2016

Lexical and sublexical analysis of single-word reading and writing errors

BOSTON UNIVERSITY

SARGENT COLLEGE OF HEALTH AND REHABILITATION SCIENCES

Thesis

LEXICAL AND SUBLEXICAL ANALYSIS OF SINGLE-WORD READING AND WRITING ERRORS

by

KATRINA ROSS

B.A., Bard College, 2013
B.S., University of Washington, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2016

Approved by

First Reader ______________________________

Swathi Kiran, Ph.D. CCC-SLP
Professor of Speech, Language, and Hearing Sciences

Second Reader ______________________________

Elizabeth Hoover, Ph.D. CCC-SLP, BC-ANCDS(A)
Clinical Associate Professor of Speech, Language, and Hearing Sciences

Third Reader ______________________________

Anne Carney, M.S. CCC-SLP
Lecturer of Speech, Language, and Hearing Sciences

Fourth Reader ______________________________

Jeff Johnson, M.S. CCC-SLP
Research Advisor

ACKNOWLEDGEMENTS

I would first like to thank my supervisors Dr. Swathi Kiran and Jeffrey Johnson. This project was only possible because of their extraordinary commitment to both clients and students, their insightful guidance, and their unyielding encouragement. They consistently shared their vast wealth of knowledge and offered their generous patience, which allowed me to grow as a clinician and a researcher. I am so grateful for the opportunities they have given me during my two years in the Aphasia Research Laboratory.

I would also like to thank my committee members Dr. Elizabeth Hoover and Anne Carney. I have been fortunate enough to work and learn with these women throughout my program at Boston University, as well. I am so grateful for their dedication and expertise as teachers, supervisors, and clinicians. Their work has been foundational in all aspects of my graduate education, and I am honored that they have played such essential roles in my clinical and academic career.

A special note of gratitude belongs to my family and friends, who have been unerringly supportive throughout this rigorous program and beyond. You have been a vital source of comfort, humor, brilliance, and joy. To my D-Team, you are 'Angels', and I am so lucky to have you as colleagues and friends. To my sisters, your willingness to listen and discuss both my interests and your own have kept me as grounded as I've ever been. To my mom and Carmen, thank you for always offering me your trust and unconditional love.

Thank you to Shreya Ramesh and Marcos Zedan for taking on the potentially daunting task of translating this project into a language completely foreign to me. I learned so much from you and was continually amazed by your innovation, your excitement, and your commitment. Furthermore, because of your work, this project will be made available everywhere and can hopefully be used to help people to understand and recover language in a new way.

Finally, I want to thank the clients I have had the opportunity to work with during my time at Boston University. Like all of the people I have mentioned, these individuals are superlatively generous, patient, and kind. They never fail to teach and amaze me, and it is because of them that this work is both meaningful and possible.

This work was completed with the generous funding from the Dudley Allen Sargent Research Fund.

Thank you.

LEXICAL AND SUBLEXICAL ANALYSIS

OF SINGLE-WORD READING AND WRITING ERRORS

KATRINA ROSS

ABSTRACT

Within a dual-route neuropsychological model, two distinct but interrelated pathways are used to read and write, known as the lexical and sublexical routes. Individuals with reading and writing deficits often exhibit impairments in one or both of these routes, and therefore must rely on the combined power of the integrated system in print processing tasks. The resultant errors reflect varying degrees of lexical and sublexical accuracy in a single production. However, no system presently exists to analyze bimodal errors robustly in both routes. The goal of this project was to develop a system that simultaneously, quantitatively, and qualitatively captures lexical and sublexical errors for single-word reading and writing tasks. This system evaluates responses hierarchically in both routes according to proximity to a target. Each response earns a bivariate score [sublexical, lexical], which is plotted along x and y axes. This scoring system was developed using data from a novel treatment study for patients with acquired alexia/agraphia. Repeated-measures multivariate analyses of variance and post hoc analyses revealed a significant treatment effect in both the lexical and sublexical systems. Qualitative analyses were also conducted to evaluate patterns of change in both the trained and untrained modalities, in the sublexical and lexical systems. Overall, the results of this study indicate that treatment-induced evolution of reading/writing responses can be comprehensively represented by this novel scoring system.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

Reading and writing research has long been dedicated to investigating the existence of and the relationship between two pathways of print processing, known as the lexical (i.e. direct, whole-word, semantic) and sublexical (i.e. indirect, individual sound-to-letter and letter-to-sound conversion, phonological/graphemic) routes. Coltheart (1980), Patterson, Coltheart, & Marshall (1985), Ellis (1982), and Ellis and Young (1988) were among the first researchers to develop so-called "dual-route processing models" (Figure 1), which are schemata that distinguish how the brain comprehends information sublexically and lexically (see also: Beeson, Rising, Kim, & Rapcsak, 2008; Caramazza, 1988; Houghton & Zorzi, 2003; Rapcsak, Henry, Teague, Carnahan & Beeson, 2007; Tainturier & Rapp, 2001).

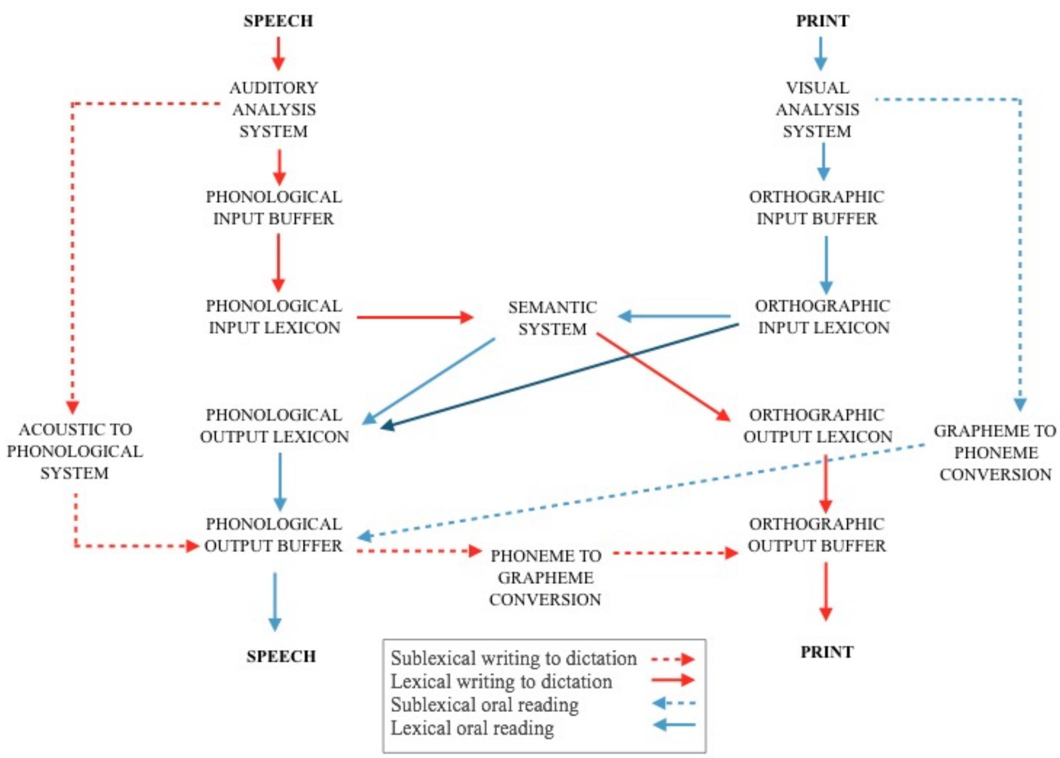


Figure 1. Dual-route model of single-word processing (adapted from Beeson, Rising, Kim, & Rapcsak, 2008, developed from Ellis & Young, 1988)

Within a dual-route processing model, the sublexical route is used for fine-grained analysis at the level of individual letters or sounds (graphemes or phonemes, respectively). Each letter or sound must be encoded discretely before connecting to construct a full word. It is, therefore, this pathway that allows readers and writers to parse low frequency words and novel words as well as nonwords. The unfamiliar graphemes and phonemes must be individually analyzed to gain access to mental representation, which all must precede word-level comprehension or production (Beeson, Rising, Kim, & Rapcsak, 2008; Howard & Gatehouse, 2006; Rapp & Caramazza, 1997). The sublexical route is often known as the phonological route, meaning that the process gains access to meaning via conscious and explicit phonological analysis operating below the full-word level.

The lexical route, on the other hand, is responsible for whole-word processing, when the brain analyzes a set of graphemes or phonemes as a complete unit. That is, when the components of a word are recognized without being decoded piece by piece, the lexical route is the primary system of access. This route is responsible for reading and writing irregularly spelled words (e.g. *height, yacht*), since their component graphemes cannot be decoded accurately with regular letter-to-sound correspondence rules. This is also the primary route of access for common or high frequency words, as the brain memorizes such forms in their entirety in order to increase the efficiency of processing. The lexical route is often called the semantic route, since this pathway maps word forms directly to word meanings. A typically functioning linguistic network uses both the sublexical and the lexical processes to decode and encode, to read and write (Beeson &

Rapcsak, 2002; Tainturier & Rapp, 2001).

In addition to the evidence supporting these systems as distinct in neurotypical individuals, the separation between sublexical and lexical processing is most often explored in the context of differential impairment, particularly in cases of alexia (reading impairment) and agraphia (writing impairment; Beauvois & Dérouesné, 1981; Shallice, 1981). Some persons with alexia/agraphia (PWA/A) experience a breakdown in the sublexical route. As this is the pathway of grapheme-phoneme correspondence (GPC), impairments of this route will likely result in impoverished nonword reading and writing since the component graphemes and phonemes no longer have independent representation in the language system. These deficits are known as *phonological*/*deep* alexias and agraphias, depending on the particular area of deficit. Both phonological and deep alexias/agraphias entail a GPC breakdown, but in the case of phonological alexias/agraphias, the connection to the semantic system remains undisturbed. This results in poor reading/writing of low frequency words and nonwords, as well as visual errors (e.g. *slid* as "mild"), but relatively few semantic errors (e.g. *slid* as "fell"). However, in cases of deep alexias/agraphias, deficits are present in both GPC and semantic access, resulting in all of the error types above (Cherney, 2004; Friedman & Lott, 2002; Hillis, Rapp, & Caramazza, 1999; Tainturier & Rapp, 2001 for discussions).

Other PWA/A who exhibit alexia and agraphia may indicate an impairment in the lexical route. This most commonly presents as difficulty recognizing or generating irregularly spelled words (e.g. Beauvois & Dérouesné, 1981; Goodman & Caramazza, 1986; Gvion & Friedmann, 2010; Roeltgen & Heilman, 1984; Shallice, 1981). Since irregular forms are memorized, the inability to comprehend or produce them is not a failing of GPC but rather an impoverished representation of the entire lexical item or access to it. Such deficits are referred to as *surface* alexias and agraphias.

