connected speech in aphasia we defined were rated with good-to-excellent interrater reliability by researchers and student clinicians. Most features demonstrated strong concurrent validity with respect to quantitative connected speech measures computed from AphasiaBank transcripts and clinical measures. Factor analysis showed that four readily interpretable underlying factors accounted for 79% of the variance in connected speech profiles. Taken together, these findings indicate that the APROCSA is a promising scheme

Figure 2. Concurrent validity of the 27 connected speech features. Pearson correlation coefficients are indicated by depth of color, and *r* values are shown for correlations with uncorrected $p < .05$. The *y* axis shows the 27 connected speech features. The *x*-axis shows the 17 quantitative measures derived from the transcription and coding of the speech samples in AphasiaBank, and the five subscores and the Aphasia Quotient from the Western Aphasia Battery–Revised (WAB-R). Auditory-perceptual rating of connected speech in aphasia (APROCSA) connected speech features were all defined such that high scores are indicative of impairment. The other measures differ in terms of their directionality. In general, the blue color scale is used to encode correlations of scores indicating impairment with scores indicating impairment, whereas the red color scale is used to encode correlations of scores indicating impairment with scores indicating sparing. Exceptions to this are three AphasiaBank quantitative measures—bound morphemes (proportion), closed class words (proportion), and pronouns (proportion)—because these measures can be perturbed in either direction in aphasia (Wilson et al., 2010). The perturbation of these scores in the "agrammatic" direction was arbitrarily defined as the direction of impairment. Yellow boxes indicate AphasiaBank measures that were considered a priori to be measuring the same or related phenomena to each connected speech feature.

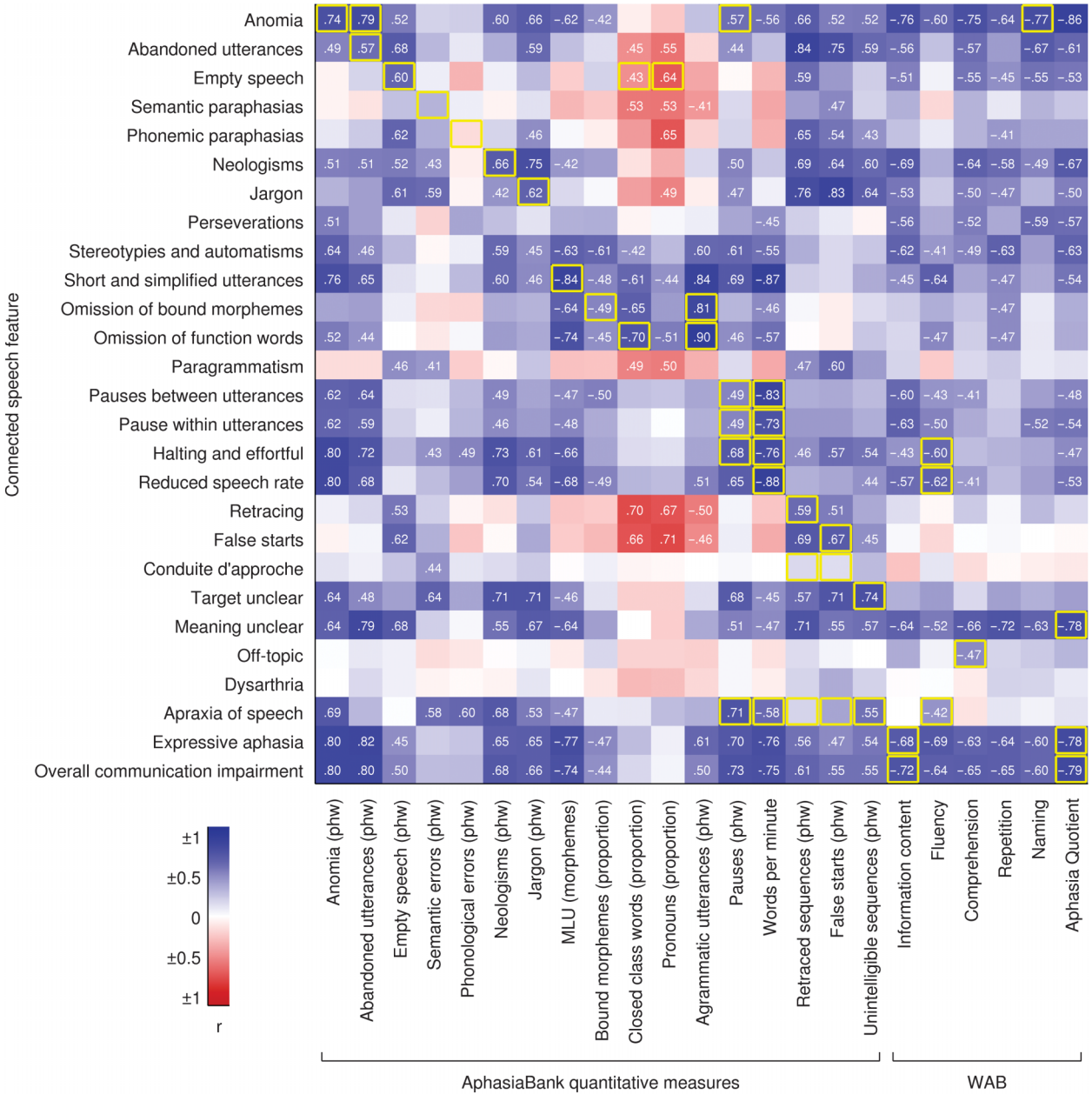
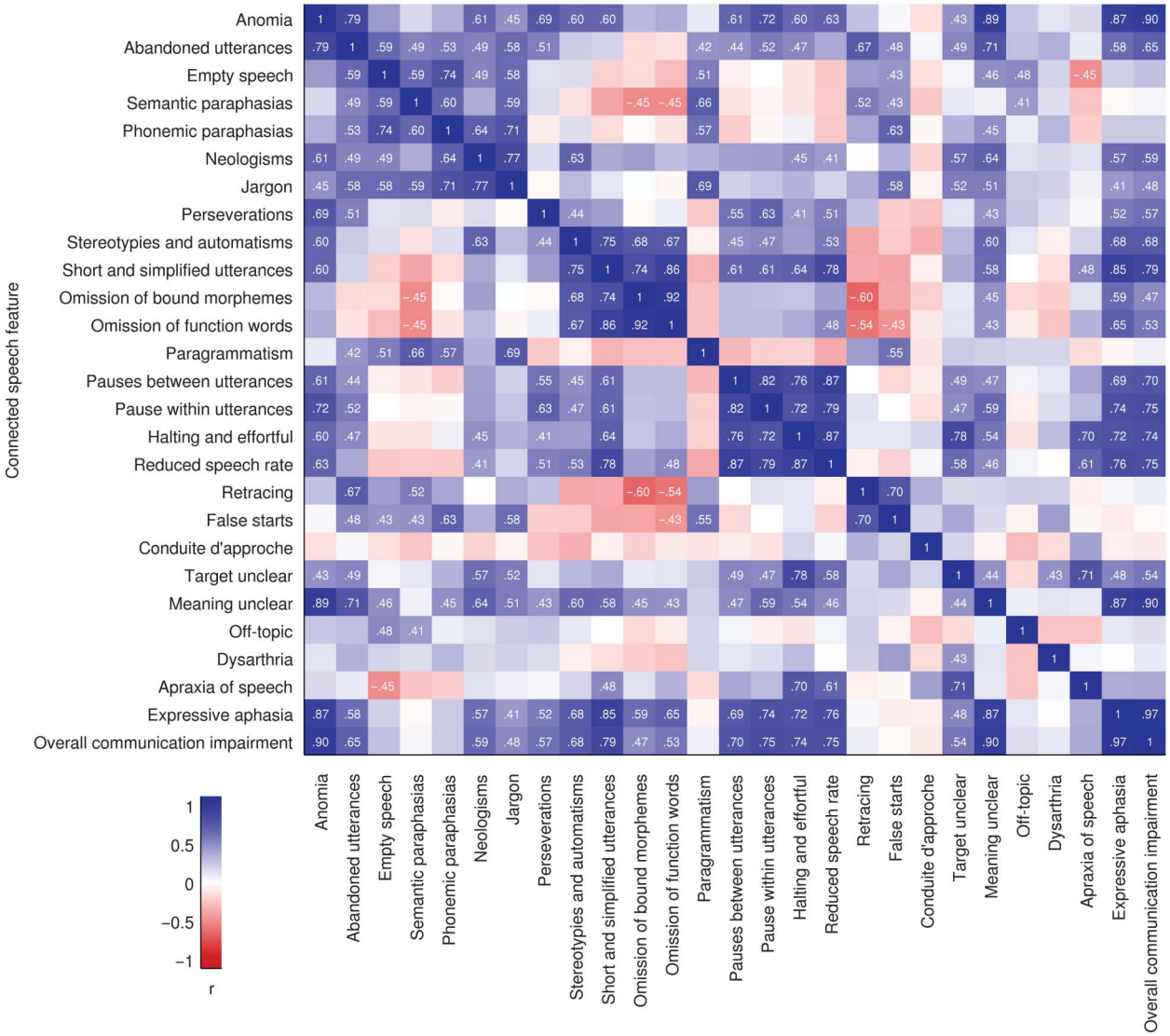


