Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 116–134

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Quantification of speech disfluency as a marker of medication-induced cognitive impairment: An application of computerized speech analysis in neuropharmacology☆

Serguei V.S. Pakhomov *, Susan E. Marino, Angela K. Birnbaum

Center for Clinical and Cognitive Neuropharmacology, 7-125F Weaver Densford Hall, 308 Harvard St. SE, Minneapolis, MN 55406, USA

Received 7 May 2011; received in revised form 4 April 2012; accepted 27 April 2012
Available online 8 May 2012

Abstract

We present the results of a study investigating the use of speech and language characteristics extracted from spontaneous spoken discourse to assess changes in cognitive function. Specifically, we investigated the use of automatic speech recognition technology to characterize spontaneous speech disfluency induced by topiramate, an anti-epileptic medication with language-related side-effects. We audio recorded spontaneous speech samples from 20 participants during several picture description tasks and analyzed the recordings automatically and manually to extract a range of spoken fluency measurements including speech discontinuities (e.g., filled pauses, false starts, and repetitions), silent pause duration, speaking rate and vowel lengthening. Our results indicate that some of these paralinguistic speech characteristics are (a) sensitive to the effects of topiramate, (b) are associated with topiramate concentrations in the blood, and (c) complement standard neuropsychological tests typically used to investigate cognitive effects of medications. This work demonstrates the use of computational linguistic tools to assess cognitive effects in a more sensitive, objective, and reproducible manner than is currently available with standard tests.

Keywords: Speech recognition; Cognitive assessment; Speech disfluency; Topiramate

1. Introduction

The field of computational linguistics has generated a wealth of automated and semi-automated approaches and systems designed to process natural language and speech. However, most of the advances in the development and validation of language technology have been focused on language produced by neurotypical adults and children under various conditions and using a variety of modalities – newspaper text, dictation, dialogue, as well as other common forms of human–human and human–computer communication. As a result, the majority of computerized language technology today is based on the reasonable assumption that the speech and language that it targets contains regularities and patterns that may be captured via explicit rules or language and acoustic models based on distributional characteristics of linguistic entities in text and speech corpora. While this assumption may be true for the majority of the typical end-users of human language technology, it is certainly not true, at least not entirely, for speech and language

☆ This paper has been recommended for acceptance by 'Björn Schuller, PhD'.
* Corresponding author. Tel.: +1 612 624 1198; fax: +1 612 625 9931.
E-mail addresses: pakh0002@umn.edu (S.V.S. Pakhomov), marin007@umn.edu (S.E. Marino), birnb002@umn.edu (A.K. Birnbaum).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.04.002

produced by people with cognitive impairment arising from traumatic brain injury, neurodegenerative disorders, or even something as common as stress, fatigue, and use of alcohol and drugs (both illicit and prescribed). In a recent publication, Krahmer (2010) outlined several areas where the fields of computational linguistics and psychology may intersect and learn from each other. Krahmer maintains that words and sequences of words can provide valuable information about the person who produces them (p. 7). This is particularly true of words (and we might add sounds and silent pauses) produced by cognitively impaired individuals.

The field of neuropsychology has traditionally relied on standard neuropsychological test batteries consisting of tests designed to assess specific functions. For example, one of the standard tests of verbal fluency involves asking the subjects to name in 60 s as many words as they can that start with a certain letter of the alphabet (phonemic fluency) or that belong to a certain semantic category such as "animals" (semantic fluency). These tests of verbal fluency are able to distinguish between people with damage to the frontal networks (e.g., Left Inferior Frontal Gyrus (LIFG) – Broca's area) resulting in decreased performance on phonemic fluency but not semantic fluency, from people with damage to the temporal networks (e.g. Anterior Temporal Lobe (ATL)) resulting in the opposite effect – decreased semantic fluency with preserved phonemic fluency. However, these neuropsychological tests of verbal fluency are not designed to assess what we commonly think of as linguistic fluency, which is a multifaceted speech characteristic consisting not only of how many words or syllables per unit time one can produce but also various prosodic features such as intonation, rhythm, and presence of disfluent events (e.g. um's and ah's, repetitions, and repairs) (Shriberg, 1994). This notion of "natural" fluency has become the focal point of attention in research on progressive aphasia – a spectrum of conditions caused by gradual deterioration of language-related cognitive networks in the brain and evident in impaired language use. Distinguishing between fluent and non-fluent subtypes of progressive aphasia is critical to developing accurate information processing models that underlie this condition and can inform its treatment (Hillis, 2007). Furthermore, it has been recognized by researchers in clinical neuropsychology that a truly comprehensive description of progressive aphasia necessitates the inclusion of prosody into language assessment instruments (Rohrer et al., 2008).

In the field of clinical neuropharmacology, one of the central questions is to determine how medications designed to treat neurological conditions interact with the brain to explain variability in response across individual patients. For example, a regularly prescribed dose of an anti-epileptic medication may be toxic to some people and have therapeutic or subtherapeutic effect in others. In treating epilepsy, neurologists typically start patients with very low doses and slowly titrate them to therapeutically effective levels. Thus, appropriate medication dosing is currently determined by a "trial-and-error" approach. An improved dosing approach would be able to identify those individuals who would not benefit or who may be hurt by a medication, so that these medications could be avoided. In addition, the capability to determine the relationships between drug concentration and response along with factors that can influence these relationships (i.e., race, age, and genetics) will enable the choice of an optimal initial and target dose in patients and thus avoid the unnecessary exposure to toxicity or seizures (in the case of epilepsy). This ability to predict drug response prospectively, and at the individual patient level, would enable a more rational approach to determining the most effective medication and its dosage. Improvements in objective and reproducible behavioral measurements of drug response are needed in order to adequately describe the relationship between concentration and response in the patient. This provides a unique niche to which computational linguistics and automated language and speech processing technology can contribute.

In the current study, we address the problem of using speech fluency analysis to measure cognitive effects of an anti-epileptic medication – topiramate (brand name – Topamax®). This medication is approved by the Food and Drug Administration in the U.S. for the treatment of epilepsy and migraine prophylaxis but is also widely utilized for the treatment of chronic pain, obesity, and alcohol addiction. Despite its effectiveness in treating epilepsy and migraine, this medication has a number of cognitive side-effects including deficits in memory and attention and, most notably, word-finding problems. Patients treated with topiramate report numerous events in which they know the word that they want to say but just cannot "get it out" (Meador et al., 2003; Mula et al., 2003). However, these cognitive side-effects are reported only by a subset of the patients taking this medication (Mula et al., 2003). It remains unclear what individual physiological and/or neuropsychological differences may account for these differential effects. Furthermore, the exact mechanism of action of this medication in the brain is currently not well understood. Being able to quantify behavioral linguistic characteristics such as speech fluency may help to predict how different individuals will respond to this drug and thus determine the most optimal dose in a more rational manner than currently exists.

In the current article, we extend Krahmer's (2010) argument to the domain of neuropsychology and demonstrate the application of several natural speech and language processing approaches to fully automated, objective, and

reproducible quantification of the effects of anti-epileptic medications on cognitive function as manifest in fluency of spontaneous speech. We also demonstrate how our cross-disciplinary approach benefits the field of computational linguistics by providing experimental data for validation of automated speech and language processing technology.

2. Background

Hesitations are part of normal spontaneous speech production and constitute 5–10% of speech output in naturally occurring conversations (Shriberg, 2001, 1994; Oviatt, 1995; Maclay and Osgood, 1959). Hesitations consist of a variety of speech events including but not limited to unfilled pauses (silences), filled pauses ("um", "ah", etc.), repetitions, vowel lengthening (Shriberg, 1994; Clark, 1996), and have a number of diverse functions. One of these functions is to allow the speaker time to plan the content and form of an upcoming utterance (Butterworth, 1980; Levelt, 1989). Thus, predictably, more hesitations are found near utterances that are less routine and are more cognitively demanding (Freud and Strachey, 1938; Goldman-Eisler, 1958; O'Connell et al., 1969). Hesitations may also be used as a conversational strategy in order to retain the listener's attention, signaling to the listener that the speaker is not yet ready to give up the floor (Maclay and Osgood, 1959; Shriberg, 1994; Erbaugh, 1987). Hesitations have also been found to be related to the contextual probability and frequency of the words following the hesitation (Lounsbury, 1954; Goldman-Eisler, 1958, 1967; Beattie, 1983; Bell et al., 1999), as well as boundaries of various units of syntactic and discourse structure (Downing, 1970; Boomer, 1965; Shriberg, 1994; Hawkins, 1971; Heeman and Allen, 1997; Cook, 1977; Grosjean et al., 1979). The functions of filled and unfilled pauses often overlap, as do the cognitive explanations for these various forms of hesitations (Maclay and Osgood, 1959; Downing, 1970; Beattie, 1980; Shriberg, 1994; Kircher et al., 2004). When both filled and unfilled pauses are considered together as a single phenomenon, they tend to appear at regular intervals, alternating with more fluent speech forming periodic temporal cycles that have been postulated to reflect cognitive cycles (Roberts and Kirsner, 2000; Merlo and Barbosa, 2010). Word fragments are not considered to be hesitations but rather are typically found as part of speech repairs – a disfluent speech event resulting from the speaker's monitoring of his/her own speech and interrupting the speech when an error is detected with subsequent correction of the problem. The interruption in speech repairs may occur in the middle of a word resulting in a word fragment in approximately 22% of all repairs (Levelt, 1989) (e.g., "I see a mo- um a mother doing dishes"). For the purposes of our study, both hesitations and disfluencies produced as part of speech errors are of interest as they may signal a disruption in cognitive processing affected by medications. In the study, we limited the labeling of repairs to only those with word fragments, grouped these distinct classes of speech events together, and refer to them as speech discontinuities in the remainder of this article.