However, despite the evident existence of two routes and the benefits of distinguishing between them in order to isolate the level of impairment, research has consistently shown that the systems are inherently and fundamentally linked (Davies, Cuetos, & Rodriguez-Ferreiro, 2010; Folk & Jones, 2010; Folk & Rapp, 2004; Folk, Rapp, & Goldrick, 2002; Hillis & Caramazza, 1991; Hillis, Rapp, & Caramazza, 1999; Rapp, Epstein, Tainturier, 2002; Seidenberg & McClelland, 1989; Tainturier, Bosse, Roberts, Valdois, & Rapp, 2013; Zorzi, Houghton, & Butterworth, 1998). One type of support for this lexical-sublexical interaction is found in research demonstrating the overlap in neurotypical individuals. For instance, Folk and Rapp (2004), studied the effects of real word priming on written spelling of nonwords in order to measure the influence of the lexical system on the sublexical output mechanism. They found that presenting a real word (lexical information) prior to a nonword-spelling task (sublexical task) effectively primed the desired spelling, confirming this overlap.

Apart from the research regarding sublexical-lexical interaction in neurotypical individuals, overlap between lexical and sublexical processing has also been repeatedly observed in individuals with aphasia. Kendall et al. (2003) provided phonological treatment to patients with alexia, thereby attempting to use the sublexical route to improve lexical function, and their work showed positive effects. Folk & Jones (2010) conducted real-word and nonword-spelling tasks under “articulatory suppression”, by asking a patient with dysgraphia (i.e. agraphia) to speak a string of nonsense syllables while writing to dictation. Among their findings was the fact that lexical substitutions are more common under the articulatory suppression task than under normal conditions, from which these authors conclude that the lexical process is more activated when the sublexical process is disrupted. Furthermore, Tainturier & Rapp (2001) discuss the overlap of modalities as well. In reading and writing impairments, they assert, "not only does dysgraphia usually accompany dyslexia but that specific types of dyslexia often co-occur with the very same types of dysgraphia" (275). For this reason, this project, as well as the treatment study upon which it is based, will address both reading and writing.

In many circumstances, from evaluation to treatment, it is appropriate to quantify patients' behaviors as either *correct* or *incorrect*. However, in order to gain insight into the qualitative aspects of their performance, it becomes necessary to describe and evaluate further. Several authors, including those mentioned thus far, have provided descriptors and classifications to elaborate on the binary correct/incorrect system when assessing participants' print processing. These qualitative analyses take patients' "incorrect" responses in particular and assess them in terms of the type and location of linguistic breakdown. Parsing error types in this manner not only clarifies the data of the particular study in which the errors occurred, but it also allows clinicians to evaluate the relationship between patient profiles and specific print processing mistakes which is essential to creating informed treatment plans. As with the studies themselves, the qualitative error analyses identify the loci of breakdowns in the sublexical, lexical, and/or combined route(s), consistent with the models of spelling and reading described above.

Table 1 presents an overview of some of the existing error scoring systems that are outlined below.

Each error type indicates a different locus of breakdown within the dual-route model. For example, by referring back to the sublexical pathways in Figure 1, it is clear that a breakdown at the orthographic output buffer would prevent grapheme production in the task of writing to dictation, while a breakdown in the phonological output buffer could prevent phoneme-to-grapheme conversion as well as preventing phoneme production in oral reading. Sublexical error classifications in the following coding systems can be considered as falling into two broad categories: individual phoneme/grapheme errors and phoneme/grapheme combination errors. The individual phoneme/grapheme types are those in which the subjects fail to encode or decode a specific selection of the components in the target. Errors in the phonological/orthographic output lexicon include deletions (e.g. "stor" for *store*), additions (e.g. "stoore"), substitutions (e.g. "spore"), and transpositions (e.g. "srote"; Falconer & Buchwald, 2013; Friedman, 1996; Rapp & Caramazza, 1997); phoneme perseveration errors are likely due to breakdowns in the input buffer, input lexicon, or output lexicon (e.g. "stote"; Lott & Friedman, 1996, 2002); and visuospatial errors indicate breakdowns in the visual analysis system (e.g. "slcrc"; Howard & Gatehouse, 2008; Rapp & Caramazza, 1997).

The other category type of sublexical error classification is a failure to organize and integrate phonological/orthographic information into a cohesive whole, which may indicate an impairment in the phonological/orthographic output lexicon or buffer. Some studies evaluate this based on the percentage of accuracy in the production as compared to the target, as in responses containing ≥50% of target phonemes or in targets containing ≥50% of response phonemes in the correct order (Howard & Gatehouse, 2006).

Another way to examine the errors that patients produce is in terms of their sublexical feasibility. Phonologically plausible errors (PPEs), as defined in Rapp & Caramazza (1997), are linguistically "legal" responses. They follow the basic sound and spelling rules of the language and they are reasonable alternate spellings (e.g. *debt* as "dett"; Beeson, 1999; Friedman, 1996; Joshi, 2006; Rapp, Folk, & Tainturier, 2000). This type of response indicates a relatively preserved GPC system, with a breakdown anywhere in the lexical route (Beeson, 1999).

Lexical system errors, on the other hand, qualify the sufficiency of access to a specific lexical-semantic item in a given trial. Such errors include production of semantic substitutions (e.g. "shop" for *store*), morphological errors (e.g. "stores"), and even lexically unrelated errors, which all indicate various degradations of the lexical system (Howard & Gatehouse, 2006; Friedman, 1996; Kim et al., 2011; Rapp & Caramazza, 1997). Additionally, other types of response that also indicate lexical deficits include circumlocutions in single-word tasks (e.g. "it's the place where you buy things" for *store*), lexicalizations in nonword tasks (e.g. "steeple" for *stemple*), and perseverations in either (Howards & Gatehouse, 2006; Lott & Friedman, 2002; Kim et al., 2011).

Table 1. Overview of Error Coding and Scoring Systems

AUTHORS	ERROR CODES AND DESCRIPTIONS	READING	WRITING	OTHER
BADECKER (1995)	Geminate (double grapheme) errors: shortening, shift, substitution, pseudo-substitution; transpositions		x	
BAHR, SILLIMAN, BERNINGER, & DOW (2012)	Constrained: PPEs Unconstrained: POMAS (phonological, orthographic, morphological errors)		x	
BAXTER & WARRINGTON (1985)	Complete and incomplete deletions ("yen" for "veneer"; "coweb" for "cobweb"); derivational + deletions ("applaus" for "applaud") ; additions ("princk" for "prick", "deprise" for "despise") ; letter order errors (e.g. "advangate"); > 50% overlap target letters and response; visual; homophones; miscellaneous		x	
BEAUVOIS & DEROUESNE (1981)	Deletions and additions of mute letters; grapheme substitutions with "the same phonetic value"; substitution + addition; substitution + deletion, transpositions/shifts; vowel simplification; miscellaneous	x	x	
BEESON (1999)	Single or multiple letter errors, correctly and incorrectly spelled semantic errors, visually similar, unclassifiable, no response		x	
BEESON, MAGLOIRE, & ROBEY (2005)	Visually similar word substitutions; unrelated words; reading rate; no response	x	x	
BEESON, RISING, KIM, & RAPCSAK (2010)	Correct/incorrect; low-probability, lexically correct errors (e.g. "jealish" for *jealous*); sublexical + lexical errors (e.g. "rythum" for *rhythm*)	x	x	
BOSE (2013)	Phonological Overlap Index (POI): NShared · 2/(LT + LE) NShared = number phonemes shared between target and error regardless of position; LT = phonemic length of target; LE = phonemic length of error *values that range from 0 to 1; 1 representing complete overlap.			x
BUCHWALD & RAPP (2003)	Substitutions, deletions, additions, transpositions, and shifts (high rate of CV preservation in substitution errors); morphological and phonologically plausible errors , semantic errors (not observed)		x	

CARAMAZZA & MICELI (1990)	Single vs. mixed errors: substitutions (+1), insertions/additions (+1/2 before and after), transpositions/exchanges (+1 for both), shift (+1 for moved letter, +1/2 before moved letter) deletions (+1); geminate consonants (sub, duplication, deletion, exchange, shift, exchange of feature and consonant, sub of one); anticipations, repetitions, ambiguous		x	
CHIALANT, DOMOTO-REILLY, PROIOS, & CARAMAZZA (2002)	Substitutions, geminate errors, letter case		x	
DAVIES, CUETOS, & RODRIGUES-FERREIRO (2010)	Correct, no response, word error (semantic, morphological, visual, visual+semantic, visual then semantic, functor substitution, or unrelated errors), nonword error, or circumlocution	x		
FALCONER & BUCHWALD (2013)	Lexical selection errors, morphological errors, 'misspellings' (PPEs), mixed errors (orthographic form and meaning)		x	
FALCONER, MINER, VELEZ, & BUCHWALD (2011)	Semantic errors, mixed errors (semantic + orthographic)		x	
FOLK & JONES (2010)	Words: PPEs, nonwords, lexical substitutions Nonwords: orthographically legal nonwords, words		x	
FRIEDMAN & LOTT (2002)	Semantic paralexias, paralexias + correct initial phoneme, circumlocutions	x	x	
FRIEDMAN (1996)	PPEs, deletions, substitutions, orthographically similar errors			
GOMEZ, RATCLIFF, & PEREA (2008)	Overview of sublexical overlap systems			x
HOWARD & GATEHOUSE (2006)	Correct, visual errors, semantic errors (shared features, associates, circumlocution, semantic followed by phonological, mixed semantic/phonological errors), phonologically related errors (if 50% ≤ target's phonemes in the response or 50%≤ response's phonemes in target in roughly the same order; failed to meet either of these criteria = unrelated), unrelated, NR, or other	x		

JÓNSDÓTTIR, SHALLICE, & WISE (1996)	Substitutions, deletions, additions, exchanges, shifts, compounds, PPEs, morphological errors/VSW, stroke errors, visual + position (with mid-word errors most common)		x	
KIRAN, BALACHANDRAN, & LUCAS (2014)	10-point error scale: NR; neologism; perseveration; unrelated word; circumlocution; semantic error; mixed error; phonemic error; correct in nontarget language; accent influence in target language			x
RAPP & CARAMAZZA (1997)	Letter error: visuospatial, stroke-feature, neither, both Letter-level errors: addition, substitution, transposition, deletion, other (case shift, etc.); word level: PPEs, semantic, visually and phonologically similar, ambiguous		x	
RAPP, EPSTEIN, & TAINTURIER (2002)	PPEs; semantic errors; morphological errors; letter-level errors (addition, substitution, transposition, deletion); perseverations of single and multi-letter sequences	x	x	

Finally, the interaction of access in lexical and sublexical errors is indicated in the qualitative descriptions in a small number of studies, critically Howard & Gatehouse (2006) and Rapp, Epstein, & Tainturier (2002). Howard & Gatehouse (2006), for instance, used the classifications of semantic followed by phonological errors, and mixed semantic-phonological errors. Rapp et al. (2002) extend the analysis of Rapp & Caramazza (1997) by investigating low frequency PPEs that still preserve partial lexical accuracy to the target. For example, spelling *bouquet* as "bouket" demonstrates the phonologically plausible and lexically correct, though low frequency, spelling of /eɪ/ as "et". However, the similarly low frequency /k/ as "q" is not preserved. Rapp et al. (2002), in describing these errors, began to extend the fact of lexical-sublexical interaction into error analysis. The lexical representation of the word "bouquet" influenced the spelling of the final syllable, but the sublexical phoneme-grapheme correspondence rules for /k/ determined the spelling of the medial consonant. This analysis is the only example so far to bridge the gap between the dual routes in error analysis. However, its greatest shortcoming lies in its lack of a means with which to objectively compare and contrast the accuracy of the two routes.