Figure 3. Patterning of connected speech features: correlation matrix. Each feature is shown on both the *x*- and *y*-axes, so the matrix is symmetric around the diagonal. Positive correlations are indicated in blue, and negative correlations are in red. Pearson *r* values are shown for correlations with uncorrected $p < .05$.



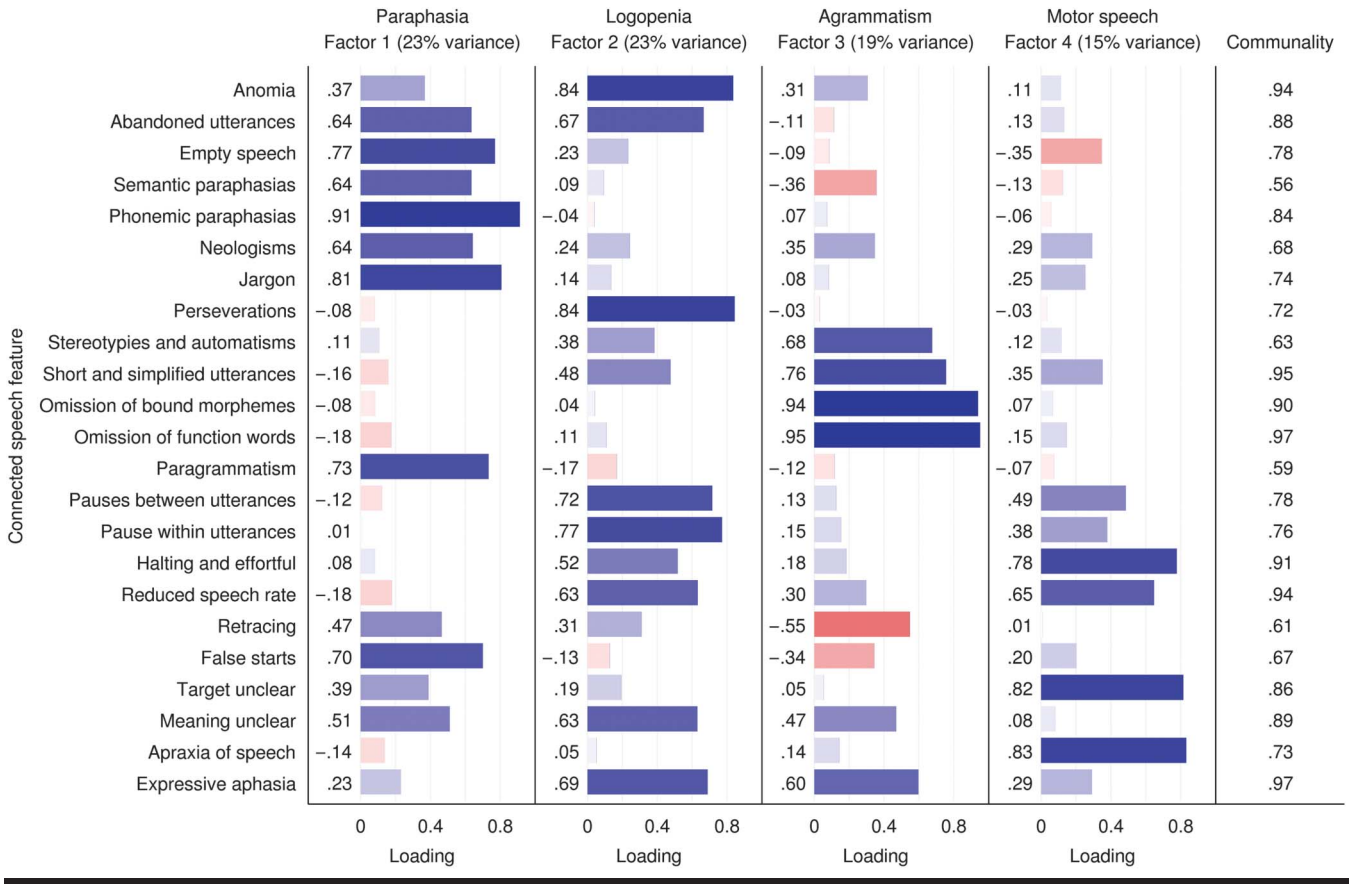
for comprehensive, efficient, reliable, and valid characterization of connected speech in aphasia.