Methods that are typically used to study hesitations and speech errors in spontaneous speech production rely on automatic speech recognition technology used in a semi-automated mode. In this mode, the speech samples are transcribed verbatim at the word level by hand and the transcripts are subsequently aligned using supervised phone-based recognition (Wheatley et al., 1992; Wightman and Talkin, 1996; Bratt and Shriberg, 2008; Huang et al., 2006) resulting in highly accurate timing information. The alignment accuracy on the SWITCHBOARD corpus, for example, was estimated to be 91% at the word level (correctly matching to within 40 ms word onsets and offsets) (Wheatley et al., 1992). The SWITCHBOARD corpus represents one of the more acoustically challenging sources of audio data to process because it contains spontaneous telephone conversations that were recorded at a 8 kHz sampling rate (limited by the telephone channel bandwidth) and contain numerous speaker overlaps, signal corruptions and non-speech noise. An example of accurate automatic alignments obtained on spontaneous dialogues recorded with close-talking microphone quality at a 16 kHz sampling rate is the TRAINS corpus (Heeman and Allen, 1993). The accuracy of automatic word-level alignment is reported to be 83% to within 50 ms boundary identification precision. The agreement between manual alignments and those that were subsequently corrected using the spectrogram for improved precision was reported to be 95%, thus establishing the ceiling for manually achievable accuracy. The developers of the TRAINS corpus also made the assumption that phone-level alignments were likely to be as accurate as word-level alignments (Heeman and Allen, 1993, p. 9). Similarly good results on alignment (81% accuracy) were achieved on the ATIS corpus that also represents spontaneous speech sampled at 16 kHz using a high quality desktop microphone. In the domain of read speech, highly accurate word and phone-level alignment have also been reported on the TIMIT corpus recorded in a controlled environment at a 16 kHz sampling rate. This corpus represents an "easier" target than those containing spontaneous dialogues and telephony speech; however, the accuracy of 93% achieved on phonetic alignment to within

20 milliseconds precision (Hosom, 2002) is very close to the ceiling performance of 95% demonstrated on the TRAINS corpus.

Phonetic and word-level alignments resulting from these semi-automated approaches have been used extensively to describe the distributional, acoustic and prosodic properties of various types of hesitations Shriberg (1994, 1996, 1999), Bell et al. (2003), Heeman et al. (2006). For example, Shriberg (1999) has shown experimentally on the ATIS corpus that vowels tend to be significantly longer in duration and may change in quality (e.g., the article "the" pronounced as [ðˑ iˑ]) in the context of a repetition as well as other types of hesitations. Hesitation-related vowel lengthening typically occurs in the last syllable of the word preceding the hesitation but may also occur in other vowels as well as continuant consonants such as [m], [n], [ŋ], [l], [j], [w], [r]. Furthermore, filled pauses (in English) consist of at least one vowel (e.g., [a, e, u]) that may sometimes be followed by nasalization typically realized as a nasal consonant (e.g., [m, n, ŋ]), and tend to be longer in duration than equivalent vowels and consonants pronounced as parts of words (Shriberg, 1999; Bell et al., 1999). This is an important observation in the context of automated quantification of speech fluency characteristics. The amount of vowel and consonant lengthening present in a given speech sample may be indicative not only of the number of filled pauses it contains but also of how hesitant the speech is overall. However, one of the disadvantages of the semi-automated approaches to prosodic feature analysis is the necessity to have a verbatim or near-verbatim transcript aligned with the audio signal. Verbatim transcription is difficult and time consuming (Pakhomov et al., 2001) and therefore may not be feasible in settings that require fast response such as clinical neuropsychological testing.

One of the current challenges facing automated identification of disfluent events is the relatively low accuracy of automatic speech recognition (ASR) systems on continuous spontaneous speech. While modern ASR systems are able to achieve accuracies close to 90% on read speech exemplified by the English Broadcast News benchmark tests, spontaneous conversational speech continues to present a considerable challenge, with benchmark accuracy results in the 50–80% range depending on the task and environment (NIST RT-09 Evaluation). Low ASR accuracy presents a problem for disfluency detection because approaches that work best on this task rely on lexical and syntactic information obtained from the raw ASR output. Also, ASR errors tend to occur around disfluent events of interest, thus making the task of improving ASR spontaneous speech-to-text transcription even more challenging.

While relatively high ASR word-error rates may preclude the use of the technology for reliable speech to text conversion, the ASR output at the subword level may still be useful in developing a measure of fluency for neuropsychological testing. For example, phone-level ASR output may be able to identify silent pause and phone boundaries with enough consistency to be indicative of durational and segmental properties of speech sounds present in a sample of speech. Thus, from the standpoint of using ASR for accurate speech to text conversion, the critical question is how good the ASR system is at correctly identifying the spoken words. From the standpoint of using ASR to measure fluency for neuropsychological testing, however, the critical question is whether the measurements derived from the ASR output have any validity in relation to the cognitive processes impaired by disease, medications or other factors. This type of validity in psychological research is sometimes referred to as *construct* validity (Cronbach and Meehl, 1955). To validate a new construct, it is necessary to show that the measurement being validated for a specific use correlates with other independent types of measurements known to capture the phenomenon being measured (criterion validity). It is also necessary to show that the new measurement is able to distinguish between groups of interest such as healthy vs. impaired subjects (concurrent validity) or can be used to predict future states (predictive validity) or events (e.g., using disfluency score to predict the likelihood of developing a neurologic disorder). In the current study, we validate a novel application of ASR for the purpose of measuring the fluency/disfluency characteristics of spontaneous speech.

3. Methods

3.1. System design

Phone-level transcription of spontaneous speech samples was performed using the Hidden Markov Model Toolkit (HTK 3.4) (Young et al., 2006). We used a standard ASR system architecture consisting of a front-end signal processing and encoding encoding module and a decoder module. The front-end signal processing was performed with HWave and HParm modules of HTK in order to convert the digitized waveform audio files to a series of Mel-frequency Cepstral Coefficients (MFCCs) parameter vectors representing discrete observations derived from the FFT-based log spectra. A

detailed description of the filterbank used by HTK to derive MFCCs can be found in the HTK Book (Young et al., 2006). The HTK decoder module (HVite) constructs a lattice of hypotheses from a set of Hidden Markov Models (HMMs) representing an acoustic model, a language model, and a dictionary. Subsequently, the token passing variation of the Viterbi search algorithm (Young et al., 1989) is used to traverse the lattice in order to find the most likely path. We trained a set of tri-phone HMMs (an acoustic model) specifically for this task and constructed a simple phone-loop language model. Both acoustic and language modeling are described in more detail in the following sections.

3.1.1. Phone inventory

We used a standard set of phones for acoustic and language modeling consisting of the following set of 39 phones including pure vowels, diphthongs and consonants (shown in TIMIT format followed by the corresponding IPA symbol):

aa [ɑ], ae [æ], ah [ʌ], ao [ɔ], aw [aw], ax [ə], ay [aj], b [b], ch [ç], d [d], dh [ð], eh [e], ey [ej], f [f], g [f], hh [h], ih [ɪ], iy [i], jh [ʄ], k [k], l [l], m [m], n [n], ng [ŋ], ow [ow], oy [oj], p [p], r [r], s [r], sh [ʃ], t [t], th [θ], uh [ʊ], uw [u], v [v], w [w], y [j], z [z], zh [ʒ]

and two special phones to represent silent pauses: (sp – short pause and sil – long pause) for a total of 41 items.

3.1.2. Acoustic model

We trained a speaker independent acoustic model on speech collected for another unrelated study that used the same microphones and similar speech tasks to the current study. A total of 73 samples from 45 older and younger subjects (25 men, 20 women, mean age 58.1 years old) were collected from speakers instructed to describe a series of pictures some of which were the same as those in the current study and some that were similar. The speech used in acoustic modeling was recorded at 16 kHz sampling rate and 16 bit resolution with Andrea Electronics microphone array with an external USB sound card. The total duration of speech used in acoustic modeling was approximately 1.3 h. This model is speaker independent only in the sense that it was not constructed or adapted for a specific speaker. The relatively small amount of training data for a speaker independent acoustic model is addressed in the Limitations section of this article. MFCC-type acoustic models for each of the three corpora were trained, each consisting of a set of five-state tied-state continuous density tri-phone HMMs. We used 19 filterbank channels and 10 cepstral features with cepstral mean normalization turned on. The audio was sampled at 10 ms intervals with a Hamming window of 25 ms. The same signal processing parameters were used during decoding.

3.1.3. Language model

A phone loop finite-state network was constructed consisting of the 39 phones in the phone inventory. The network contained mandatory utterance initial and final nodes and an optional silent pause (sil) following each of the phones as defined by the BNF-style grammar in Fig. 1. This grammar allows any sequence of phones mixed with silences in any order and thus results in an ASR system that relies only on the acoustic model to determine the most likely phone sequence corresponding to the input speech signal. The dictionary for this language model was constructed by using each phone symbol in the phone inventory as its own pronunciation.