The solution to this disparity between quantitative and qualitative analyses was found in a study by Kiran, Balachandran, & Lucas (2014). Kiran et al. examined the nature of naming errors in bilingual individuals with aphasia. The significance of this study in relation to the current topic of error analysis is that the researchers developed a hierarchy of accuracy. Responses were classified by error type and then those types were graded based on their proximity to the target, so all responses obtained values reflecting their relative accuracy. When the target item was produced in the target language – a successful execution of the task – the response earned the highest score (in this case a 10.5). At the other end of the hierarchy, a "don't know" response in the non-target language, as the least accurate, earned the lowest score (1).

The benefit of Kiran et al.'s scoring system is three-fold. First, it allows for efficient, qualitative analysis of a subject's success, since high scores intuitively correspond to strong, positive results. Second, it provides a point of reference when tracking an individual's change over time or as a function of rehabilitation. If scores increase from pre-treatment to post-treatment, this coding hierarchy provides a quantitative measure of that improvement while maintaining qualitative value. Third, unlike a binary scoring system, a graded scale reflects how even error responses approach the target, without being entirely "correct". This particular system, however, was developed for naming and so does not account for the parameters specific to print processing access routes or modalities (i.e. lexical/sublexical distinctions, grapheme-phoneme correspondence).

Despite the evidence supporting the existence of both lexical and sublexical routes—as well as the research demonstrating how the two systems necessarily interact in any given print task—there is no coding system that allows for a thorough, simultaneous analysis of the lexical and sublexical processing that underlies a single response. Furthermore, in order to analyze errors and error progression, responses within the dual processes should not only be categorized, but should be hierarchically scored to judge relative accuracy. Finally, this code must be applicable to errors in both reading and writing, since the two modalities interact just as the two routes do. The existing coding systems are useful when endeavoring to understand the locus, nature, and extent of impairment, but none of them provides sufficient information to achieve all of the following important criteria: (a) track changes objectively below the level of binary accuracy, (b) to compare lexical to sublexical performance in a single item and across items, and (c) to compare reading performance to writing performance.

In the current project, a novel scoring system was developed in order to meet these needs. This scoring system, therefore, aims to classify and evaluate the lexical and sublexical error evolution of patients with acquired alexia and agraphia, in both reading and writing. By analyzing errors in both routes and both modalities, assessing errors hierarchically according to proximity to the target, and tracking error progression over time, this study aims to qualify and quantify evolution in performance. This coding system, if proven to capture performance and progress effectively, may then be used to analyze error evolution across patient types, modalities of response, and access routes. This will allow clinicians to easily identify error patterns in order to inform therapy tasks, quantify progress outside of other formal measures, and justify insurance reimbursement.

Data from eight participants in a reading and writing treatment study at the Boston University Aphasia Research Laboratory were collected. In this project, the following research questions were addressed:

1. *Does a novel scoring system focused on hierarchical, two-dimensional evaluation of print processing errors capture significant differences in error type as a result of treatment?*

Eight patients received treatment for either alexia or agraphia with training focused on oral reading or writing to dictation, respectively. Their performance in both modalities was evaluated before, immediately after, and in a follow-up phase of treatment, and all responses were evaluated using the sublexical and lexical scoring hierarchies developed in the present project. It was hypothesized that a novel scoring system that analyzes errors a) hierarchically according to proximity to a target, b) in both print processing routes and modalities, and c) over time, would reveal significant effects of treatment in multiple ways. Due to the presumed interaction between sublexical and lexical systems (Davies, Cuetos, & Rodriguez-Ferreiro, 2010; Folk & Jones, 2010; Folk & Rapp, 2004; Hillis, Rapp, & Caramazza, 1999; Rapp, Epstein, Tainturier, 2002; Tainturier, Bosse, Roberts, Valdois, & Rapp, 2013), as well as the bi-dimensional nature of this treatment approach, it was believed that both the sublexical and lexical scores would increase as a result of treatment. Furthermore, due to the bimodal interaction of reading and writing (Chialant, Domoto-Reilly, Proios, & Caramazza, 2002; Folk, Rapp, & Goldrick, 2002), it was believed that treatment effects would be observed in the trained modality, as well as in the untrained modality to a lesser degree (Orjada & Beeson, 2005).

2. Do these lexical and sublexical error scores capture discrete, qualitative error progression below the level of item-wise accuracy?

It was hypothesized that qualitative (i.e. descriptive and visual) analyses would illustrate discrete changes and patterns of progress in the two processing systems for individual patients and for the group as a whole. The two scoring systems are organized hierarchically (Kiran, Balachandran, & Lucas, 2014), so progress in treatment should be evident in errors progressing from low to high scores.

3. Does this coding system demonstrate reliability as evaluated by agreement with an automated scoring system?

Finally, to evaluate the reliability of this scoring system, an automatic coding software program in Python was developed in a related project. The manual (i.e. clinician) coding of patient responses was then supplemented and examined by the automatic scores generated by this computer software. It was hypothesized that automated scores would correlate highly with manually generated results for all patients at all times, since the automated systems were developed according to the rules outlined in the present study.

METHODS

Subjects

Eight individuals with alexia/agraphia (PWA/A) and a primary diagnosis of aphasia subsequent to a cerebrovascular accident (CVA), ages 44–74 (mean 62.75), participated in the treatment study from which data were drawn. For additional demographic data, please refer to Table 1. Inclusion criteria included diagnoses of alexia and agraphia and the ability to complete standardized testing using the Psycholinguistic Assessment of Language Processing in Aphasia (Kay, Lesser, & Coltheart 1992), Boston Naming Test (Goodglass, Kaplan, & Weintraub, 2001), and the Western Aphasia Battery-Revised (Kertesz, 2006). One participant was excluded from this study, due to ceiling performance at baseline (achieving >65% in probes for both modalities). Other exclusion criteria consisted of the inability to complete pre-testing or if they exhibited severe enough alexia/agraphia as to be unable to execute any of the treatment tasks (e.g. inability to manipulate pen; hearing impairment preventing writing to dictation). Neither dysarthria nor apraxia of speech was grounds for exclusion; one participant had an apraxia diagnosis. (P4), another exhibited a mild flaccid dysarthria (P6) since overt production was not a requirement of this study. All participants discontinued any independent reading and writing treatment for the duration of the study, though other speech and language services continued as previously arranged.

Table 2. Patient Demographics

PARTICIPANT	P1	P2	P3	P4	P5	P6	P7	P8
TREATMENT MODALITY	W	W	R	R	W	W	W	W
AGE (YEARS)	53	60	74	46	67	71	77	67
SEX	M	M	M	M	M	F	M	F
HANDEDNESS	L	R	R	R	L	L	R	R
EDUCATION (YEARS)	12	16	20	12	16	18	14	18
TIME POST ONSET (MONTHS)	192	30	73	20	110	52	177	72
APHASIA TYPE	Broc.	Wern.	Wern.	Broc.	Wern.	Anom.	Anom.	Broc.
ALEXIA TYPE	Deep	Deep	Deep	Phon.*	Deep	Phon.	Phon.	Phon.
AGRAPHIA TYPE	Deep	Deep	Deep	Phon.	Deep	Deep	Deep	Deep
WAB AQ	52.6	59.7	37.6	46.9	37	80.6	90.6	67.4
WAB LQ	50.8	42.6	39.5	52.8	32.1	75.9	79.6	68.1

W = writing; R = reading; Broc. = Broca's; Wern. = Wernicke's; Anom. = Anomic; phon. = phonological; WAB = Western Aphasia Battery – Revised (Kertesz, 2006); AQ = Aphasia Quotient; LQ = Language Quotient. *based on limited productions due to high number of non-responses in oral reading subtests at baseline; likely attributable to pt's moderate-severe apraxia of speech

Procedure

Stimuli

Each of the participants trained in a single modality (oral reading; writing to dictation) on a set of 16 words (e.g. *honey*). The 16 trained items were selected from a larger set of baseline probe items. The selection was made based on item frequency, length, and the existence of a semantic associate (SA; e.g. *bee*) as well as a phonologic/orthographic neighbor (PN/ON; differing from the trained item by one letter, e.g. *money*), that all matched the target for frequency and length. (See Johnson, Ross, & Kiran, 2015 for full discussion of stimuli development and selection), which were probed before, during, and after treatment to measure for generalization effects (n = 48). Additionally, 10 irregularly spelled words (e.g. *laugh*) were also probed as control items.

The Treatment Protocol

The protocol for treatment in the study by Johnson, Ross, & Kiran (2015; Appendix 1A and 1B) consisted of the following steps: 1)*lexical decision between real and nonword, 2) attempted production of target word in trained modality (reading; writing), 3) verbal repetition/written copy of target from a direct model, 4)* picture selection of target, 5) semantic feature generation or decision, 6) GPC for target letters and distractor letters, 7) spelling of target with letter tiles, 8) GPC for letters of word in context, 9) PGC for letters of the target, 10) production of target (reading; writing), 11) verbal repetition/written copy of target; 12) delayed production of target (reading; writing). Steps *1) and *4) of treatment administration were removed for P5 – P8 in order to improve efficiency. Participants completed these steps for all 16 trained items as many times as possible during each session, and they continued to cycle through their trained list in this manner until the end of the treatment phase. Due to individual variability in pacing and accuracy, the number of repetitions in a session (and therefore the number of overall exposures to each item) was not controlled.