Interrater Reliability

Interrater reliability was good or excellent for most features for both researchers and student clinicians, so long as scores were averaged across multiple raters. Not surprisingly, scores from single raters were less reliable, even when those raters were experienced researchers.

As expected, experienced researchers were generally more reliable than student clinicians, but a subset of student clinicians performed comparably to researchers. This suggests that some students with adequate training and clinical experience in aphasia are capable of being excellent raters, without necessarily needing to have extensive training or experience in quantitative connected speech analysis. Note that all students were second-year master's students in speech-language pathology, who had completed graduate coursework in aphasia and had at least 25 hr of clinical experience in aphasia; thus, they were far from naive listeners. We do not think it would be feasible for less knowledgeable listeners to reliably rate APROCSA features, because the feature definitions (see Table 1) presuppose substantial knowledge of linguistics and aphasia. Further research is warranted to determine to what extent the performance of inexperienced raters, such as student clinicians, can be improved by additional training and also to investigate the extent of training required for certified speech-language pathologists to become excellent raters.

Figure 4. Patterning of connected speech features: factor analysis. Only 23 of the 27 features were used, because there were only 24 patients. A four-factor rotated model provided the most explanatory account of the data. The factors were labeled Paraphasia, Logopenia, Agrammatism, and Motor Speech. Loadings of each feature on each factor are shown and accompanied by bars: positive in blue and negative in red. Communality indicates the proportion of variance of each feature that was explained by the four factors.