In addition to using the phone-loop language model, fixed-order word-level language networks were constructed for each individual picture description sample from the verbatim transcripts. Thus, these networks represent *sample-dependent* (in contrast to domain or speaker-dependent) language models that were used in forced alignment. Word sequences were encoded via Backus-Naur Form (BNF) grammars in which each word in the verbatim transcript was entered in a fixed sequence with optional silences between words. The dictionary for word-level models was constructed by selecting pronunciations from the Carnegie Mellon University (CMU) dictionary (cmudict.0.6). Pronunciations for

```
$phoneme = sil | zh | z | y | w | v | uw | uh | th | t | sh | s | r | p | oy | ow | ng | n
         | m | l | k | jh | iy | ih | hh | g | f | ey | eh | dh | d | ch | b | ay | ax
         | aw | ao | ah | ae | aa ;

(sil < $phoneme > sil)
```

Fig. 1. Phone-loop language network in BNF format.

Fig. 2. Examples of picture description stimuli: Boston Diagnostic Aphasia Examination (left) and Minnesota Test for Differential Diagnosis of Aphasia (right).

unknown words were automatically approximated with the MBRDICO ID3-based letter-to-phoneme conversion toolkit trained on the CMU dictionary.

Thus, the first system (based on phone-loop network) constitutes completely free phoneme recognition without forced-alignments, whereas the second system constitutes phoneme recognition based on forced alignment with word-level verbatim transcripts Fig. 2.

3.2. Study design

3.2.1. Participants

Twenty-five ($n = 25$) native English-speaking, healthy volunteers (8 women, 17 men) between 18 and 50 years of age were recruited for this study. During initial screening, persons who reported clinically significant cardiovascular, endocrine, hematopoietic, hepatic, neurologic, psychiatric, or renal disease; current or a history of drug or alcohol abuse; the use of concomitant medications known to affect topiramate, lorazepam, or that alter cognitive function; prior adverse reaction or prior hypersensitivity to topiramate, lorazepam or related compounds; demonstrated a positive pregnancy test (administered to all women before enrollment and before each experimental session); received any investigational drug within the previous thirty days, were excluded from participation in the study.

3.2.2. Experiment design

This paper presents a method developed for data analysis of a previous study. The study design involved recruitment at two different study sites – University of Minnesota (UMN) and University of Florida (UF). Each site's Institutional Review Board (IRB) Human Subjects' Committee approved this protocol. Using a randomized, double-blind, crossover design, subjects at UMN received 100 mg oral topiramate (TPM), 2 mg oral lorazepam (LZP) (brand name – Ativan®), and an inactive placebo (PBO), while those at UF received only TPM and PBO. LZP is a sedative that was used as an active control in order to distinguish between language-specific effects of TPM and generalized sedative effects. The study design at both sites is illustrated in Fig. 3.

Each UMN subject participated in five sessions, whereas each UF subject participated in four sessions. Each potential subject participated in a telephone-screening interview in which the risks associated with the study were discussed and an IRB-approved consent form was subsequently mailed to those subjects that met the inclusion criteria. On the first visit (WK1 pre-baseline) the study staff went over the terms of the consent form with the subject one more time and the subject supplied a brief demographic, medical and medication history. Each participant was then given a neuropsychological test battery lasting approximately 1 hour after which they were randomly assigned to a study treatment sequence (UMN: TPM, LZP, PBO; UF: TPM, PBO). During the second visit (WK2), subjects were administered the drug (or placebo) and after 1 or 1.5 h (UMN and UF, respectively) they were given the same neuropsychological battery as in the first session, but with alternate versions of each test. After completing the battery, vital signs were recorded and a blood sample was drawn to determine drug concentrations in the blood. Participants

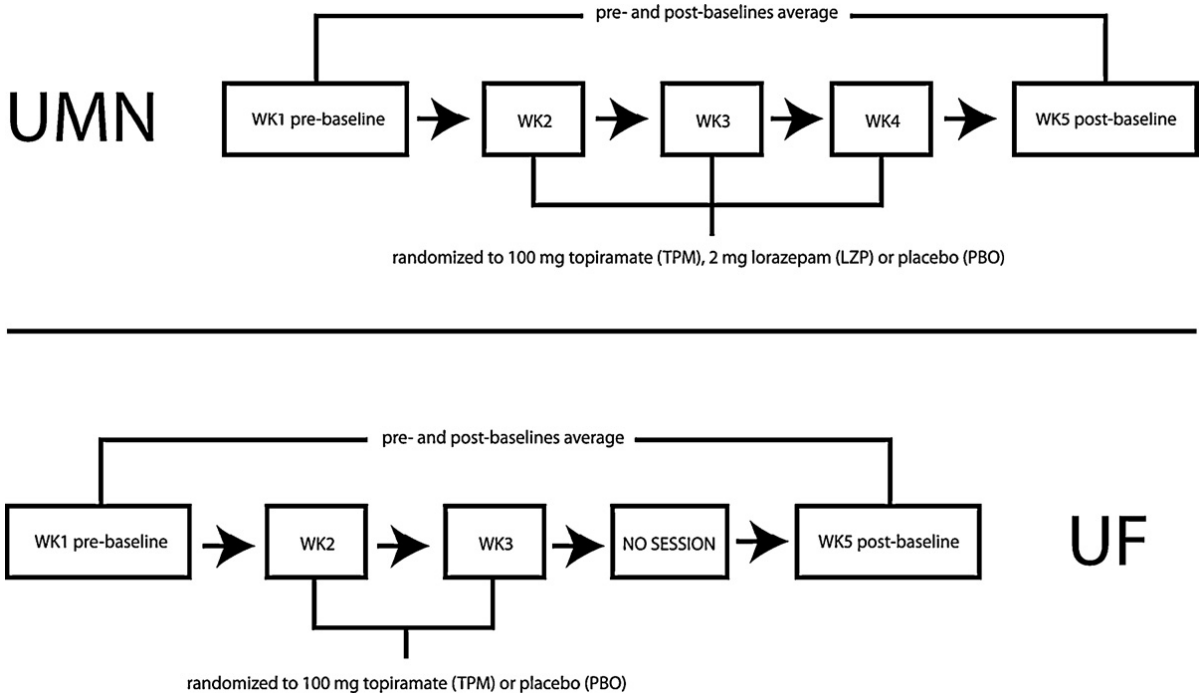

Fig. 3. Study design illustration.

returned one week later (WK3) in order to repeat the testing protocol after switching drugs according to the fixed randomization scheme. For subjects at UMN, the third treatment session (WK4) occurred seven days after the second drug administration. Since there was one less treatment at UF than at UMN, subjects at UF did not have session number 4 (NO SESSION). One week after WK4 session for UMN and two weeks after WK3 session for UF, subjects returned for a second baseline (WK5 post-baseline) testing session. The two baseline scores were averaged in order to correct for any practice effects that might have occurred across the entire study period, which was kept constant across sites despite the difference in number of treatments. The average baseline score was used to compute the percent change score of each test administered under drug conditions as the ratio of the difference between the score obtained during the experimental session from the average baseline score divided by the average baseline score.

3.2.3. General corpus characteristics

Due to a variety of reasons including adverse drug effects and the presence of drug from the previous treatment arm, five subjects either had missing data or were excluded from analysis resulting in a set of 20 remaining subjects. The experiments resulted in a corpus of 89 audio recordings collected from the picture description task lasting a total of 123.6 min. Some of the recordings ($n = 11$) were lost due to operator errors and equipment malfunction. The mean length of a picture description recording was 83.3 s (SD = 52.5). The shortest recording was 20.2 s and the longest – 291.5 s. Descriptive statistics summarizing key corpus characteristics are shown in Table 1. Table 2 shows the distribution of speech and silent pause duration and raw counts of speech discontinuities across different experimental conditions – Baseline, TPM, LZP and PBO.

A subset of 29 out of 89 recordings was used to evaluate the reliability of transcription. This subset contained a total of 4633 words (mean = 159.8, SD = 87.4) and was transcribed once by each transcriptionist and subsequently used to calculate inter-rater agreement.

3.2.4. Cognitive measures

The testing battery administered during each of the study sessions included the following measures of cognition directly relevant to the assessment of fluency:

Table 1
Descriptive corpus statistics. The "Total", "Min", and "Max" columns contain values computed over all available picture descriptions. The mean and SD values represent the average and standard deviation of the average computed across the picture descriptions.