In Johnson et al. (2015)'s study, 58-item probes (16 trained items, 16 SAs, 16 PNs/ONs, 10 irregular) were administered at the beginning of every session, once a week in reading, once a week in writing. This allowed for weekly monitoring of treatment effects and generalization, to both untrained words and the untrained modality. The other time points at which data were collected include immediately post-treatment and 6–8 weeks later. Three full probe sets in both modalities were collected immediately post-treatment, and one full set in both modalities was collected at the follow-up date. Pre, post, and follow-up baselines were exclusively analyzed in this study, though the data collected during treatment are available for future analysis and research.

Of note, P5, P6, P7, and P8 trained in the writing modality after the protocol had been slightly modified, as discussed above. Additionally, P6, P7, and P8 demonstrated a ceiling effect when reading single words, so these 3 participants read 2- to 4-word phrases containing the target item or the target in a compound word (e.g. trained item "*hair*", target phrase "*hair*dresser and barber"). In order to account for such responses in this project, only the target word (or the compound derived from the target) was scored using the typical sublexical and lexical codes.

The Scoring Hierarchies

The error scoring system developed in this study serves two primary goals: first, it quantifies the lexical and sublexical parameters of a single response; second, it can equally be applied to the processes of reading and writing. Both dimensions afford higher scores as responses approach their targets.

Table 3. Sublexical and lexical scoring hierarchies

SUBLEXICAL			LEXICAL	
Correct length, >50% overlap with target	**Target** (e.g. "pie")	**S9**	**Target** (e.g. "pie")	**L9**
	GPE (e.g. "pai" for *pie*) or **PPE** (e.g. "pea" for *pie*)	**S8**	**Target + morphological error** (e.g. "pies")	**L8**
	Addition (e.g. "piel")	**S7**	**Related word** (e.g. "crust")	**L7**
	Transposition (e.g. "pei")	**S6**	**Unrelated word** (e.g. "snake")	**L6**
	Substitution (e.g. "bie")	**S5**	**Related description/circumlocution/gesture** (e.g. "dessert filled with fruit")	**L5**
	Deletion (e.g. "-ie")	**S4**	**Perseveration – real word** (repetition of a target or response within the previous 3 items)	**L4**
	Multiple errors (e.g. "paee")	**S3**	**Nonword** (e.g. "piel")	**L3**
Correct length ≤50% overlap with target (e.g. "mle")		**S2**	**Unrelated description** (e.g. "it's a place you go sometimes")	**L2**
Incorrect length ≤50% overlap with target (e.g. "rmle")		**S1**	**Perseveration – nonword** (>50% overlap with a target or response within the previous 3 items)	**L1**
No response		**S0**	**No response**	**L0**
*GPE = graphemically plausible error (written modality); PPE = phonemically plausible error (reading modality); all examples = "response" in quotations for the target word *pie* in italics				

The sublexical system involves a fine-grained assessment of the individual phonemes or graphemes in each response. In Table 3, examples of each type of response are provided. Responses are presented within "quotation marks", and the target item in all examples is the word *pie,* presented in italics. Scores range from ***no response*****,** which yields a sublexical score of zero [S0], to the ***target*** (i.e. *pie*) [S9]. Other responses that do not meet either of those criteria were then assigned a value between the two poles as follows, described from most accurate to least. A score of [S8] indicates a ***graphemically/phonemically plausible error (GPE/PPE)***, as described by Rapp, Epstein, & Tainturier (2002). An example of a ***GPE*** in the written modality would be spelling the word as "pai", which would result in the same sounds (/paɪ/). In reading, an example of a ***PPE*** would be "pea" (or /pi/), since the letters "-ie" can also be pronounced as such. ***GPE/PPE*** is considered to be the closest response to the target in the sublexical system because it indicates that participants successfully identified the most likely sound to be associated with a given letter, despite the "deep," "opaque" nature of English orthography (see Joshi, 2006; Rapp, Folk, & Tainturier, 2000). It should, however, be noted that in this particular reading and writing study, such errors may not be common. This is because, in determining the stimuli for treatment, the developers selected words based on the primary criteria of having both a semantic associate and an orthographic neighbor, as well as words that were matched for length and frequency. As a result, the most common grapheme-phoneme correspondences were not always trained, and so ***GPEs*** and ***PPEs*** should not be expected as often as they might be in a different treatment protocol.

All scores between [S7] and [S3] indicate that the response overlaps by more than (>)50% with the target item. A maximal overlap system was used to determine this 50% cutoff, similar to that discussed in Gomez, Ratcliff, & Perea (2008). [S7] is the score associated with a single grapheme or phoneme ***addition***. For example, "piel" would be counted as an ***addition***, as would "pies", "spie" and "plie". This is considered to be the closest approximation of the target after a ***GPE/PPE*** because it is the only sublexical score indicating that the target was produced in full and in order. This order may have been disrupted, as in the case of mid-word additions, however, the entirety of the target still must be produced with the component graphemes and phonemes in the correct relative sequence in order to earn this score. Below a single ***addition*** is a ***transposition*** [S6], which is defined as possessing all the components of the target with either (a) two of those pieces having been switched in place, or (b) one of those elements having been shifted to another place in the word (as in the classification system used by Caramazza &

Miceli, 1990). In the case of the target *pie*, the graphemes "i" and "e" may be transposed as in "pei", the phonemes /aɪ/ and /p/ may be transposed "ipe", or a shift may occur, such as "p" moving to the end to produce "iep".

Responses that are accurate to the target except for a single ***substitution*** earn the sublexical score of [S5]. This is the first score in which the participant still correctly produced more than half of the target, but now has inaccurately retrieved one element (e.g. substituting the letter or sound "b" for "p", resulting in "bie"). However, the response does represent the missing element, essentially preserving the shape and structure of the target with an erroneous grapheme/phoneme. [S4], on the other hand, indicates the ***deletion*** of a single grapheme/phoneme without any attempt to replace the empty space (e.g. deletion of "p" for "ie"). The last production that overlaps by more than 50% with the target is [S3], in which ***multiple errors*** have been made, regardless of whether these errors are all of the same type (e.g. ***multiple additions*** in the case of "piies") or of several different types (e.g. an ***addition*** and a ***substitution***, as in the above example "paee").

Responses that overlap with half of the target or less (***≤50%***) are then further distinguished by their length. If the production has the ***correct number of graphemes/phonemes with ≤50% overlap***, it earns a score of [S2], as this type of response retains at least the appropriate length of the target. The example "mle" is such a production, and it is intuitively distant from the target. In this case, the production also contains one correct grapheme, though this need not be the case. If a response has the correct number of graphemes/phonemes, it obtains a score of [S2]. Finally, if the response overlaps by ***≤50%*** and is also the ***incorrect length***, it is considered to have the least sublexical similarity to the target, and earns the lowest numerical, [S1], as with the example "rmle", which is clearly maximally dissimilar to the target.

The lexical dimension, just as in the sublexical dimension, has codes that range from ***no response*** (NR, [L0]), to the ***target*** [L9]. A score of [L8], [L7], or [L6] indicates a single-word response that differs by its lexical proximity to the target. Thus, immediately below the correct score is the score for a ***morphological error*** [L8]. The previous example of a ***sublexical addition***, "pies", would earn the ***lexical*** score for ***morphological error***. These items indicate strong lexical and semantic activation, with preservation of the meaning and the majority of the whole-word form. It deviates from the target only in tense, voice (active or passive), or number (singular or plural), with all other elements of the word preserved. Below that is a lexically or semantically ***related word*** [L7] (such as "crust" for pie) followed by an ***unrelated word*** [L6] (such as the word "snake", which does not have any features in common with the target). These three response types ([L6–L8]) earned the highest scores because they indicate that participant (a) appropriately produced the goal behavior of the probe (i.e. oral reading, writing to dictation), (b) demonstrated some lexical access (Beeson et al., 2008; Ellis & Young, 1988; Falconer & Buchwald, 2013; Hillis, 2002), and (c) attempted to generate a novel response according to the elements of the stimulus.

Responses that involved a type of behavior other than what the probe called for (e.g. description, gesture, nonword production, perseveration) received lower scores [L5–L0]. The score [L5] is allocated to ***accurate descriptions, gestures, and drawings*** of the target. This is because, while such a production does not entail the target task of GPC/PGC, a description/gesture/drawing does require accurate decoding in order to access the appropriate semantic network (Beeson & Rapcsak, 2002). Related descriptions/gestures/drawings, therefore, earn lower scores than unrelated real words because the emphasis is on lexicality rather than semantics. [L4] is the code for a ***real-word perseveration***. Perseverations in this system are judged when the entirety of a previous item (either a target or a response) is replicated within 3 trials. Therefore, if the incorrect and unrelated response "snake" for the target *pie* were once again produced two items later as "snake" or "snakes", it will be counted as a ***perseveration***. Although this is closer to the target in manner of production than a description, gesture, or drawing, it is judged to be farther from the target in lexical access.

Below real word perseverations are ***nonwords*** [L3]. This type of response does not contain lexical value; however, it is unclear whether the breakdown occurred prior to or following access to the lexicon (Beeson & Rapcsak, 2002). Additionally, a nonword still indicates a novel production in response to a target, as well as the generation of a discrete set of graphemes/phonemes as the task calls for. Most of the sublexical examples given in the previous section would have earned a score of [L3], such as "pai," "pei," "bie," "paee," and "mle". Inaccurate descriptions, imprecise gestures, and irrelevant drawings all earn the score of [L2] as these responses do not demonstrate any lexical access. Finally, a score of [L1] is assigned to nonword perseverations, because neither does this response carry lexical value nor does it demonstrate attention to the relevant target item. The criterion for a nonword perseveration is >50% overlap with one of the three preceding stimulus or response items.