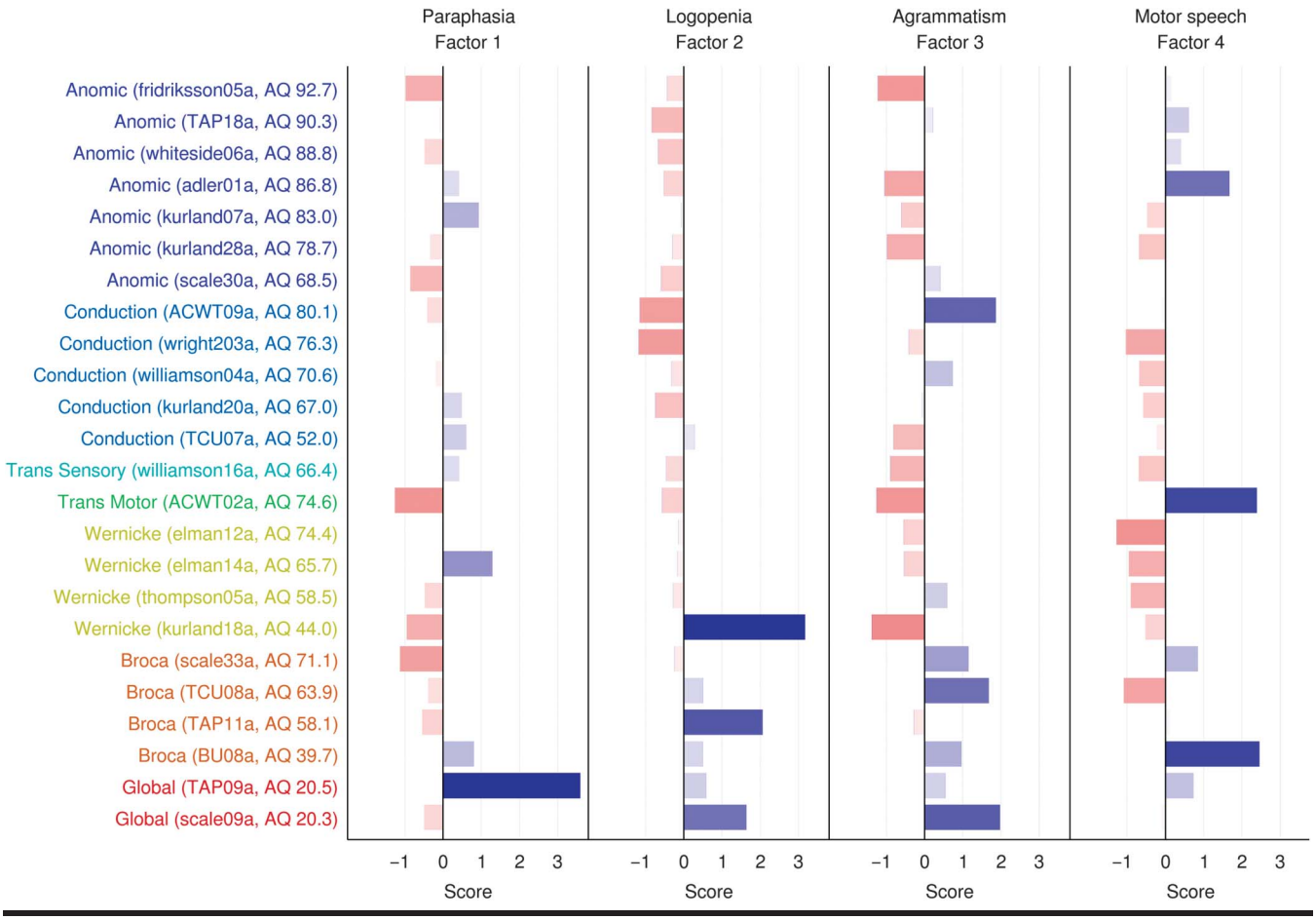


The specific design of our study was intended to evaluate two ways in which the APROCSA might be used in practice in research contexts; that is, a group of key personnel on a study team might rate all patients (which is feasible because of how quickly the APROCSA can be scored), or student clinicians might rate patients in the role of research assistants. In the former case, the same set of raters would rate each patient, whereas in the latter case, turnover of research assistants must be anticipated, which is why we modeled these two situations in our experimental design. Our results showed that both of these approaches are feasible in principle. Because there was considerable variability among the student clinicians in terms of their ability to provide ratings in concordance with ratings of experienced researchers, potential raters would need to be trained and screened for this ability.

Not all APROCSA features were scored with the same degree of interrater reliability. The *Phonemic paraphasias* feature was relatively unreliable, which probably reflects the well-known difficulty of differentiating between phonemic and apraxic errors. It is notable though that two other features impacted by phonological errors—*Neologisms* and *Jargon*—showed excellent interrater reliability, as did the *Apraxia of speech* feature. Another relatively unreliable feature was *Off-topic,* which might have been difficult to judge because it requires the rater to make inferences with respect to context, which was sometimes lacking in the 5-min excerpts.

Previous studies have investigated the interrater reliability of several quantitative and qualitative connected speech measures. Rochon et al. (2000) reported interrater reliability for 12 measures derived from Quantitative Production Analysis (Saffran et al., 1989); ICCs ranged from 0.89 to 0.98, indicating excellent reliability. Of note, these reflect ratings of samples that were already segmented into utterances; in practice, additional variability would arise from this step. Excellent interrater reliability for Quantitative Production Analysis was also reported by Gordon (2006). To our knowledge, the interrater reliability of connected speech measures derived from CHAT coding of speech samples has not been investigated for individuals with aphasia, only for neurologically normal participants (Hancock, Stutts, & Bass, 2015; Richardson & Dalton, 2016). Interrater reliability of CHAT transcription (but not coding) of

Figure 5. Characteristics of individuals with aphasia. For each of the 24 patients, the scores on each of the four factors are shown. Patients are ordered by aphasia type per the Western Aphasia Battery–Revised (WAB-R), with less severe types first and then by descending aphasia quotient (AQ) within type.



aphasic connected speech has been reported to be excellent (Fergadiotis & Wright, 2011), but that study included only individuals with mild or moderate anomic aphasia or conduction aphasia; it can be anticipated that aphasias involving coexisting apraxia of speech would be transcribed less reliably. For qualitative scales, interrater reliability appears to depend on who is rating the samples. With experienced raters, excellent reliability has been reported for the profile of speech characteristics of the BDAE (Goodglass & Kaplan, 1972), the fluency scale of the WAB (Kertesz, 1979), and four qualitative measures that contributed to a hybrid quantitative–qualitative approach to connected speech assessment (Wagenaar et al., 1975). However, much less consistency was observed in studies where qualitative scales from the BDAE and the WAB were rated by community speech-language pathologists (Gordon, 1998; Trupe, 1984).

The interrater reliability of APROCSA compares favorably to the interrater reliability of auditory-perceptual rating of motor speech disorders. Darley et al. (1969a) concluded that the agreement and reliability they observed were adequate for purposes of clinical assessment and research aimed at identifying perceptual features that play a prominent role in the presenting speech disorder. Bunton et al. (2007) carried out a study in which 20 raters evaluated 47 patients of varying dysarthria types on 38 perceptual features based on a 40-s sample of conversational speech. Differences between patients accounted for 36%–62% of the variance depending on the variable, which corresponds to partial r from .60 to .79, indicating that the reliability for most measures was in the good range.

Concurrent Validity

Most of the APROCSA features showed strong concurrent validity with respect to quantitative measures of connected speech calculated from AphasiaBank transcripts and/or standard clinical measures derived from the WAB-R, as reflected by strong correlations between features and measures representing similar or related constructs. It is noteworthy that such high correlations were observed between APROCSA features and quantitative measures,

when the former are so much quicker to obtain. The total time taken for three researchers to rate a 5-min speech sample was only about 10 min per rater (i.e., 30 min total), whereas transcription and coding of aphasic connected speech with CHAT takes approximately 30 min per minute of recorded speech for an experienced coder (Davida Fromm, personal communication).

There were just a few APROCSA features that did not show strong correlations with any expected transcript- or battery-based measures. These instances may reflect weaknesses in the APROCSA features, the transcript- or battery-based measures, or both. Neither *Semantic paraphasias* nor *Phonemic paraphasias* correlated with related transcript-based measures *Semantic errors (per hundred words)* or *Phonological errors (per hundred words)*. In part, this probably reflects challenges in scoring these APROCSA features; as noted above, the *Phonemic paraphasias* feature was one of the least reliable, and *Semantic paraphasias* showed only good, not excellent, reliability when averaged across the three researchers. However, limitations in the relevant CHAT-based measures probably contribute even more to the failure to observe correlations. In particular, real word errors in CHAT are coded [*s] (semantic error) even if they may have resulted from a phonemic paraphasia that happened to result in a real word (MacWhinney, 2000). This implies that to derive more robust measures of semantic and phonemic errors from CHAT transcripts, it would be necessary to manually disambiguate [*s] errors with respect to some phonemic criteria; we did not attempt this, because we were only concerned with transcript-based measures that could be derived from existing transcriptions and codes. Another issue is that many errors that we attributed to apraxia of speech are coded with [*p] (phonological error) and sometimes [*s] (semantic error) codes. This is a limitation of CHAT, which provides few tools for the transcription or analysis of apraxic phenomena, nor does any other transcription system we are aware of (see Vermeulen et al., 1989, for further discussion). *Conduite d'approche* was another feature that did not correlate with a priori measures *Retraced sequences (per hundred words)* or *False starts (per hundred words)*. The lack of correlations in this case may reflect the fact that these are only somewhat related measures; there is no explicit code for *Conduite d'approche* in CHAT. Finally, *Off-topic* correlated significantly but not strongly with the WAB-R Comprehension subscore. We had chosen this a priori measure based on the assumption that many instances when a patient's speech appears to be off topic actually reflect comprehension impairments on the part of the patient; however, clearly *Off-topic* and *Comprehension* are only tangentially related constructs.