Corpus characteristic	Total	Min	Max	Mean (SD)
Words (W)	13,027	37	514	146.4 (98.8)
Speech duration (SpD) (s)	3954	9.9	151.5	44.4 (29.5)
Silence duration (SilD) (s)	3462	10.4	146.2	38.9 (25.7)
Silent pauses (SIL)	4361	8	143	40.1 (25.7)
Filled pauses (FP)	581	0	37	6.5 (6.2)
Repetitions (REP)	90	0	9	1.0 (1.7)
Word fragments (WF)	78	0	5	0.9 (1.2)
Speech discontinuities (SD = FP + REP + WF)	749	0	38	8.4 (7.5)

Word-level verbal tests: Phonemic (Controlled Oral Word Associations (COWA)) and Semantic Fluency (animals/clothes) Tests (Loonstra et al., 2001; Benton et al., 1994)
Discourse-level verbal tests: Picture description task (Goodglass and Kaplan, 1983)

During the test of phonemic fluency, the subjects were asked to say out loud as many words as they could that started with a specific letter of the English alphabet in 60 seconds while avoiding the use of proper names. A trained psychometrist administering the test counted the number of valid responses from the subjects as the phonemic fluency score. In each session the subjects underwent three trials consisting of the following alternating sets of three letters (A, F, S) or (B, H, R). The phonemic fluency score was computed by averaging the scores across the three letters in a set. During the picture description test, the subjects were presented with one of four standard pictures and were asked to describe them verbally. The pictures were selected from standard test batteries designed to assess aphasia: (1) Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983); (2) Minnesota Test for Differential Diagnosis of Aphasia (Schuell, 1965); (3) Nicholas–Brookshire "Rescue" picture description task (Nicholas and Brookshire, 1995); (4) Nicholas–Brookshire "Birthday" picture description task (Nicholas and Brookshire, 1995). During both pre- and post- baseline assessments, the subjects were asked to describe the Boston Diagnostic Aphasia Examination picture. The picture description tests did not have a time limit and were administered using the following standardized instructions script: "I am going to show you a picture. Please describe everything you see going on in the picture. Try to speak in sentences". The recording was stopped if the subject indicated completion or was silent for longer than 30 s. Both the phonemic fluency and the picture description tests were administered to all participating subjects at UMN and UF sites.

All tests were administered by a trained examiner at each site. All language based tests including the picture descriptions and phonemic fluency tests were audio recorded using an Andrea Electronic Array Microphone at 44.1 kHz downsampled to 16 kHz (16 bit PCM) and subsequently transcribed verbatim including hesitations.

Table 2
Distribution of means and total counts for general corpus characteristics presented in Table 1 across experimental conditions (Baseline, TPM, LZP, and PBO).

	Avg. baseline[a]		TPM		LZP		PBO	
	Total	Mean	Total	Mean	Total	Mean	Total	Mean
W	2631	131.6	2949	147.4	2251	204.6	3082	154.1
SpD	831	41.5	901	45.1	658	59.8	887	44.4
SilD	734	36.7	839	42.0	541	49.1	729	36.5
SIL	765	38.3	854	42.7	542	49.3	807	40.4
FP	128	6.40	150	7.50	74	6.72	122	6.10
REP	14	0.70	30	1.50	13	1.82	19	0.95
WF	16	0.78	27	1.35	8	0.73	15	0.75
SD	158	7.88	207	10.35	95	8.64	156	7.80

[a] Average baseline is the mean of speech event counts in pre- and post-baseline recordings. Therefore, the totals in this table do not add up to the totals in Table 1 that represents a sum of all counts and durations across all experimental sessions.

3.2.5. Speech measures

We calculated a total of five different manual and automatic measurements related to spontaneous speech characteristics elicited with the picture description task.

Manual measures: One of the measures was based on manual transcriptions of the spoken narratives subsequently aligned with the audio signal. Two transcriptionists were instructed to identify and transcribe all speech and non-speech events including words, filled pauses, repetitions, repairs, lipsmacks, breaths, coughing, and laughter. Other unclassifiable speech and non-speech events were transcribed as noise. Examples of the verbatim transcripts are provided below.

> "he is flying a different kite and the do- a dog is watching him as he's flying it FP_um and assume this is his family"
>
> "T_NOISE we have a scenario where my T_LIPSMACK gosh where do i start well there's a tree a tree T_COUGH in the center"

In this example, non-speech events and speech noise are marked with a prefix "T_", filled pauses are marked with the "FP_" prefix, and word fragments in the reparandum of repairs are marked with "-". Repetitions are counted as the number of repeated one-, two-, and three-word sequences. Thus, the two examples above contain 43 words, one filled pause, one repair, and one repetition.

Inter-rater agreement between the two transcriptionists was calculated on the subset of 29 randomly selected picture descriptions. We used the NIST 'sclite' tool with "dynamic programming" and "time-mediated alignment" options turned on. In addition to inter-rater agreement on transcription, we assessed the differences in word boundary assignments performed fully manually and with forced alignment. The manual word boundary determination used the Praat system (Boersma and Weenink, 2009) for acoustic analysis and relied on using the spectrogram in addition to voice and the waveform cues. Automatic forced alignments were performed with HTK automatic speech recognition toolkit. The comparisons between manually and automatically determined boundaries were performed on the portions of the transcripts where the two transcriptionists agreed with each other.

Based on the manual verbatim transcriptions, we defined a manual measure of speech discontinuity as the ratio of hesitations to the total number of words in the transcript (SDR). Another manual measure of mean vowel length (MVL) was calculated based on phone durations extracted by hand from vowels in a random sample of words produced by each participant in the study in each of the experimental conditions. Vowels occurring as part of filled pauses were excluded in computing MVL. The words for each recording were obtained by randomly sampling the beginning, middle and the end of the recording resulting in a data set containing 1766 vowel instances representing a vocabulary of 313 unique words. To obtain these phone durations, a linguistics graduate student with experience in acoustic phonetics manually corrected phone boundaries produced via automatic alignments shown in Fig. 4 and calculated the durations of vowels in each of the sampled words. All phone durations were normalized to word length and log-transformed after normalization.

Automatic measures: In addition to the SDR and MVL measures based on manual annotations, we defined three measures derived from the output of the ASR system as shown in Fig. 4.

One of these automatic measures comprised the mean of log-transformed silent pause duration (MPD). Log-transformation was used based on the results of a recent study by Hird and Kirsner (2010) showing that pause durations computed in real time on spontaneous speech samples have a long-tailed distribution that may mask an underlying bimodal distribution evident when log time is used to measure duration. A silent pause was defined as a silent segment in the raw ASR output longer than 150 ms in duration. This cutoff was chosen conservatively to avoid counting phonetically conditioned pauses such as the release phase in the phonation of a word-final stop consonant that may last up to 100 ms in duration (Forrest et al., 1989; Levelt, 1989). We also defined an automatic measure of phone lengthening (APL) that was based on the mean duration of pure vowels and nasal continuant consonants that were extracted from the raw ASR output and log-transformed similarly to silent pause durations. Finally, we defined a measure of speaking rate (SPR) computed as the ratio of words in the verbatim transcript to the total duration of non-silent segments in each recording. Non-silent segments were defined as those representing intelligible speech or disfluencies – speech and non-speech noise were treated as silent segments. As with cognitive measures, we used the relative change from averaged baselines to make group comparisons between LZP, TPM and PBO conditions.

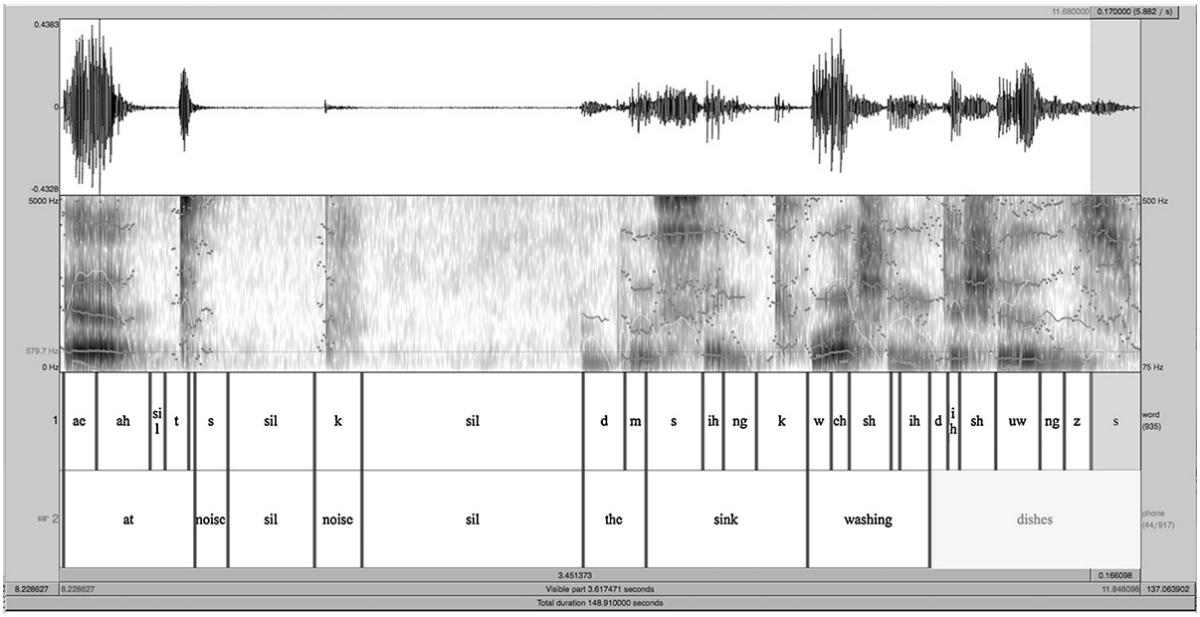

Fig. 4. An example of the phone-loop recognizer's output (tier 1) superimposed over the manual transcription forced aligned with the same audio (tier 2).