Data Coding and Scoring

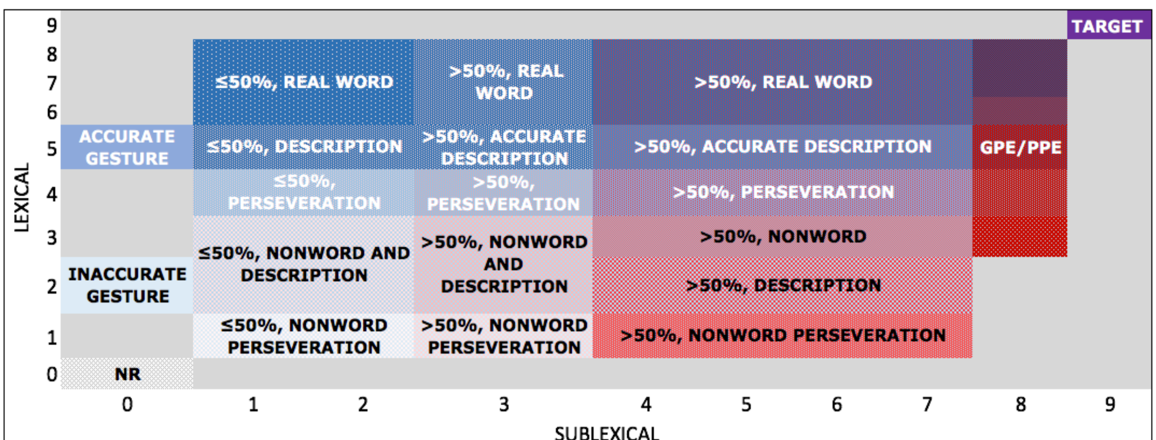
All responses were scored for each of the participants (i.e. 58 words x 2 modalities x 7 time points) according to the lexical and sublexical dimensions described above. The irregularly spelled control words (e.g. *yacht,* 10 per patient) were not included in the final analyses as neither writing nor reading for these words was predicted to change as a result of treatment. Therefore, the total number of scored responses included ((58 total words - 10 irregulars =) 48 x 2 x 7 x 8 = 5,376).

Pre-treatment and post-treatment matrices were generated for each patient at every time point and modality, resulting in a [Sublexical, Lexical] value for all target words. For example, P1's productions of the target *brush* had 3 pre-treatment scores in writing (e.g. [S1, L6], [S1, L3], [S0, L0]) and 3 in reading (e.g. [S9, L9], [S3, L6], [S9, L9]), as well as 3 post-treatment scores in both modalities, and 1 follow-up score in both modalities. These scores were averaged within a time point to obtain a single item score for pre, post, and follow-up performance in reading and writing. In the case of P1, for instance, his average pre-treatment score for *brush* in writing was [S0.7, L3.0] and average pre-treatment score in reading was [S7, L8]. Neither of these represents the error type, though, they are simply mean values.

Statistical Analysis

To answer Research Question 1, regarding the effect of treatment on lexical and sublexical accuracy, a composite average of all 48 items at each time point in both modalities was taken for each patient. For instance, P1 earned an average pre-treatment writing response of [S1.9, L2.8], indicating that his pre-treatment writing responses were distant from their targets lexically and even more distant sublexically. Repeated-measures multivariate analyses of variance (rMANOVAs) were used to compare pre-treatment to post-treatment and post-treatment to follow-up performance in both the trained and untrained modalities. Time was used as the within-subject factor and patient as the between-subject factor, while sublexical and lexical scores were the dependent variables.

In order to generate visual representations of sublexical/lexical interaction, the two scores for each response were used as coordinates along *x* and *y* axes (e.g. "tow" for *brush*, which earned the score [S1, L6], had an *x* value of 1 and a *y* value of 6; Refer to Figure 2). Each response earning that particular value (i.e. [S1, L6]) was then counted



NR = no response; ≤50%/>50% indicate graphemic/phonemic overlap between the target and response; GPE/PPE = graphemically/phonemically plausible error

Figure 2. *Schematic overview graph of error types. Quadrant Descriptors: lower left = low sublex., low lex. ; lower right = high sublex., low lex.; upper left = low sublex., high lex.; upper right = high sublex., high lex.*

and plotted, resulting in a single graph for each time point and modality displaying the number of each type of production. Overall, the four quadrants of the resultant graphs represent different strengths and weaknesses of response type. A schematic representation of this is presented in Figure 2, with the density of blue shading indicating lexical accuracy and the density of red representing sublexical accuracy. The lower left quadrant of the graph reveals low lexical and sublexical accuracy, indicating a response with very little similarity to the target. The top left quadrant represents strong lexical accuracy with low sublexical accuracy. The bottom right quadrant denotes high sublexical proximity to the target but low lexical value. Finally, the top right quadrant displays those responses that were closest to the target in both parameters.

In order to answer Research Question 2, regarding the types and patterns of errors made by each patient throughout treatment, 14 graphs were made for each patient, representing their performance at every point in the study ((3 pre + 3 post + 1 follow-up) x 2 modalities). These graphs were translated into heat maps, which are gradient color representations that change color based on the value of a cell, so that the most frequent response types were highlighted. Finally, the three pre- and three post-treatment maps were summed, resulting in a total of 6 maps per patient (3 time points x 2 modalities), each representing the distribution of their 144 productions (48 items at each time point x 3 baselines). Sums were used, rather than means, in order to preserve the qualitative value of the lexical and sublexical parameters, since this descriptive information was lost when the responses were averaged together (e.g. P1's average *brush* score of [S0.7, L3.0] could no longer be interpreted as a specific sublexical or lexical response type).

Finally, in order to address Research Question 3, concerning the reliability and objectivity of lexical and sublexical scoring, computerized versions of both systems were developed. Two undergraduate computer science students at Boston University were recruited for this related project in which they translated the existing scoring systems into code programs using Python 2.7.10 (Van Rossem, 2015). The dictionary FuzzyWuzzy (Cohen, 2014) was used to assess GPEs and PPEs in the sublexical system; the Natural Language Toolkit (Bird, Loper, & Klein, 2009) was used to distinguish real words from nonwords and to determine relatedness in the lexical system. With these scoring programs, the user inputs the target item(s), followed by the client's response(s), and the system returns the appropriate score. At present, the two programs remain in separate systems, however work is ongoing to combine the two.

For the present project, these automatically generated scores were compared to manually generated scores for bidirectional agreement and accuracy for all responses. In cases of the coding program identifying a clinician error, the data were adjusted to reflect the appropriate score. In cases of an automated system error (e.g. identifying an unrelated word as 'related'), the clinician manually overrode the response. A discussion of the nature of these discrepancies is included below.

RESULTS

Research Question 1: Does a novel scoring system focused on hierarchical, two-dimensional evaluation of print processing errors capture significant differences in error type as a result of treatment?

1.1 Trained Modality, Pre-Treatment to Post-Treatment

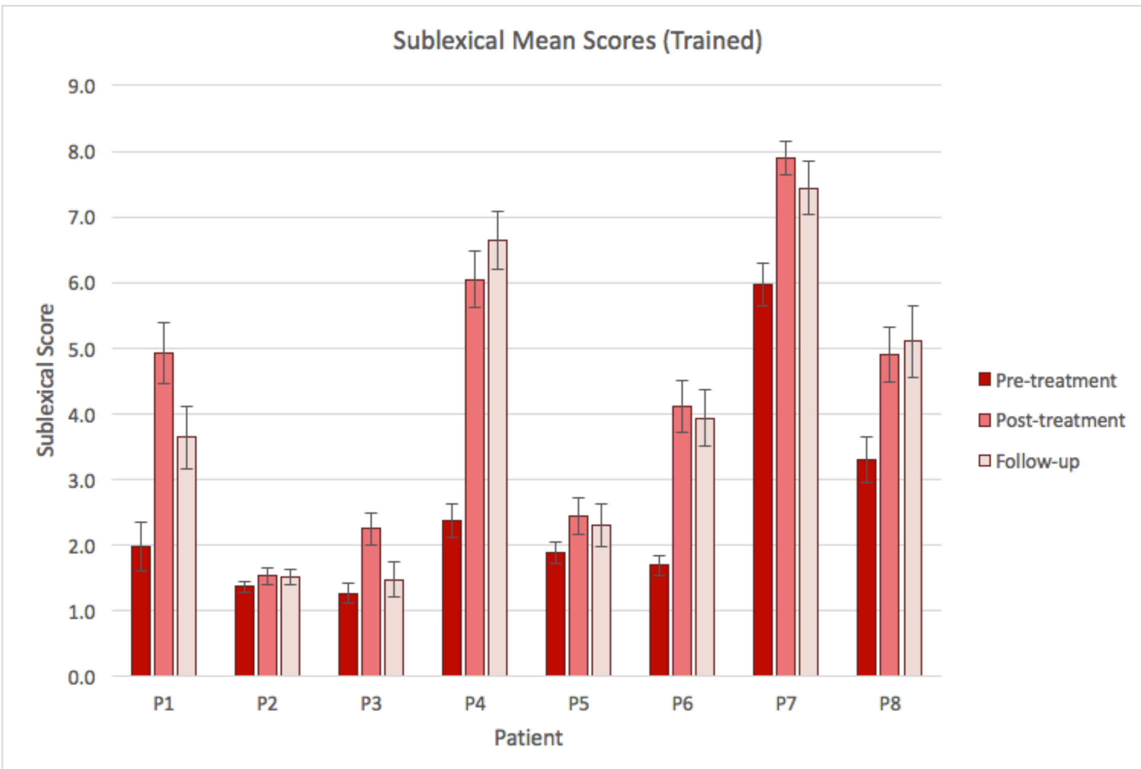


Figure 3. Sublexical mean scores for patients' trained modalities (pre-treatment, post-treatment, follow-up); P3 and P4 trained in reading

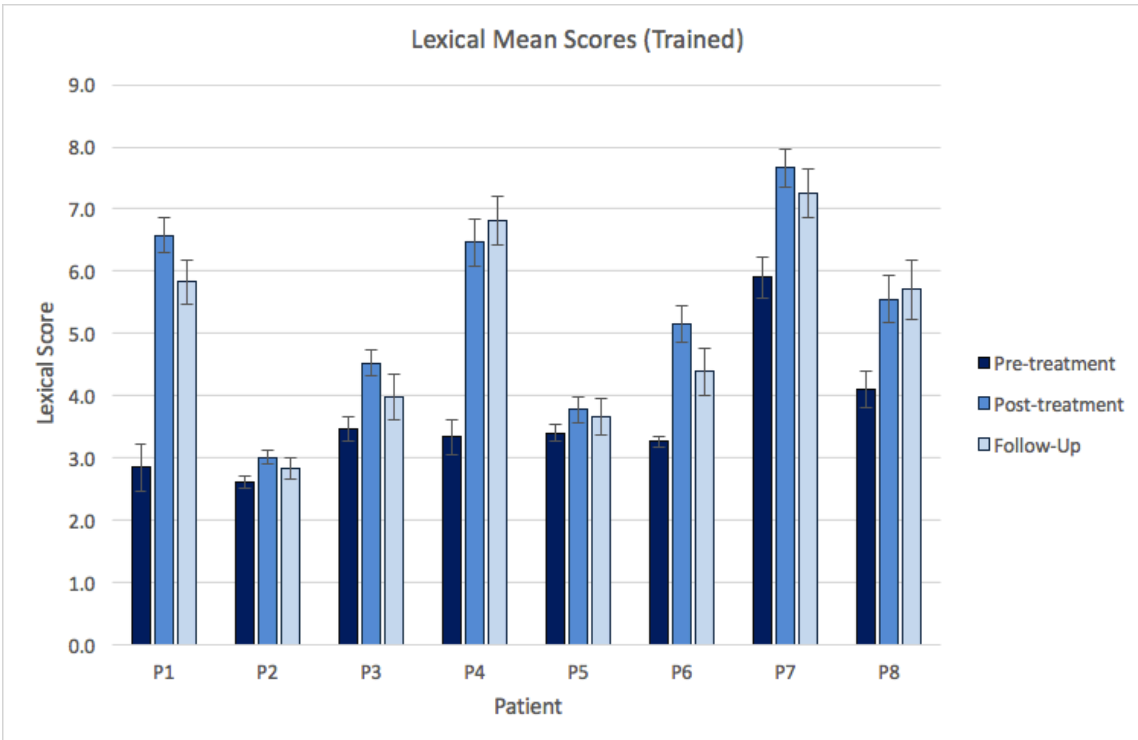


Figure 4. *Lexical mean scores for patients' trained modalities (pre-treatment, post-treatment, follow-up); P3 and P4 trained in reading*