Four APROCSA features—*Paragrammatism, Perseverations, Stereotypies and automatisms,* and *Dysarthria*—did not have related AphasiaBank measures. Although there are CHAT codes that could capture paragrammatism ([* m], [* f], [+gram]), perseverations ([+per]), and stereotypies ([*n:uk:s]), they were rarely or never used in the AphasiaBank transcripts of our selected samples, with the exception of [+gram], which was used almost exclusively for agrammatic utterances. A clinical diagnosis of dysarthria was provided in AphasiaBank, but this was not a viable correlate for the *Dysarthria* feature due to its binary nature; moreover, only two of the 24 patients we rated presented with dysarthria.

Only one previous study to our knowledge has compared qualitative and quantitative measures of connected speech in aphasia (Grande et al., 2008). In that study, five quantitative measures derived from transcripts were compared to the qualitative speech rating scales from the Aachen Aphasia Test (Huber et al., 1984) for their utility in detecting treatment-induced changes in 28 individuals with aphasia. Although the authors argued that the quantitative measures were more sensitive to change, this conclusion was undercut by the fact that many of the changes were not in the direction associated with improvement, suggesting that the criterion for inferring a critical difference may not have been set appropriately.

Patterns of Feature Co-occurrence

As expected, many APROCSA features clustered together as evidenced by strong pairwise correlations between features reflecting common underlying impairments. A factor analysis showed that four underlying factors, which we labeled Paraphasia, Logopenia (paucity of speech), Agrammatism, and Motor Speech, accounted for much of the variance. Every feature contributed to one or more of the four factors, as indicated by high communality scores, and there were only a handful of features that were weighted heavily on more than one factor.

The Paraphasia factor loaded most heavily on *Abandoned utterances, Empty speech, Semantic paraphasias, Phonemic paraphasias, Neologisms, Jargon, Paragrammatism, Retracing,* and *False starts*. These features are generally reflective of selection errors in semantic, syntactic, and phonological domains, which were often self-corrected. Many of these features would be characteristically associated with fluent aphasias (Benson, 1967), but they can, of course, occur in nonfluent aphasias, too.

The Logopenia (paucity of speech) factor loaded heavily on *Anomia, Abandoned utterances, Perseverations, Pauses between utterances, Pauses within utterances, Halting and effortful speech,* and *Reduced speech rate,* reflecting prominent word-finding difficulties leading to slow, labored speech. Accordingly, the *Meaning unclear* and *Expressive aphasia* features were also highly weighted. Note that the label of this feature should not be confused with logopenic progressive aphasia, which is associated with many of these features, but also with phonological encoding impairments (Wilson et al., 2010), which in our analysis loaded instead on the Paraphasia factor.

The Agrammatism factor loaded most heavily on three features clearly associated with agrammatism: *Short and simplified utterances, Omission of bound morphemes,* and *Omission of function words*. Also highly weighted were *Stereotypies and automatisms* and *Expressive aphasia*. Agrammatism was sharply distinguished from

paragrammatism, with negative weightings on factors associated with paragrammatism such as *Paragrammatism* and *Retracing*.

The Motor Speech factor loaded most heavily on four factors clearly associated with apraxia of speech: *Halting and effortful, Reduced speech rate, Target unclear,* and *Apraxia of speech,* and to a lesser extent on pausing between and within utterances.

Taken together, the Logopenia, Agrammatism, and Motor Speech factors represented a parcellation of nonfluency into three components, each of which was clearly distinct. Much previous research has shown that fluency is a multifactorial concept reflecting many dimensions of speech production (Benson, 1967; Goodglass, Quadfasel, & Timberlake, 1964; Gordon, 1998; Kerschensteiner, Poeck, & Brunner, 1972) and that these dimensions can dissociate (Miceli, Mazzucchi, Menn, & Goodglass, 1983; Thompson et al., 2012; Wilson et al., 2010). Our data-driven approach suggests that, in a poststroke cohort, there are specifically three major dimensions of nonfluency. The other factor, Paraphasia, was not simply the opposite of nonfluency (Benson, 1967) but could occur alone or in conjunction with any or all of the nonfluent dimensions.

Although many researchers have carried out factor analyses and other multivariate analyses of measures derived from aphasia batteries (Butler, Lambon Ralph, & Woollams, 2014; Goodglass & Kaplan, 1972; Kertesz, 1979; Mirman et al., 2015; Swinburn, Porter, & Howard, 2004), to our knowledge only a handful of studies have reported multivariate analyses of variables derived from connected speech analysis. Benson (1967) qualitatively rated 10 measures of connected speech, but his analysis revolved around summing the measures to obtain a single fluency rating, which he argued was bimodally distributed; see also Howes (1967) and Kerschensteiner et al. (1972) for other early approaches along similar lines.

Two Dutch studies based primarily on quantitative analyses of connected speech in partially overlapping poststroke aphasia cohorts reported factor analyses similar to our approach (Vermeulen et al., 1989; Wagenaar et al., 1975). We will compare our findings to the latter of these studies, because it has several methodological advantages over the former (see Vermeulen et al., 1989, for discussion). Vermeulen et al. (1989) applied factor analysis to 17 features derived from quantitative analysis of connected speech, along with a confrontation naming measure. There were many similarities between their variables and the 23 features we included in our factor analysis, but there were also important differences. For example, some features in our analysis corresponded to multiple variables in theirs (e.g., we had only one *Phonemic paraphasias* feature, whereas they had separate measures for transpositions, additions, substitutions, and consonant cluster reductions). Conversely, several of our features had no counterparts in their approach, specifically *Abandoned utterances, Stereotypies and automatisms, Perseverations, Omission of bound morphemes, Paragrammatism, Pauses between utterances, Pauses within utterances, Halting and effortful, Retracing,* and *Meaning unclear*.