3.2.6. Physiological measures

All subjects in this study were given the same dose of both TPM (100 mg) and LZP (2 mg); however, there are still large individual differences in the amount of medication that is present the bloodstream and delivered to the target receptors in the brain. Therefore, in order to account for the variability in drug response, it is necessary to determine the amount of the drug present in the bloodstream after it has been metabolized by the body. A single blood sample was collected for analysis immediately following each drug testing session, which occurred approximately 2 hours after drug administration at UMN and 3 h at UF. Samples were immediately centrifuged and the plasma frozen at −80 °C until analysis. Plasma concentrations of TPM were measured using a simultaneous assay, developed by Subramanian et al. (2008) for the measurement of nine anti-epileptic drugs. Another fast and sensitive analytical method was developed for the analysis of subject plasma samples containing LZP. The assay was based on a method previously published by Zhu and Luo (2005). Analyses were performed on a coupled high performance liquid chromatograph–electrospray tandem mass spectrometer (Waters Micromass Quatro Ultima). The assay's limit of quantification (LOQ) was 1 ng/mL, and the run time for each sample was 6 min. The assay was validated according to FDA guidelines (FDA, 2001). Samples were tested for both TPM and LZP to determine the concentration of the treatment drug and to ensure that the drug from the previous treatment was not present.

3.3. Statistical analysis

Spearman's rank correlation was used to test for correlations between speech, neuropsychological, and physiological measurements. Standard two-tailed *t*-test based on the binomial distribution was used to test for differences between group means. Paired *t*-test was used to compare variable means on different experimental conditions (TPM vs. PBO) because subjects' visits were randomized to these conditions and, thus, each subject was on both PBO and TPM but at different time points. Simple linear regression modeling was used to test for associations between continuous variables. All statistical calculations were performed using the R statistical package (version 2.9).

4. Results

We first present the results of calculating inter-rater agreement and the accuracy of automatic word boundary identification, followed by the results of comparisons of automatically derived speech variables to manually derived

Table 3
Results of comparison of speech characteristics and drug effects. Comparisons with TPM were based on 20 subjects. Comparisons with LZP were based on 11 subjects.

Speech variable	Group means				Correlations	
	Baseline	TPM	LZP	PBO	TPM	LZP
Manual						
SDR	0.068	0.085	0.062	0.048	0.47 (0.03)	0.34 (0.31)
MVL (ms)	107	114	94	97	0.26 (0.27)	0.76 (0.01)
Automatic						
MPD (ms)	89.9	91.8	91.1	86.22	0.21 (0.37)	0.25 (0.45)
SPR (w/s)	3.25	3.31	3.42	3.55	0.11 (0.64)	0.06 (0.85)
AVL (ms)	21.7	22.9	22.1	22.8	−0.08 (0.75)	0.43 (0.19)
AVL-C2 (ms)	15.03	15.6	14.8	14.8	0.02 (0.94)	−0.30 (0.37)
Cognitive						
Ph-fluency	15.42	11.51	15.14	15.00	−0.34 (0.14)	0.54 (0.09)

variables, as well as the results of testing both manually and automatically derived variables for association with drug effects. This section concludes with the presentation of results of traditional cognitive testing known to be sensitive to effects of TPM on cognition. The group means for various arms of the study and correlations with individual blood concentrations of TPM and LZP are summarized in Table 3.

4.1. Inter-annotator agreement and word boundary identification

The transcriptionists agreed 85.7% of the time. The comparison of word beginning boundaries (WBB) and word end boundaries (WEB) identified with forced alignment and manually showed that the mean WBB difference was 55 ms (SD = 172) and WEB was 60 ms (SD = 188). Approximately 86% of the word-onset boundaries and 80% of the word-offset boundaries were identified with better than 50 milliseconds precision, which is comparable to the accuracy of alignments in the TRAINS corpus (Heeman and Allen, 1993).

4.2. Automatically assessed disfluency

The automatically obtained mean log-transformed pause duration (MPD) measurements significantly and negatively correlated with the manually assessed disfluency measurements – SDR (rho = −0.24, $p = 0.023$). The mean difference in log-transformed pause durations from baseline measurements on TPM was positive (relativechange = 0.005) but was not significantly different from the negative mean relative change of −0.008 from baseline on PBO ($p = 0.206$). As a quality control step, we compared pause duration measurements obtained in this fully automated mode to pause durations measured semi-automatically based on forced alignments between the audio and the verbatim transcriptions and did not find a statistically significant difference (log-transformed MPD mean of 6.53 vs. 6.42, $p = 0.256$). Furthermore, automatic and semi-automatic mean pause durations were highly correlated (rho = 0.75, $p < 0.001$, $n = 89$). Similarly, counts of silent pauses greater than 100 ms in duration were also highly correlated between the automatic and semi-automatic measurements (rho = 0.96, $p < 0.001$, $n = 89$).

The automatically obtained measurement of phone lengthening (APL), was tested in two stages. In the first stage, we correlated mean durations of all available vowels and consonants with manually assessed SDR. In the second stage, we computed mean durations of a subset of vowels and consonants (phone composite variables) consisting only of those phones that had statistically significant correlations with manually assessed SDR. The results of the first stage are presented in Table 4.

The results in Table 4 show that statistically significant correlations were found between SDR and most of the monophthong vowels [ah, aa, ih, iy, uh, eh], as well as the voiced nasal bilabial continuant consonant [m]. The relationship between the phone composite variable AVL-C2 consisting of [ah, aa, ih, iy, uh, eh, m] and SDR is shown in Fig. 5. In addition to these correlations with SDR, we fitted a linear regression model with the manually assessed

Table 4
Spearman's rank correlations durations of individual phones and classes of phones with manual measurements of speech discontinuity (SDR).

Vowels			Consonants		
Phone	rho	*p*-Value	Phone	rho	*p*-Value
Individual phones					
ah	0.55	<0.001	w	0.13	0.238
uw	0.06	0.599	r	−0.01	0.904
ao	0.14	0.185	jh	0.05	0.664
ax	0.08	0.435	dh	0.07	0.516
aw	0.21	0.050	d	−0.01	0.943
oy	0.20	0.061	y	−0.13	0.236
ih	0.43	<0.001	k	0.02	0.858
ay	0.01	0.897	ch	−0.08	0.432
uh	0.27	0.012	g	0.20	0.060
ey	0.11	0.310	f	−0.06	0.561
iy	0.35	<0.001	th	−0.09	0.379
ae	0.13	0.226	t	−0.05	0.658
aa	0.22	0.043	n	0.07	0.502
eh	0.21	0.043	zh	0.00	0.983
ow	0.10	0.339	m	0.24	0.026
			v	−0.23	0.028
			s	−0.10	0.335
			l	0.14	0.180
			p	0.04	0.734
			b	0.09	0.383
			z	−0.15	0.163
			hh	0.00	0.969
Phone classes					
All Vowels	0.51	<0.001			
All Consonants	0.05	0.654			
All phones	0.40	<0.001			
AVL-C1 [ah, aa, ih, iy, uh, eh]	0.59	<0.001			
AVL-C2 [ah, aa, ih, iy, uh, eh, m]	0.62	<0.001			

disfluency as the outcome and both AVL-C2 and MPD as predictors to determine if the phone lengthening variable (AVL-C2) predicts SDR independently of mean pause duration (MPD). The resulting model is shown below:

$$\mathrm{SDR} = -0.60763 + 0.20541 \times \mathrm{AVL} - C2 + -0.03872 \times \mathrm{MPD}$$

Both AVL-C2 and MPD variables had a statistically significant effect on the SDR (AVL-C2: $p = 0.009$, MPD: $p < 0.001$, $R-\mathrm{squared} = 0.42$, $\mathrm{df} = 86$) showing that vowel lengthening contributed information in addition to that provided by the mean pause duration.

In the second stage, we tested the phone composite variable AVL-C2 to determine if this variable had different means in the TPM and PBO conditions. The comparison of AVL-C2 means in the two conditions showed that the mean AVL-C2 duration on TPM increased from the baseline by 1.3% whereas the duration of PBO decreased from baseline by 0.4%. This difference is small but statistically significant ($p < 0.001$, $\mathrm{df} = 19$) indicating that subjects on TPM had a slight but highly consistent tendency towards phone lengthening as compared to PBO. We also determined that vowel lengthening was distinct from the speaking rate by doing the same comparisons between groups using the SPR variable. The speaking rate increased from the baseline in both conditions (TPM: 6.3%, PBO: 0.7%); however, the difference was not statistically significant ($p = 0.538$). Furthermore, we also tested manually assessed vowel durations (MVL) for correlation with automatically assessed vowel durations (AVL) and found a relatively strong and significant correlation between the two variables ($\mathrm{rho} = 0.64$, $p = 0.003$, $n = 20$).

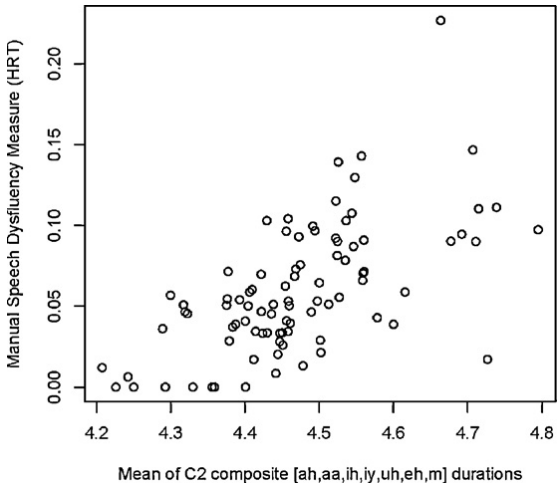

Fig. 5. Illustration of the relationship between automatically extracted phone durations and manually determined SDR.