Patients' single-word responses in the trained modality were scored and averaged to obtain one sublexical (Figure 3) and one lexical (Figure 4) value for pre-treatment as well as for post-treatment. Specifically, the two dependent measures were averaged lexical scores and averaged sublexical scores. The results from a repeated-measures MANOVA reveal a significant main effect of time point, ($F(2, 375) = 139.97$, $p < .001$, Wilks' $\lambda = 0.57$), as well as a significant interaction effect of time point and patient, ($F(14, 750) = 9.80$, $p < .001$, Wilks' $\lambda = 0.72$). Additionally, there is a significant between-subjects effect of patient, ($F(14, 750) = 25.99$, $p < .001$, Wilks' $\lambda = 0.45$). These results indicate that patients changed as a result of treatment, and that some patients benefitted more from treatment than others.

Univariate within-subject contrasts confirmed a significant main effect of time point for both lexical (*F(1,376)* = 242.72, *p* < .001) and sublexical scores (*F(1,376)* = 254.59. *p* < .001), as well as a significant interaction effect between treatment time point and patient for both scores (lexical *F(7,376)* = 14.88, sublexical *F(7,376)* = 14.49; *p* < .001).

Post hoc Bonferroni corrections compared amongst patients' pre- to post-treatment scores in both the lexical and sublexical parameters at a *p* value of < .05. Overall, the interaction effect of patient and time reveals that P1, P4, and P7 experienced significantly greater changes in both parameters as a result of treatment than the other subjects in this study, while P2 and P5 exhibited the smallest changes.

1.2 Trained Modality, Post-Treatment to Follow-Up

For the second time point comparison (post-treatment to follow-up; Figures 2 and 3), a repeated-measures MANOVA showed a main effect of time point was significant (*F(2, 375)* = 3.11, *p* < .05, Wilks' λ = 0.98), as did the between-subjects effect of patient, (*F(14, 750)* = 1.77, *p* < .05, Wilks' λ = .98). There was also a significant interaction effect of time point and patient (*F(14, 750)* = 26.69, *p* <.001, Wilks' λ = .45).

Within-subject contrasts revealed that the lexical and sublexical systems exhibit differential significant changes. There was a significant effect of time point in the lexical parameter from post to follow-up, (*F(1,375)* = 5.92, *p* < .05) as well as in the sublexical parameter (*F(1,375)* = 4.16, *p* < .05). However, when evaluating the interaction effect of time point and patient, only the sublexical system reached significance, (*F(7, 750)* = 2.74, *p* < .05). These results reveal that, even after treatment, there were significant or nearly significant changes in both the lexical and sublexical parameters. P1, P3, and P7 showed some decline in the sublexical parameters, but the majority of other patients maintained or slightly improved their performance.

In post hoc Bonferroni corrections, it was revealed that most patients remained stable from post-treatment to follow-up. However, P1, P6, and to a lesser degree P3 exhibited slight decreases in their scores.

2.1 Untrained Modality, Pre-Treatment to Post-Treatment

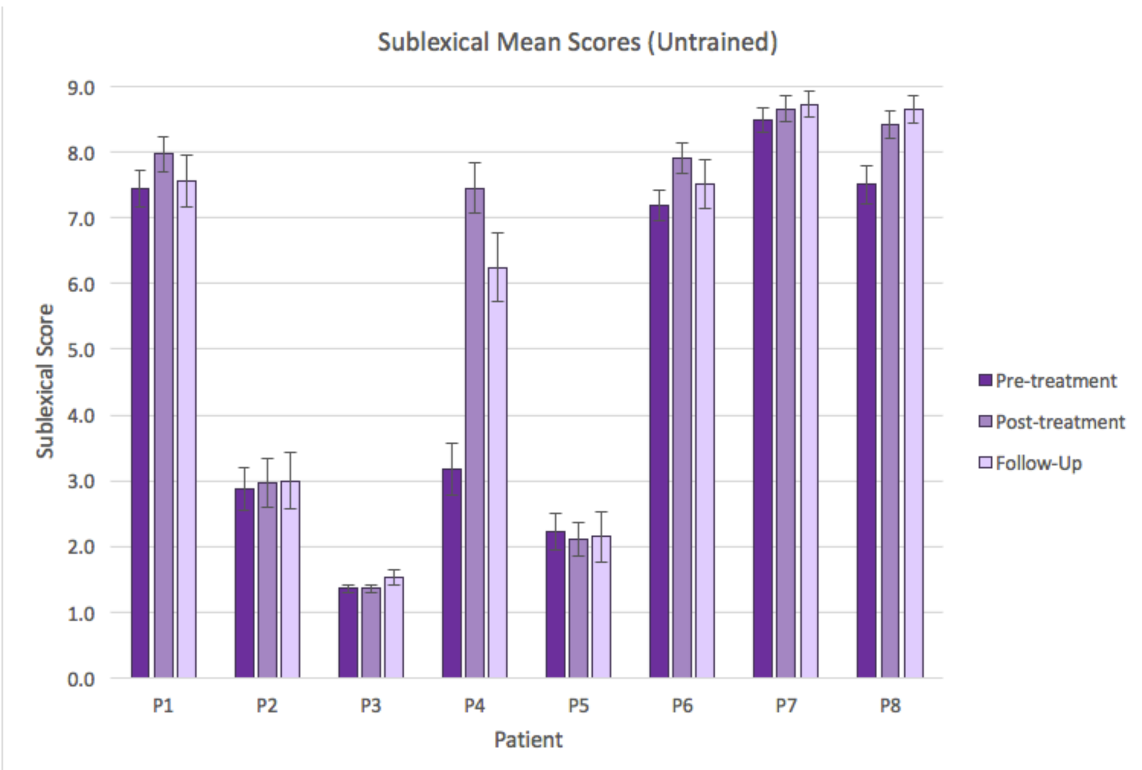


Figure 5. Sublexical mean scores for patients' untrained modalities (pre-treatment, post-treatment, follow-up); P3 and P4 untrained in writing

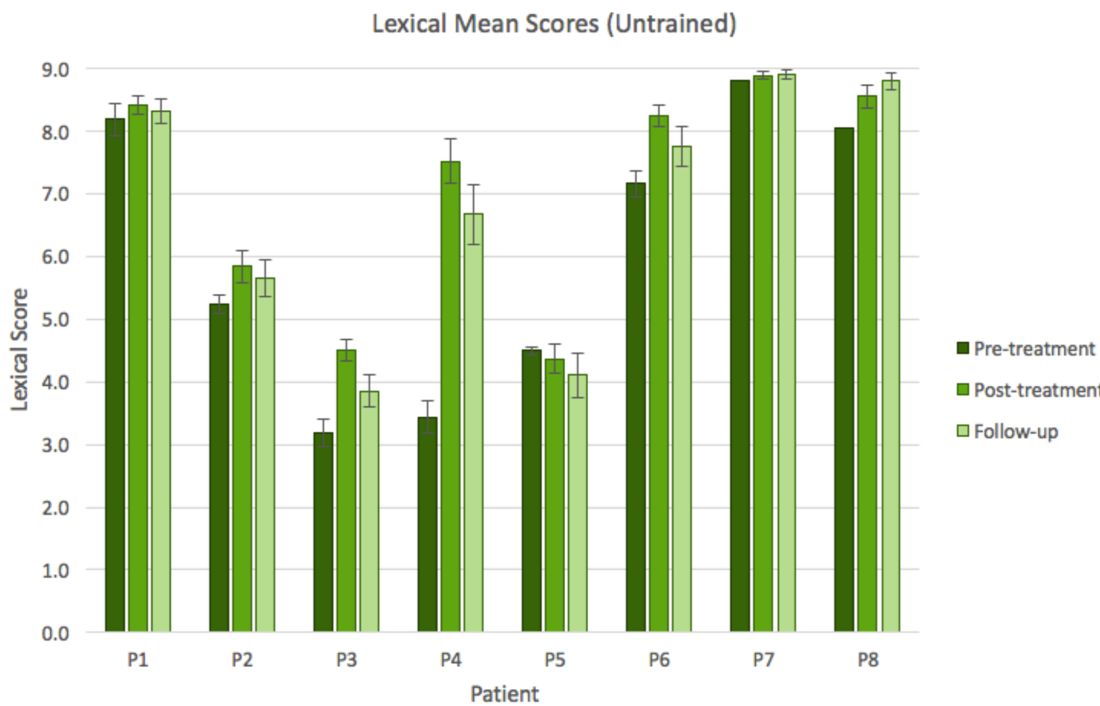


Figure 6. Lexical mean scores for patients' untrained modalities (pre-treatment, post-treatment, follow-up); P3 and P4 untrained in writing

In order to assess generalization to the patients' untrained modality (Figures 5 and 6), repeated measures MANOVA analyses were conducted on averaged values for pre-treatment and post-treatment as reported in section 1.1. The first rMANOVA analysis, from pre to post, revealed that a significant main effect of time point was present in the untrained modalities, ($F(2, 375) = 59.23$, p <0.001, Wilks' $\lambda = 0.76$), as well as a significant effect of patient ($F(14, 750) = 76.74$, p <0.001, Wilks' $\lambda = 0.17$) . Additionally, the interaction effect of time and patient reached significance, as well, ($F(14, 750) = 18.92$, p <0.001, Wilks' $\lambda = 0.55$).