Vermeulen et al.'s (1989) factor analysis yielded five latent factors, which they labeled Syntactic ability, Phonological paraphasia, Neologistic paraphasia, Articulatory impairment, and Vocabulary, and which together accounted for 49.5% of the variance. There was generally a close correspondence between our factors and their factors: Our three nonfluent factors each corresponded to one of theirs: Logopenia to Vocabulary, Agrammatism to Syntactic ability, and Motor Speech to Articulatory impairment, whereas our Paraphasia factor was parcellated into phonological and semantic factors in Vermeulen et al.'s analysis, labeled *Phonological paraphasia* and *Neologistic paraphasia.* For the most part, the variables loading on each factor were similar across the two analyses, with just a few notable exceptions: speech rate loaded on Syntactic ability for Vermeulen et al., but on Logopenia and Motor Speech in our analysis; false starts (which they termed *literal perseverations*) loaded on Articulatory impairment for Vermeulen et al., but on Paraphasia in our analysis; and Empty speech loaded on Vocabulary for Vermeulen et al., but on Paraphasia in our analysis. Although these divergences will need to be resolved with larger and more diverse data sets, the rather striking correspondences between the two analyses suggest that the factors identified reflect coherent underlying deficits that account for many of the observable features of connected speech in chronic poststroke aphasia.

We observed remarkably different factor profiles among patients who met WAB-R criteria for each aphasia type. Although variability within aphasia types is expected and has been extensively documented (Kertesz, 1979), the factor profiles derived from the APROCSA often did not match classic conceptions of the nature of language production in each aphasia type (Goodglass, 1993). Based on these observations, we listened again to the samples of many of the patients to subjectively evaluate whether their factor loadings faithfully reflected the major characteristics of their connected speech. In all cases, we were satisfied that the factor loadings indeed provided an accurate picture. Interested readers can readily assess this claim by listening to the samples we rated, which are freely available to aphasia researchers and clinicians on AphasiaBank, and considering them in relation to the factor profiles of each patient shown in Figure 5.

Limitations

Our study had several noteworthy limitations. First and foremost, the auditory-perceptual approach is inherently subjective. Raters may bring preconceived notions about which features are likely to pattern together. The definitions of the features, as presented in this article, are brief and leave room for interpretation, and the specific ways that raters interpret each feature will depend not only on their training and experience but also on the specific manner in which they are trained to score the APROCSA.

Second, although we attempted to define a set of features that would cover the range of deficits that commonly occur in aphasias of diverse etiologies, it is not difficult to conceive of additional features that may be useful. For example, none of the 27 features captures "disinclination to speak," which is a hallmark of transcortical motor aphasia and can occur in other nonfluent aphasias (Greenwald, Nadeau, & Gonzalez Rothi, 2000). Other features could arguably be subdivided; for instance, the *Halting and effortful* feature subsumes prosodic disturbances, which could have been rated separately. The APROCSA would also need to be modified for languages that are significantly typologically different from English; for instance, the *Omission of bound morphemes* feature would have no relevance in a highly isolating language, whereas languages with richer inflectional morphology than English may require more elaboration of features capturing morphological disruption.

Third, although we established adequate interrater reliability of APROCSA, we did not investigate test–retest reproducibility. In other words, if two speech samples were acquired from the same patient on two different occasions, would they be scored the same way? Test–retest reproducibility will depend on factors such as the extent to which diagnostic behaviors of interest occur sufficiently frequently in a 5-min sample to quantify their prevalence (Boles & Bombard, 1998); differences in connected speech features based on the topic of conversation or the nature of the elicitation task (Armstrong, 2000); the extent to which varying situational, motivational, physiological, and cognitive factors impact connected speech features; and in some individuals, multiple available registers, such as a patient with Broca's aphasia who alternated between telegraphic and nontelegraphic modes of communication (Bastiaanse, 1995).

Fourth, the patient sample consisted only of individuals with chronic poststroke aphasia. Although there is no reason to think that interrater reliability and concurrent validity would be different for aphasias of other etiologies, this will need to be established. On the other hand, the factors that emerged from the factor analysis almost certainly do reflect the nature of the sample. For example, the Paraphasia factor loaded on connected speech features including semantic paraphasias, phonemic paraphasias, and empty speech, indicating that these features tend to co-occur in chronic poststroke aphasia, but it has been previously shown that these features dissociate in different variants of primary progressive aphasia (Wilson et al., 2010).

Fifth, the number of patients in our study was quite small for factor analysis. Although the clear interpretability of the factors and the convergence of our findings with those of Vermeulen et al. (1989) were encouraging, it will be necessary to perform a similar analysis with a much larger group of patients to determine whether or not these specific factors are robust. Ideally, an analysis with a group of patients that is not only larger but is also diverse in terms of etiology would probably yield a larger set of factors that might provide insights into the similarities and differences between aphasias arising from different underlying causes.

Clinical Applications

Clinicians may find the APROCSA scoresheet to be a helpful tool for systematically quantifying the prevalence of many of the features that commonly occur in connected speech in aphasia. After rating a patient on each feature, the patient's profile can be considered in relation to the loadings of the four factors shown in Figure 4: Paraphasia, Logopenia, Agrammatism, and Motor Speech. We believe that conceptualizing a patient's connected speech in terms of these four dimensions, rather than in terms of the traditional fluent/nonfluent dichotomy, will lead to a greater appreciation of the patient's strengths and weaknesses, which can inform the development of treatment goals and approaches. The APROCSA is an assessment of connected speech, not a comprehensive aphasia assessment, so it should be used in conjunction with an aphasia battery that includes constrained assessments of speech and language production such as confrontation naming, repetition, and a motor speech evaluation, as well as assessments of comprehension of words and sentences.