4.3. Manually assessed disfluency

The ratio of speech discontinuities to words (SDR variable) in the transcripts of the subjects' speech produced on the picture description task correlated with the pharmacokinetic measures of TPM plasma concentrations (rho = 0.47, $p = 0.035$, $n = 20$). The correlation with LZP plasma concentration was not significant (rho = 0.26, $p = 0.44$, $n = 11$). There was a statistically significant difference in the SDR means between the PBO condition showing a 3% decrease in SDR and the TPM condition showing a 44% increase in SDR ($p = 0.002$). With LZP, we observed a 12% decrease in SDR and a decrease of 5% on PBO; however, this difference was not significant ($p = 0.585$). In summary, the subjects were significantly more disfluent on TPM than on PBO and the degree of disfluency on TPM was responsive to the amount of medication present in the bloodstream.

The manually assessed vowel lengthening (MVL) measure did not correlate with TPM concentrations but there was a significant difference between the MVL means in the TPM vs. baseline ($p < 0.001$), and LZP vs. baseline ($p < 0.001$) conditions, but not between the PBO and baseline ($p = 0.201$). Vowel lengthening of 10 and 16 ms on average, as compared to baseline, was observed for both TPM and LZP, respectively. There was a strong correlation between MVL and LZP concentrations (rho = 0.76, $p = 0.006$, $n = 11$).

4.4. Hesitation vs. disfluency

In a post hoc analysis, we combined filled pauses, repetitions and vowel elongations to form a measure of hesitancy separate from a measure of speech errors consisting of word-fragment-containing repairs. We did not find either significant correlations with TPM concentrations or a significant difference between relative change from baseline for TPM and PBO group means ($p = 0.129$) in the rate of speech repairs. The correlation between the measure of hesitation that included filled pauses, repetitions and vowel elongations and TPM levels was 0.41 ($p = 0.07$), which is slightly lower than 0.47 ($p = 0.03$) with the measure of speech discontinuities (filled pauses, repetitions and repairs). The differences between the means of relative change from baseline in the TPM and PBO groups were significant for the hesitations ($t = 2.69$, df = 19, $p = 0.014$), but not for speech repairs ($t = 2.09$, df = 19, $p = 0.055$). However, the difference in the means for speech repairs is on the border of significance defined at 0.05 Type I error probability level. Thus, we think that the rate of speech repairs may be affected by TPM but they may not occur often enough in our data sample to be treated separately from hesitations.

4.5. Cognitive measures

The performance on the standard neuropsychological test of phonemic fluency decreased by 24% on TPM relative to the baseline, as compared to 1.4% on PBO ($p < 0.001$). However, there was no statistically significant correlation

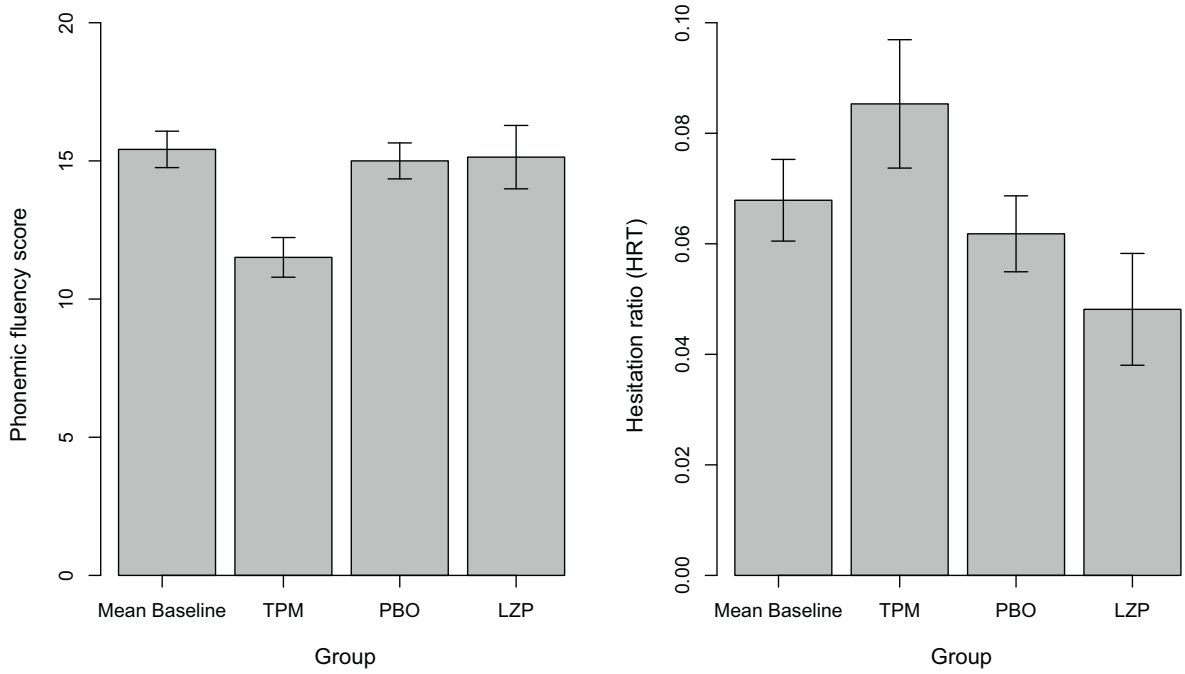

Fig. 6. Means and 95% CIs for phonemic fluency vs. SDR measurements within different study conditions (Groups: mean of two baselines; TPM – topiramate; PBO – placebo; LZP – lorazepam).

between the scores on the phonemic fluency test and the physiological measures of TPM plasma concentrations (rho = −0.38, $p = 0.10$, $n = 20$). The performance on the semantic fluency test did not seem to be affected by TPM, as no statistically significant differences were found between the TPM and the PBO conditions.

A comparison between the means for phonemic fluency and SDR scores across the baseline, TPM, PBO and LZP experimental conditions is illustrated in Fig. 6. The figure shows that phonemic fluency scores are lowest and the SDR scores are highest in the TPM condition. These significant group means differences contrast poor correlations between phonemic fluency and TPM concentrations in the blood and significant correlations between blood concentrations and spoken fluency reflected in the SDR measurement. We further address the significance of this finding in Section 5.

When calculated across all individual subject testing sessions (i.e., baseline, TPM, PBO or LZP), phonemic fluency scores had a statistically significant but weak correlation with SDR scores (rho = −0.24, $p = 0.025$, $n = 89$), while semantic fluency test scores did not show a significant correlation with SDR (rho = −0.17, $p = 0.102$, $n = 89$). However, after separating the scores by experimental condition, only the PBO group had a statistically significant correlation between phonemic fluency and SDR (rho = −0.57, $p = 0.008$, $n = 20$). The phone composite variable AVL-C2 also was not significantly correlated with phonemic fluency test scores (rho = −0.20, $p = 0.064$) and was significantly but weakly correlated with semantic fluency test scores (rho = −0.25, $p = 0.017$). Neither semantic nor phonemic fluency test scores significantly correlated with the log-transformed mean pause durations (MPD).

5. Discussion

In this paper we have demonstrated a novel application of spoken human language technology to the domain of neuropsychological testing and neuropharmacology. Our fully automated approach, based on measuring durational characteristics of silent pauses and phones output by an ASR system, was able to produce measurements that correlated with manually assessed disfluencies consisting of filled pauses, repetitions and false starts. Furthermore, our findings indicate that the notion of spontaneous speech disfluency measurable via automated or semi-automated approaches is associated with objective physiological measurements and are consistent with what is known to date about the cognitive effects of TPM. In particular, the use of TPM has been reported to result in word finding difficulties (Mula et al., 2003) that are likely to manifest themselves via a more effortful and hesitant speech. Thus, these findings

offer support for the *criterion* and *concurrent* components of construct validity (Cronbach and Meehl, 1955) of our approach. Further validation is currently underway and includes the *predictive* component of construct validity in which we will determine if the disfluency response at a lower (100 mg) dose of TPM is predictive of the response at a higher (200 mg) dose.

Automated or even semi-automated tools for assessing spontaneous speech fluency on tasks that are closer to what people do in real life than standard neuropsychological tests provide a novel dimension to the assessment of global cognitive functioning. Standard neuropsychological tests are reportedly anxiety-producing for the subjects, which may alter their cognitive state and produce results that do not reflect functioning under normal conditions. Automated tools based on spontaneous speech task described in this paper may provide a less anxiety-provoking and more ecologically valid approach to measuring and monitoring cognitive effects of medications and neurodegenerative disorders such as dementia. Patients with these disorders frequently present with aphasia but may also have relatively subtle changes in language production such as alterations in hesitation behavior while describing a picture. Recent research on primary progressive aphasia, a language disorder characterized by the atrophy of frontal and temporal brain networks, points out that early diagnosis of the non-fluent variant of primary progressive aphasia remains the subject of much debate and that ". . .developing reliable and objective measures that capture patients early in the disease process is very important . . ." to addressing this challenge (Gorno-Tempini et al., 2011). In our previous work, we have successfully used an approach similar to the one described in this paper that relied on the alignment between audio and verbatim transcripts of picture descriptions by patients with primary progressive aphasia to distinguish between speech, motor, and semantic deficits (Pakhomov et al., 2010). Also, other groups have successfully investigated the use of similar approaches to the diagnosis of patients with mild cognitive impairment (MCI), an early disorder associated with Alzheimer's disease (Roark et al., 2011), and probable Alzheimer's disease (Singh et al., 2010). Durational speech characteristics including silent pause duration, duration of phonation with and without pauses, as well as the ratio of pauses to words, have been found to be associated with diagnoses of various types of dementia in these previous works. The work presented in this paper examines additional speech fluency characteristics such as hesitation rate and vowel lengthening and their relation to the cognitive effects of medications that may be just as subtle at lower doses as early effects of dementia and primary progressive aphasia.