Post hoc within-subject analyses also confirmed that all effects were still observed in both parameters. Specifically, there was a treatment effect for lexical, ($F(1,182)$ =

118.71, $p < .001$), and sublexical, ($F(1,130) = 70.32$), $p < .001$), as well as an interaction effect of time and patient, (lexical, $F = 30.29$, $p > .001$, and sublexical, $F = 28.36$, $p > .001$).

Overall, post hoc analysis revealed that some patients experienced greater changes as a result of treatment in the untrained modality than others. P4 exhibited the strongest changes, with P6 and P8 also showing slightly smaller two-dimensional gains. Additionally, P3 exhibited significantly greater gains in the lexical parameter than other patients, though this was not seen in his sublexical scores.

2.2 Untrained Modality, Post-Treatment to Follow-Up

Finally, the main effect of time point from post-treatment to follow-up in the untrained modality was revealed to be significant, ($F(2, 375) = 4.59$, $p < .05$, Wilks' $\lambda = 0.98$), as was the between-subjects effect of patient, ($F(14,750) = 51.33$, $p < .001$), Wilks' $\lambda =.17$). There was a significant interaction effect of time and patient ($F(14, 750) = 2.20$, $p = .007$, Wilks' $\lambda = 0.92$).

As was revealed in post hoc analyses of the trained modality, the two processing systems were differentially significant in this second time point contrast. While for lexical systems there was significant change as an effect of time ($F = 4.27$, $p < .05$), this was not true for the sublexical system ($F = 1.26$, $p = .26$). The time-patient interaction effect was significant in both the sublexical and lexical parameters, though stronger sublexically ($F = 2.99$, $p = .005$) than lexically ($F = 2.45$, $p = .018$). This indicates that different patients experience significantly different changes, and these are most substantially different in the sublexical parameter.

Finally, comparing individual patients' performance in post hoc analysis, the majority of relationships between patients were consistent in nature to the relationships they exhibited in the trained modality. While most patients remained stable from post-treatment to follow-up, P4 exhibited a decline in both dimensions, and P3 and P6 remained stable sublexical parameter but experienced a decline in the lexical parameter.

Research Question 2: Do these lexical and sublexical error scores capture discrete changes as a result of treatment in both systems from pre- to post-treatment to follow-up?

Heat maps were used to demonstrate the distribution of error types made by each participant. Pre-treatment and post-treatment graphs reflect summed scores from the 3 baselines (n = 144), while the follow-up graphs reflect a single probe set (n = 48). Consequently, the concentration of color represents the proportional frequency of response, regardless of the absolute numeric value. Therefore, coordinates that are most densely populated are the darkest red, with gradient transition to orange and yellow as the count decreases until reaching ultimately green, which represents a count of zero.

P1: 48 y.o. male, trained in writing

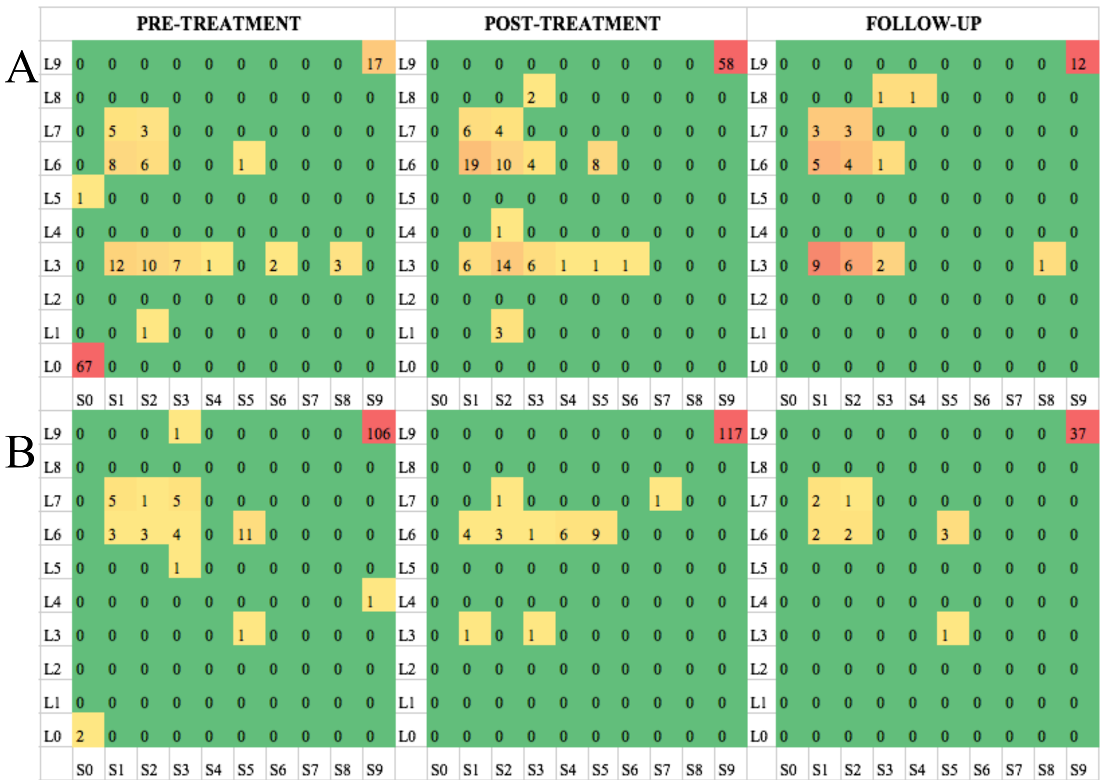


Figure 7. P1 heat maps reflecting the trained modality of writing (A) and untrained modality of reading (B)

Prior to treatment in the writing modality (Figure 7A), P1 most frequently earned a score of [S0, L0], indicating no response (n = 67). For those items which he did attempt, his errors were primarily nonwords with varying degrees of sublexical accuracy ([S1, L3], n = 12; [S2, L3], n = 10; [S3, L3], n = 7). He also produced many real words, though all but one of these productions shared less than 50% overlap with their targets ([S1–S2, L6–L7]). He successfully produced the target in 17/144 cases.

Following treatment, P1 demonstrated a drastic whole-word shift. This was particularly evident in the fact that he attempted all items (eliminating [S0, L0] scores) and in that he accurately produced 58 items, thus improving his pre-treatment score by

41. While [S1, L3] responses remained common, P1 demonstrated an increase in [S2, L3], indicating that his nonword productions more closely approximated the sublexical features of the target. He also exhibited a substantial increase in [S1, L6] and [S2, L6] responses, which indicates a lexical improvement with successful production of real words, though these remained maximally dissimilar to their targets sublexically. The map of P1's follow-up performance reflects that he maintained concentrated scores in accurate responses ([S9, L9]), sublexically dissimilar nonwords (S1–S2, L3]), and sublexically dissimilar unrelated real words ([S1–S2, L6]).

In the untrained modality of reading (Figure 7B), P1 demonstrated significantly higher accuracy than in writing at all time points, with nearly all [S9, L9] values. His most common error type was a real word response with a single sublexical error, particularly substitution ([S5, L6]). Following treatment, his whole word performance improved, as he eliminated [S0, L0]s and accurately produced the target on 11 additional occasions. His error types remained consistent with pre-treatment, as nearly all were lexically unrelated words with one sublexical deletion or substitution ([S4, L6] and [S5, L6]). Then during follow-up testing, although P2 maintained a high percentage of accurate responses, his errors reflect reduced sublexical accuracy in his real word productions, with ≤50% overlap in the majority of both related and unrelated words ([S1–S2, L6–L7])

P2: 58 y.o. male, trained in writing

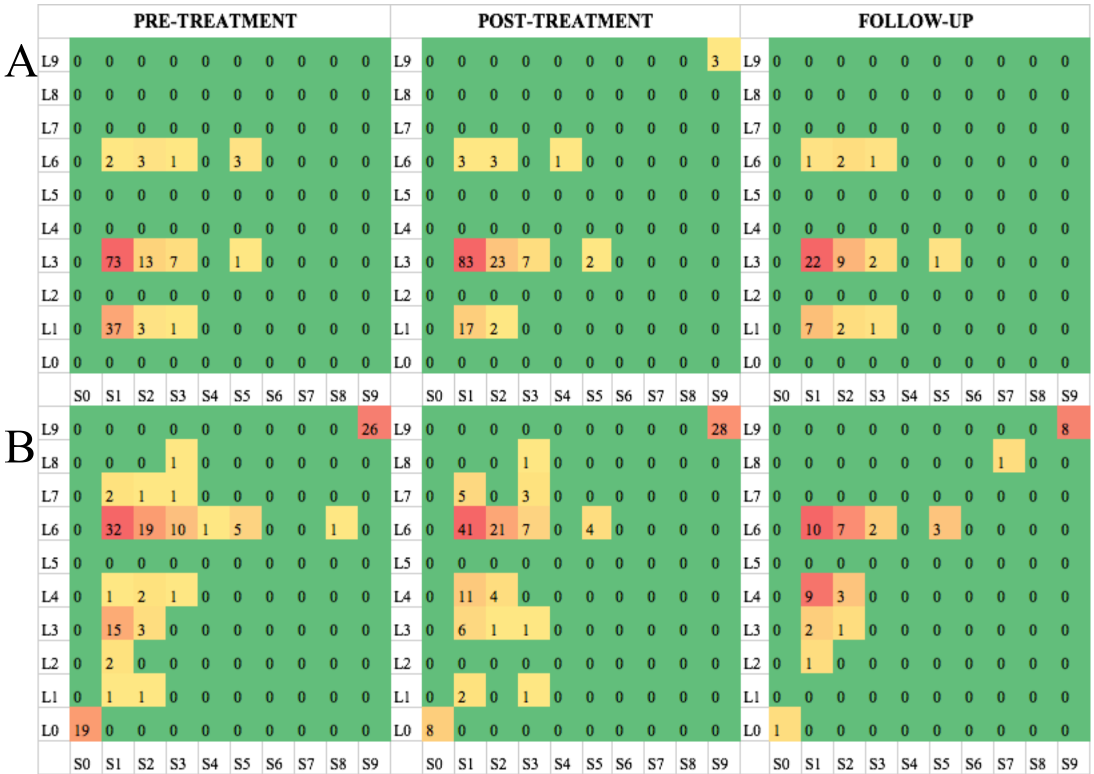


Figure 8. P2 heat maps reflecting the trained modality of writing (A) and untrained modality of reading (B)

P2 produced primarily nonwords with minimal sublexical accuracy prior to treatment in the written modality (Figure 8A) before treatment. His most frequent error types were [S1, L3] and [S1, L1] – nonwords and nonword perseverations with no more than 50% orthographic overlap with the target. He also produced a fair number of nonwords with correct orthographic length ([S2, L3], n =13), though these were less common. Finally, he did not produce any of the targets ([S9, L9]). Following treatment, this final feature changed, as he accurately wrote 3 items. Regarding his errors, he continued to predominantly produce nonwords with ≤50% overlap, however he reduced the number of perseverative errors by more than half ([S1, L1], n = 17, as compared to n

= 37 before treatment). He also demonstrated an increased number of nonwords with the correct number of graphemes ([S2, L3]), indicating an improvement in length of item responses. P2 made very limited gains in treatment, which did not reach the level of significance in binary scoring and which is reflected in these heat maps. His errors were consistent at follow-up testing, and his whole-word gains were not maintained.