Our findings motivate the future development of a clinical tool based on the APROCSA. The auditory-perceptual approach may be particularly attractive in assessment of acute poststroke aphasia due to its ease of administration, as patients often are unable to withstand prolonged testing and may present with multiple comorbidities (e.g., dysarthria, dysphagia) that require evaluation. Development of a clinical tool will require several steps, including (a) development of a formal training module based on speech samples from individuals with aphasia that are collected with appropriate informed consent and institutional review board oversight for this purpose; (b) development of an app (e.g., web, Android, iPad) so that feature scores can be entered and factor scores computed easily; (c) factor analysis of a larger sample of individuals with aphasias of diverse etiologies, in order to obtain more robust factor scores; and (d) quantification of interrater reliability and test–retest reliability for clinicians with different levels of experience and determination of the extent of training required.

Research Applications

The APROCSA is sufficiently developed to be used by researchers to characterize connected speech in research contexts such as studies of treated or spontaneous recovery from aphasia or lesion-symptom mapping. We have provided guidelines on applying the APROCSA in research contexts, as well as a MATLAB script to facilitate training and factor analysis on our website, http://www.aphasialab.org/aprocsa.

Conclusion

The APROCSA shows potential to be a comprehensive, efficient, reliable, and valid approach for characterizing connected speech in aphasia. It can be applied in research contexts as described in this article, and with further development, it has potential to become an easy-to-use clinical tool that combines the best features of quantitative linguistic analysis and qualitative rating scales for assessment of connected speech in aphasia.

Acknowledgments

This research was supported in part by the National Institute on Deafness and Other Communication Disorders Grants R01 DC013270 (awarded to S. M. W.), R21 D016080 (awarded to S. M. W.), and R01 DC007646 (awarded to P. M. B.). We thank the student raters who evaluated the speech samples; Marja-Liisa Mailend, Chelsea Bayley, and Audrey Holland for their advice and input in the development of the auditory-perceptual rating of connected speech in aphasia; Davida Fromm and Brian MacWhinney for their assistance in utilizing the AphasiaBank database; the individuals with aphasia who consented to share their speech samples through AphasiaBank; and the researchers who made these data available to the community.

References

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology, 14,* 875–892.

Bastiaanse, R. (1995). Broca's aphasia: A syntactic and/or a morphological disorder? A case study. *Brain and Language, 48,* 1–32.

Bastiaanse, R., Edwards, S., & Kiss, K. (1996). Fluent aphasia in three languages: Aspects of spontaneous speech. *Aphasiology, 10,* 561–575.

Benson, D. F. (1967). Fluency in aphasia: Correlation with radioactive scan localization. *Cortex, 3,* 373–394.

Boles, L., & Bombard, T. (1998). Conversational discourse analysis: Appropriate and useful sample sizes. *Aphasiology, 12,* 547–560.

Boyle, M. (2014). Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research, 57,* 966–978.

Boyle, M. (2015). Stability of word-retrieval errors with the AphasiaBank stimuli. *American Journal of Speech-Language Pathology, 24,* S953–S960.

Bunton, K., Kent, R. D., Duffy, J. R., Rosenbek, J. C., & Kent, J. F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language, and Hearing Research, 50,* 1481–1495.

Butler, R. A., Lambon Ralph, M. A., & Woollams, A. M. (2014). Capturing multidimensionality in stroke aphasia: Mapping principal behavioural components to neural structures. *Brain, 137,* 3248–3266.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969a). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12,* 246–269.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969b). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research, 12,* 462–496.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). *Motor speech disorders*. Philadelphia, PA: Saunders.

Doyle, P. J., Goda, A. J., & Spencer, K. A. (1995). The communicative informativeness and efficiency of connected discourse by adults with aphasia under structured and conversational sampling conditions. *American Journal of Speech-Language Pathology, 4,* 130–134.

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management* (3rd ed.). St. Louis, MO: Elsevier/Mosby.

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25,* 1414–1430.

Goodglass, H. (1993). *Understanding aphasia*. San Diego, CA: Academic Press.

Goodglass, H., & Kaplan, E. (1972). *The assessment of aphasia and related disorders*. Philadelphia, PA: Lea & Febiger.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *The Boston Diagnostic Aphasia Examination–Third Edition* (BDAE-3). Baltimore, MD: Lippincott, Williams & Wilkins.

Goodglass, H., Quadfasel, F. A., & Timberlake, W. H. (1964). Phrase length and the type of severity of aphasia. *Cortex, 1,* 133–153.

Gordon, J. K. (1998). The fluency dimension in aphasia. *Aphasiology, 12,* 673–688.

Gordon, J. K. (2006). A quantitative production analysis of picture description. *Aphasiology, 20,* 188–204.

Grande, M., Hussmann, K., Bay, E., Christoph, S., Piefke, M., Willmes, K., & Huber, W. (2008). Basic parameters of spontaneous speech as a sensitive method for measuring change during the course of aphasia. *International Journal of Language & Communication Disorders, 43,* 408–426.

Greenwald, M. L., Nadeau, S. E., & Gonzalez Rothi, L. J. (2000). Fluency. In L. J. Gonzalez Rothi, B. Crosson, & S. E. Nadeau (Eds.), *Aphasia and language: Theory to practice* (pp. 31–39). New York, NY: Guilford.

Hancock, A. B., Stutts, H. W., & Bass, A. (2015). Perceptions of gender and femininity based on language: Implications for transgender communication therapy. *Language and Speech, 58,* 315–333.

Haravon, A., Obler, L., & Sarno, M. (1994). A method for microanalysis of discourse in brain-damaged patients. In R. Bloom, L. Obler, S. De Santi, & J. Ehrlich (Eds.), *Discourse analysis and applications: Studies in adult clinical populations* (pp. 47–80). Hillsdale, NJ: Erlbaum.

Howes, D. (1967). Some experimental investigations of language in aphasia. In K. Salzinger & S. Salzinger (Eds.), *Research in verbal behaviour and some neurophysiological implications*. New York, NY: Academic Press.

Huber, W., Poeck, K., & Willmes, K. (1984). The Aachen Aphasia Test. *Advances in Neurology, 42,* 291–303.