The results of linear regression on manually assessed speech discontinuity show that the mean duration of automatically recognized phones (vowels in particular) contributes information that accounts for variability in discontinuity not accounted for by the duration of silent pauses. This finding is consistent with prior experimental research in speech disfluency showing that vowel lengthening is one of the many ways in which speakers express hesitation in spontaneous speech (Shriberg, 1999; Bell et al., 1999). Our study validates a fully automated methodology for the assessment of vowel lengthening and shows that, in certain experimental environments (e.g., exposure to a cognitive-impairing medication), the phenomenon of vowel lengthening is robust enough to overcome any imprecision in the phone boundary identification by an automatic speech recognizer. These results are particularly encouraging because the automatic speech recognition system we used in this study is relatively simple and constitutes our baseline approach, thus further improvements to the acoustic and the language model are likely to strengthen these results.

We found that both LZP and TPM negatively affect speech fluency but not to the same extent. Higher TPM concentrations result in significantly more disfluent speech (as measured by SDR); however, the association of disfluency with LZP is less pronounced. Medically, the positive correlation between TPM and speech disfluency is consistent with reports of patients complaining of word finding problems while on TPM. We would also expect administration of LZP to have some effect on speech fluency. LZP belongs to a class of benzodiazepines that are used to relieve anxiety but also may have generalized sedative effects. The anxiolytic property of LZP would be expected to result in more fluent speech, whereas the sedative property may cancel out or even counteract that leading to increased variability in speech disfluency. The increased variability, coupled with the relatively small sample size of subjects on LZP, makes these results more difficult to interpret. However, the fact that a MVL, a manual measurement of vowel duration, is highly correlated with LZP concentration is at least suggestive that LZP indeed does have a pronounced effect on spontaneous speech. The correlation results between AVL, the automated counterpart of MVL, and LZP concentrations are not as pronounced, but are consistent with MVL findings. The correlation between AVL-C2 subset of phones and LZP concentrations is not consistent with MVL and AVL findings. This is likely due to the fact that AVL-C2 subset contains the vowel "ah" and the consonant "m" that compose the pronunciations of filled pauses and, therefore, lengthening observed with this subset correlates with SDR, which in turn correlates better with TPM concentrations than LZP

concentrations. Filled pauses were explicitly excluded from the computation of the MVL variable and thus may have been responsible for this discrepancy. These preliminary results indicate that LZP and TPM may have differential effects on speech disfluencies and vowel lengthening; however, further work with larger samples must be done to test this hypothesis.

Another important finding was that the concentrations of disfluencies in spontaneous speech narratives, captured by the SDR variable, correlated with concentrations of TPM in the blood, an objective physiological measure, whereas the standard phonemic fluency test scores did not. Moreover, we did not find a strong correlation between the phonemic or semantic fluency test scores and any of the automatically or manually determined discourse disfluency measurements; however, both the phonemic fluency test scores and all of the discourse disfluency assessments demonstrated significant differences in group means between subjects on TPM vs. PBO. These results lead us to believe that while both the standard phonemic fluency and the discourse disfluency measurements can reliably distinguish between *groups* of individuals on TPM vs. PBO, only the discourse disfluency measurements are sensitive to the *individual* drug concentrations. Therefore, we believe that the standard test of phonemic fluency and the disfluency in spontaneous speech reflect different underlying cognitive processes with differential sensitivity to exposure to TPM. Producing words that start with a certain letter of the alphabet places more demand specifically on working memory and executive control in addition to lexical retrieval, phonological encoding and motor planning; however, the fluency of a spontaneous speech narrative may be more reflective of global cognitive functioning and indicative of how well various cognitive systems interact to accomplish the complex task of thought organization and expression in speech (Lezak et al., 2004).

One of the possible concerns with any study including ours that involves multiple visits over time is the possibility of learning effects. We expected subjects to have more difficulty with the pictures in the first visit when the stimulus and the task were novel to the subjects and much less difficulty due to practice and learning effects during the last visit. Therefore, we expected the average score between the two baseline visits to be more representative of the true baseline performance on the task. Furthermore, the very possibility of having learning effects in our experiments makes our results stronger by making the overall design more conservative. For example, prior work on TPM has demonstrated that this drug has significant cognitive side effects. Thus, we would expect to observe worsening of function during the visits in which subjects receive TPM. A possible learning effect would tend to help subjects improve in their performance thus making it all the more difficult for us to detect any worsening of cognitive function. The fact that we did find worsening of cognitive function despite the possible learning effects in the TPM group strengthens our findings.

While the overall correlation between the phonemic fluency test scores and the SDR measurements was weak, we did find a relatively strong correlation between these measurements in the PBO condition. This finding is likely reflective of the placebo effect. Despite blinding, all subjects underwent informed consent by which they became familiar with the details of the study and possible side-effects of the medications. Thus, the subjects may have been exerting more effort in the experimental conditions (vs. baselines) in order to counter the anticipated negative effects of the medication. Therefore, in the experimental sessions where the subject received a placebo, this extra effort may have resulted in improved performance across all tests leading to a more clear divergence between spontaneous speech fluency and phonemic fluency.

In summary, our study found that paralinguistic disfluency characteristics obtained from spontaneous speech narratives may serve as a sensitive way to assess cognitive effects of medications such as TPM at the level of individual patients in addition to coarse-grained group-level distinctions. The use of automatic speech recognition technology to characterize spontaneous speech disfluency offers an additional advantage of objectivity and reproducibility of the measurements. Finally, discourse-level tests such as the picture description task are also more likely to reflect how one is able to function in daily activities such as interpersonal verbal communication and, therefore, are more ecologically valid than standard neuropsychological tests of verbal fluency. Thus, our study has also defined a unique niche for human speech and language technology in the realms of neuropsychological testing and neuropharmacology.

6. Limitations

This is a relatively small amount of training data for a typical speaker-independent acoustic model; however, according to studies by Lamel et al. (2002) and Moore (2003), the relationship between the amount of speech training data and word error rates is close to linear with a gradual slope. For example, Lamel's results show that that an ASR

system with an acoustic model trained on 1 hour of speech in a fully supervised mode performs at 33.3% word error rate. Adding another 32 hours of speech improves the word error rate by 12–20.7%; adding another 34 h only improves the performance by 2%. These data suggest that an ASR system trained on 1.3 h of speech would have a word error rate of a little less than 30% (or better than 70% accuracy) and adding significantly more data results in diminishing returns. Also, the speech collected to construct the acoustic model for this study was highly representative of the speech collected during subject testing, and thus this *task-dependent* model is likely to perform well for the purposes of this study as evidenced by the accuracy of forced alignment results; however, further improvements can clearly be made.

The use of healthy volunteers can be perceived as limiting the generalizability of the results of these studies to patients. However, since individuals who are prescribed drugs such as TPM are, by the very nature of their condition, often prone to cognitive impairments, i.e., persons with epilepsy, the use of patients in this initial study would diminish our ability to isolate the cognitive effects of the treatment from that of the underlying disorder. In addition, we excluded subjects who were not native speakers of English, have been diagnosed with a speech and/or language impairment/disability, who have uncorrectable low vision or have a dominant left hand (to control for brain lateralization of language). We anticipate, however, that as our speech analysis technology advances, and speech elicitation stimuli improve, future studies will be able to accommodate speakers of common languages other than English, as well as those who have language/speech disabilities.

Acknowledgements

This work was supported by the United States National Institute of Neurological Disorders and Stroke (NINDS) Grant R01-AG026390 (Birnbaum); a Faculty Research Development Grant from the University of Minnesota Academic Health Center (Pakhomov, Marino and Birnbaum). We would also like to thank Chamika Hawkins-Taylor for help with study coordination and administering neuropsychological tests, the staff of the University of Minnesota Clinical and Translational Science Institute (CTSI) for help with study logistics, University of Minnesota student Eden Kaiser and Dustin Chacon for help with speech transcription. Last but not least, we would like to thank the reviewers of this manuscript for their detailed and constructive critiques that helped improve the manuscript.