Kent, R. D., Kent, J. F., Duffy, J., & Weismer, G. (1998). The dysarthrias: Speech-voice profiles, related dysfunctions, and neuropathology. *Journal of Medical Speech-Language Pathology, 6,* 165–211.

Kerschensteiner, M., Poeck, K., & Brunner, E. (1972). The fluency–nonfluency dimension in the classification of aphasic speech. *Cortex, 8,* 233–247.

Kertesz, A. (1979). *Aphasia and associated disorders: Taxonomy, localization, and recovery*. New York, NY: Grune & Stratton.

Kertesz, A. (2007). *Western Aphasia Battery–Revised*. New York, NY: Grune & Stratton.

Linnik, A., Bastiaanse, R., & Höhle, B. (2016). Discourse production in aphasia: A current review of theoretical and methodological challenges. *Aphasiology, 30,* 765–800.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25,* 1286–1307.

Marini, A., Andreetta, S., del Tin, S., & Carlomagno, S. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology, 25,* 1372–1392.

McCarron, A., Chavez, A., Babiak, M., Berger, M. S., Chang, E. F., & Wilson, S. M. (2017). Connected speech in transient aphasias after left hemisphere resective surgery. *Aphasiology, 31,* 1266–1281.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

Miceli, G., Mazzucchi, A., Menn, L., & Goodglass, H. (1983). Contrasting cases of Italian agrammatic aphasia without comprehension disorder. *Brain and Language, 19,* 65–97.

Miller, J. F., Andriacchi, K., & Nockerts, A. (2015). *Assessing language production using SALT software: A clinician's guide to language sample analysis* (2nd ed.). Middleton, WI: SALT Software.

Mirman, D., Chen, Q., Zhang, Y., Wang, Z., Faseyitan, O. K., Coslett, H. B., & Schwartz, M. F. (2015). Neural organization of spoken language revealed by lesion-symptom mapping. *Nature Communications, 6,* 6762.

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36,* 338–350.

Prins, R., & Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. *Aphasiology, 18,* 1075–1091.

Richardson, J. D., & Dalton, S. G. (2016). Main concepts for three different discourse tasks in a large non-clinical sample. *Aphasiology, 30,* 45–73.

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

Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language, 37,* 440–479.

Shewan, C. M. (1988). The Shewan Spontaneous Language Analysis (SSLA) system for aphasic adults: Description, reliability, and validity. *Journal of Communication Disorders, 21,* 103–138.

Strand, E. A., Duffy, J. R., Clark, H. M., & Josephs, K. (2014). The Apraxia of Speech Rating Scale: A tool for diagnosis and description of apraxia of speech. *Journal of Communication Disorders, 51,* 43–50.

Swinburn, K., Porter, G., & Howard, D. (2014). *Comprehensive Aphasia Test,* Hove, UK: Psychology Press.

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,* 297–321.

Thompson, C. K., Cho, S., Hsu, C.-J., Wieneke, C., Rademaker, A., Weitner, B. B., . . . Weintraub, S. (2012). Dissociations between fluency and agrammatism in primary progressive aphasia. *Aphasiology, 26,* 20–43.

Trupe, E. H. (1984). Reliability of rating spontaneous speech in the Western Aphasia Battery: Implications for classification. In R. Brookshire (Ed.), *Clinical aphasiology: Proceedings of the conference* (pp. 55–69). Minneapolis, MN: BRK Publisher.

Vermeulen, J., Bastiaanse, R., & Van Wageningen, B. (1989). Spontaneous speech in aphasia: A correlational study. *Brain and Language, 36,* 252–274.

Wagenaar, E., Snow, C., & Prins, R. (1975). Spontaneous speech of aphasic patients: A psycholinguistic analysis. *Brain and Language, 2,* 281–303.

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., . . . Gorno-Tempini, M. L. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain, 133,* 2069–2088.

Yagata, S. A., Yen, M., McCarron, A., Bautista, A., Lamair-Orosco, G., & Wilson, S. M. (2017). Rapid recovery from aphasia after infarction of Wernicke's area. *Aphasiology, 31,* 951–980.

Yorkston, K. M., & Beukelman, D. R. (1980). An analysis of connected speech samples of aphasic and normal speakers. *Journal of Speech and Hearing Disorders, 45,* 27–36.

Appendix

APROCSA Rating Form

Rate connected speech using the following scale:

Score	Severity	Description
0	Not present	Not present or within the range of healthy older speakers
1	Mild	Detectable but infrequent
2	Moderate	Frequently evident but not pervasive
3	Marked	Moderately severe, pervasive
4	Severe	Nearly always evident

Connected speech feature	0	1	2	3	4
Anomia	not present	mild	moderate	marked	severe
Abandoned utterances	not present	mild	moderate	marked	severe
Empty speech	not present	mild	moderate	marked	severe
Semantic paraphasias	not present	mild	moderate	marked	severe
Phonemic paraphasias	not present	mild	moderate	marked	severe
Neologisms	not present	mild	moderate	marked	severe
Jargon	not present	mild	moderate	marked	severe
Perseverations	not present	mild	moderate	marked	severe
Stereotypies and automatisms	not present	mild	moderate	marked	severe
Short and simplified utterances	not present	mild	moderate	marked	severe
Omission of bound morphemes	not present	mild	moderate	marked	severe
Omission of function words	not present	mild	moderate	marked	severe
Paragrammatism	not present	mild	moderate	marked	severe
Pauses between utterances	not present	mild	moderate	marked	severe
Pauses within utterances	not present	mild	moderate	marked	severe
Halting and effortful speech production	not present	mild	moderate	marked	severe
Reduced speech rate	not present	mild	moderate	marked	severe
Retracing	not present	mild	moderate	marked	severe
False starts	not present	mild	moderate	marked	severe
Conduite d'approche	not present	mild	moderate	marked	severe
Target unclear	not present	mild	moderate	marked	severe
Meaning unclear	not present	mild	moderate	marked	severe
Off-topic	not present	mild	moderate	marked	severe
Expressive aphasia	not present	mild	moderate	marked	severe
Apraxia of speech	not present	mild	moderate	marked	severe
Dysarthria	not present	mild	moderate	marked	severe
Overall communication impairment	not present	mild	moderate	marked	severe