References

Krahmer, E., 2010. What computational linguists can learn from psychologists (and vice versa). Computational Linguistics 36, 586–594.
Shriberg, E., Preliminaries to a theory of speech disfluencies (PhD thesis), PhD thesis, University of California, Berkeley, 1994.
Hillis, A., 2007. Aphasia: progress in the last quarter of a century. Neurology 69, 200–213.
Rohrer, J., Knight, W., Warren, J., Fox, N., Rossor, M., 2008. Word-finding difficulty: a clinical analysis of the progressive aphasias. Brain 131, 8–38.
Meador, K., Loring, D., Hulihan, J., Kamin, M., Karim, R., 2003. Differential cognitive and behavioral effects of topiramate and valproate. Neurology 60, 1483–1488.
Mula, M., Trimble, M., Thompson, P., Sander, J., 2003. Topiramate and word-finding difficulties in patients with epilepsy. Neurology 60, 1104–1107.
Shriberg, E., 2001. To "errrr" is human: Ecology and acoustics of speech disfluencies. Journal of the International Phonetic Association 31, 153–169.
Oviatt, S., 1995. Predicting spoken disfluencies during human–computer interaction. Computer Speech and Language 9, 19–35.
Maclay, H., Osgood, C., 1959. Hesitation phenomena in spontaneous English speech. Word 15, 19–44.
Clark, H., 1996. Using Language. Cambridge University Press, Cambridge, England.
Butterworth, B., 1980. Evidence from pauses in speech. In: Butterworth, B. (Ed.), Language Production. Academic Press, London.
Levelt, W., 1989. Speaking: From Intention to Articulation. MIT Press, Cambridge, MA, USA.
Freud, S., Strachey, J., 1938. The psychopathology of everyday life. In: Freud, S., Brill, A., Arden, A. (Eds.), The basic writings of Sigmund Freud. Modern Library, New York.
Goldman-Eisler, F., 1958. Speech analysis and mental processes. Language and Speech 1, 59–75.
O'Connell, D., Kowal, S., Hörmann, H., 1969. Semantic determinants of pauses. Psychological Research 33, 50–67.
Erbaugh, M., 1987. A uniform pause and error strategy for native and non-native speakers. In: Tomlin, R. (Ed.), Coherence and Grounding in Discourse. John Benjamins, pp. 109–130.
Lounsbury, F., 1954. Transitional probability, linguistic structure, and system of habit-family hierarchies. In: Osgood, C., Sebeok, T. (Eds.), Psycholinguistics: A Survey of Theory and Research Problems. Waverly Press, Baltimore.
Goldman-Eisler, F., 1958. Speech production and the predictability of words in context. Quarterly Journal of Experimental Psychology 10, 96–106.
Goldman-Eisler, F., 1967. Sequential temporal patterns and cognitive processes in speech. Language and Speech 10, 122–132.
Beattie, G., 1983. Talk: An Analysis of Speech and Non-verbal Behavior in Conversation. Open University Press, Milton Keynes.

Bell, A., Jurafsky, D., Cynthia, E.F.-L., Girand, D., Gildea, 1999. Forms of English function words – effects of disfluencies, turn position, age and sex, and predictability. In: Proceedings of International Congress of Phonetic Sciences (ICPhS-99), vol. 1, San Francisco, CA, pp. 395–398.
Downing, B., Syntactic structure and phonological phrasing in English, PhD thesis, University of Texas, Austin, 1970.
Boomer, D., 1965. Hesitation and grammatical encoding. Language and Speech 17, 11–16.
Hawkins, P., 1971. The syntactic location of hesitation pauses. Language and Speech 14, 277–288.
Heeman, P., Allen, J., 1997. Intonational boundaries, speech repairs, and discourse markers: Modeling spoken dialog. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 254–261.
Cook, M., 1977. The incidence of filled pauses in relation to part of speech. Cognitive Psychology 37, 210–242.
Grosjean, F., Grosjean, L., Lane, H., 1979. The patterns of silence: performance structures in sentence production. Cognitive Psychology 11, 58–81.
Beattie, G., 1980. Encoding units in spontaneous speech: some implications for the dynamics of conversation. In: Temporal Variables in Speech: Studies in Honour of Frieda Goldman–Eisler, Mouton, The Hauge, pp. 131–143.
Kircher, T., Brammer, M., Levelt, W., Bartels, M., McGuire, P., 2004. Pausing for thought: Engagement of left temporal cortex during pauses in speech. NeuroImage 21, 84–90.
Roberts, B., Kirsner, K., 2000. Temporal cycles in speech production. Language and Cognitive Processes 15, 203–222.
Merlo, S., Barbosa, P., 2010. Hesitation phenomena: a dynamical perspective. Cognitive Processing 11, 251–261.
Wheatley, B., Doddington, G., Hemphill, C., Godfrey, J., Holliman, E., McDaniel, J., Fisher, D., 1992. Robust automatic time alignment of orthographic transcriptions with unconstrained speech. In: Proceedings of ICASSP'92, Vol. 1, San Francisco, CA, pp. 533–536.
Wightman, C.W., Talkin, D., 1996. The aligner: Text to speech alignment using Markov models and a pronunciation dictionary. In: van Santen, J. (Ed.), Speech Synthesis. Springer-Verlag New York, Inc.
Bratt, H., Shriberg, E., Algemy, a tool for prosodic feature analysis and extraction, Personal Communication, 2008.
Huang, Z., Chen, L., Harper, M., 2006. An open source prosodic feature extraction tool. In: Proceedings of Language Resource and Evaluation Conference, Genoa, Italy.
Heeman, P., Allen, J., 1993. The TRAINS 93 Dialogues, Technical Report, University of Rochester, Rochester, NY.
Hosom, J., 2002. Automatic phoneme alignment based on acoustic–phonetic modeling. In: Proceedings of International Conference on Spoken Language Processing (ICSLP), Vol. 1, Boulder, CO, pp. 357–360.
Shriberg, E., 1996. Disfluencies in switchboard. In: Proceedings International Conference on Spoken Language Processing, Philadelphia, PA, pp. 11–14.
Shriberg, E., 1999. Phonetic consequences of speech disfluency. In: Proceedings International Congress of Phonetic Sciences, Vol. 1, San Francisco, CA, pp. 619–622.
Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., Gildea, D., 2003. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. Journal of Acoustical Society of America 113, 1001–1024.
Heeman, P., McMillin, A., Yaruss, J.S., 2006. An annotation scheme for complex disfluencies. In: Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP), Vol. 1, Pittsburgh, pp. 4–7.
Pakhomov, S., Schonwetter, M., Bachenko, J., 2001. Generating training data for medical dictations. In: Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL'2001), Pittsburgh, PA.
Cronbach, L., Meehl, P., 1955. Construct validity in psychological tests. Psychological Bulletin 52, 281–302.
Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., 2006. The HTK Book Version 3.4. Cambridge University Press, Cambridge, England.
Young, S., Russell, N., Thornton, J., 1989. Token Passing: A Simple Conceptual Model for Connected Speech Recognition Systems, Technical Report, University of Cambridge: Department of Engineering, Cambridge, England.
Loonstra, A., Tarlow, A., Sellers, A., 2001. Cowat metanorms across age, education, and gender. Applied Neuropsychology 8, 161–166.
Benton, A., Hamsher, S., Sivan, A., 1994. Multilingual Aphasia Examination. AJA Associates, Iowa City.
Goodglass, E., Kaplan, H., 1983. The Assessment of Aphasia and Related Disorders. Lea and Febiger, Philadelphia, PA.
Schuell, H., 1965. Minnesota Test for Differential Diagnosis of Aphasia. University of Minnesota Press, Minneapolis, MN.
Nicholas, L., Brookshire, R., 1995. Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. Journal of Speech and Hearing Research 38, 145–156.
Boersma, P., Weenink, D., 2009. Praat: Doing Phonetics by Computer (version 5.1.05) [Computer Program].
Hird, K., Kirsner, K., 2010. Objective measurement of fluency in natural language production: a dynamic systems approach. Journal of Neurolinguistics 23, 518–530.
Forrest, K., Weismer, G., Turner, G., 1989. Kinematic, acoustic, and perceptual analyses of connected speech produced by parkinsonian and normal geriatric adults. Journal of the Acoustical Society of America 85, 2608–2622.
Subramanian, M., Birnbaum, A., Remmel, R., 2008. High-speed simultaneous determination of nine antiepileptic drugs using liquid chromatography-mass spectrometry. Ther Drug Monitor 30.
Zhu, H., Luo, J., 2005. A fast and sensitive liquid chromatographic–tandem mass spectrometric method for assay of lorazepam and application to pharmacokinetic analysis. Journal of Pharmaceutical and Biomedical Analysis 39, 268–274.
FDA, 2001. Guidance for Industry: Bioanalytical Method Validation. United States Department of Helath and Human Services, Rockville, MD.
Gorno-Tempini, M., Hillis, A., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S., Ogar, J., Rohrer, J., Black, S., Boeve, B., Manes, F., Dronkers, N., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B., Knopman, D., Hodges, J., Mesulam, M., Grossman, M., 2011. Classification of primary progressive aphasia and its variants. Neurology 76, 1006–1014.
Pakhomov, S.V.S., Smith, G.E., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., Knopman, D.S., 2010. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. Cognitive and Behavioral Neurology 23.

Roark, B., Mitchell, M., Hosom, J., Hollingshead, K., Kaye, J., 2011. Spoken language derived measures for detecting mild cognitive impairment. IEEE Transactions on Audio, Speech, and Language Processing (e-pub ahead of print).

Singh, S., Bucks, R.S., Cuerden, J.M., Ew, B.B., 2010. Evaluation of an objective technique for analysing temporal variables in dat spontaneous speech. Aphasiology 15, 571–583.

Lezak, M., Howieson, D., Loring, D., 2004. Neuropsychological Assessment, 4 ed. Oxford University Press, New York.

Lamel, L., Gauvain, J.-L., Adda, G., 2002. Lightly supervised and unsupervised acoustic model training. Computer Speech and Language 16, 115–129.

Moore, R.K., 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. In: Proceedings of Eurospeech, Geneva, pp. 2582–2584.