P2 exhibited far more variable responses in the untrained modality of reading (Figure 8B). Before treatment, he most commonly produced unrelated words with less than 50% sublexical target overlap ([S1–S2, L6]). He accurately responded to 26 items ([S9, L9]), though he was noted to omit many responses as well ([S0, L0]). In post-treatment testing, his most common error remained the same ([S1–S2, L6]), though he also demonstrated a shift from sublexically dissimilar nonwords ([S1, L3]) to sublexically dissimilar perseverations of real words ([S1, L4]). Finally, he reduced the number of omissions by half ([S0, L0] n = 8). In follow-up testing, sublexically dissimilar perseverated words and unrelated words remained P2's dominant error types ([S1–S2, L4] and [S1–S2, L6]), though he also produced the target on 8 occasions.

P3: 72 y.o. male, trained in reading

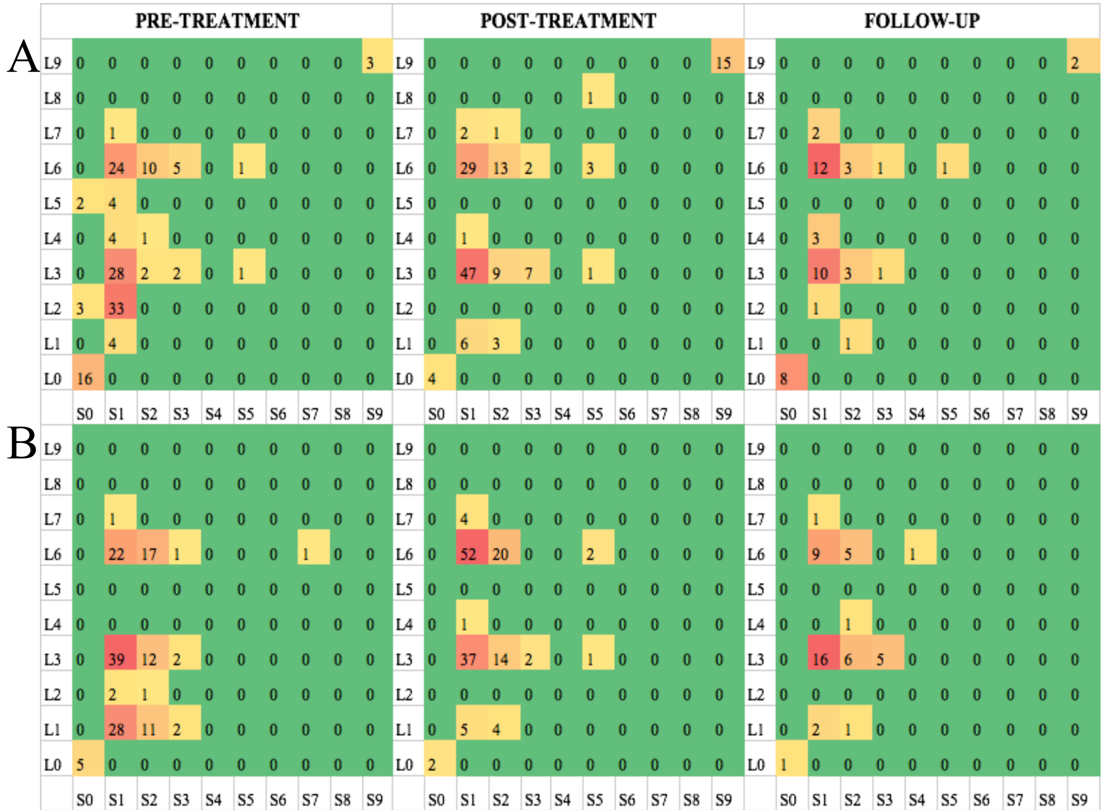


Figure 9. P3 heat maps reflecting the trained modality of reading (A) and untrained modality of writing (B)

P3 was one of the two participants to train in the reading modality, and he demonstrated highly variable responses across time points (Figure 9A). Before treatment, he omitted 16 items ([S0, L0]) and accurately produced 3 ([S9, L9]). His most common error types were all maximally dissimilar in the sublexical parameter with a range of lexical accuracy, from descriptions ([S1, L2]) to nonwords ([S1, L3]) and unrelated words ([S1, L6]). A treatment effect is evident in the shift from omissions to accurate responses ([S0, L0], n = 4; [S9, L9], n = 15), as well as in the types of responses P3 gave. Namely, he completely eliminated descriptions ([L2] and [L5]) in post-treatment. Instead, he produced mostly nonwords with ≤50% overlap ([S1–S2, L3]), as well as a slightly increased number of unrelated real words with minimal phonetic accuracy ([S1–S2, L6]).

At the final time point of follow-up testing, P3 presented with essentially the same types of errors. His whole word accuracy returned to a level resembling pre-treatment, producing a high percentage of omissions and few accurate targets. However, he successfully inhibited the descriptive responses he generated prior to treatment, and thus he predominantly read targets as nonwords and unrelated words with minimal sublexical accuracy ([S1, L3] and [S1, L6]).

Prior to beginning therapy, P3's productions in the untrained modality (Figure 9B) were primarily nonwords and nonword perseverations with ≤50% orthographic accuracy ([S1–S2, L3] and [S1–S2, L1]). He also produced many lexically unrelated words that were sublexically inaccurate ([S1–S2, L6]). Following the treatment phase, notable progress occurred in that he produced significantly fewer nonword perseverations. Instead he generated primarily unrelated real words with low sublexical accuracy ([S1–S2, L6]), as well as nonwords ([S1–S2, L3]).

P4: 44 y.o. male, trained in reading

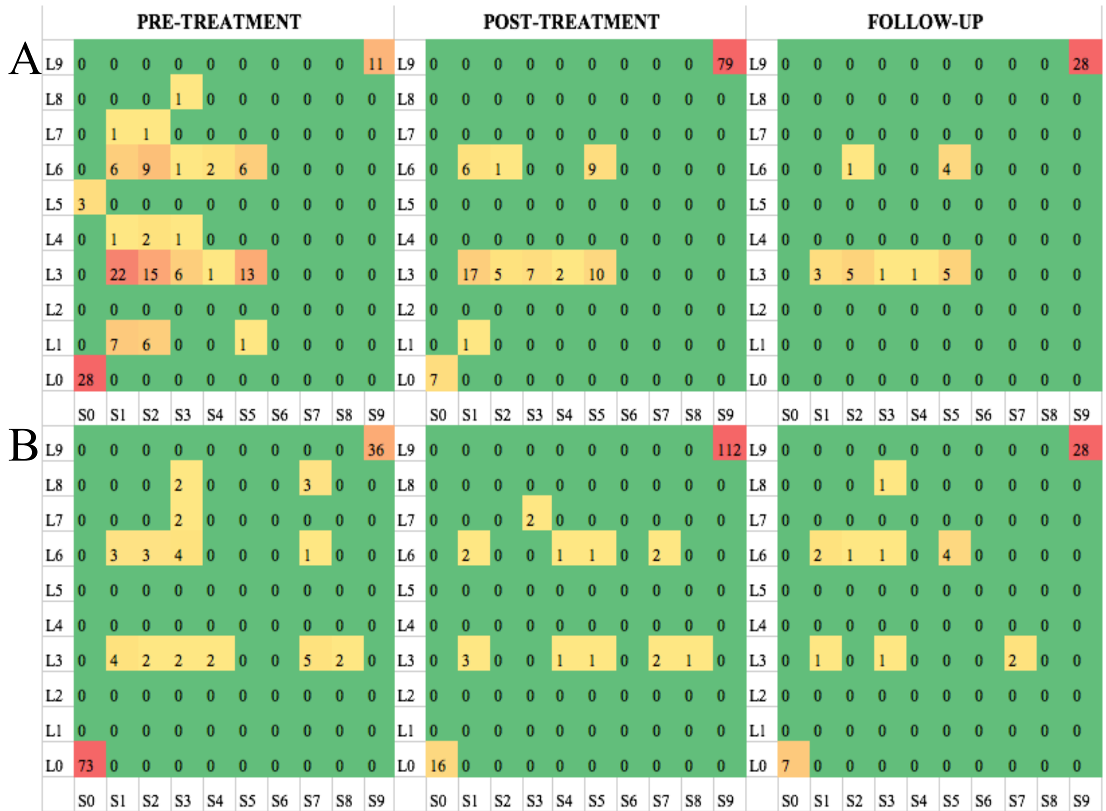


Figure 10. P4 heat maps reflecting the trained modality of reading (A) and untrained modality of writing (B)

In pre-treatment testing of reading (Figure 10A), P4 demonstrated a wide array of error types, with the most frequent score of [S0, L0] by a small margin. The other frequent error types included nonwords with minimal sublexical overlap ([S1, L3]), nonwords of accurate length ([S2, L3]), and nonwords with a single phonemic substitution ([S5, L3]).

At post-treatment, two main shifts are immediately apparent. First, P4 substantially increased whole-word accuracy, reading 79 items and omitting only 7. Second, while response scores were still concentrated in the minimally-overlapping nonword and single-substitution nonword coordinates, the patient consolidated his types of production to only nonwords ([L3]) and lexically unrelated real words ([L6]) of differential sublexical accuracy. Additionally, he increased the number of real words with a single-letter substitution ([S5, L6]). Follow-up testing revealed that accurate responses continued to be his most common production, while nonwords with various levels of sublexical overlap accounted for most of the rest ([S1–S5, L3]).

In the untrained modality of writing (Figure 10B), P4 demonstrated a strong "all or nothing" response tendency, with twice as many omissions as correct responses ([S0, L0], n = 73; [S9, L9], n = 36). Of his remaining responses, errors were broadly distributed across the systems. However, following the treatment period, he showed a drastic whole-word shift in accuracy, tripling the number of correct productions, considerably reducing the frequency of omissions, and nearly eliminating all other errors. These gains were maintained at follow-up testing, on the whole, although he exhibited a slight recurrence in the percentage of omissions.

P5: 66 y.o. male, trained in writing

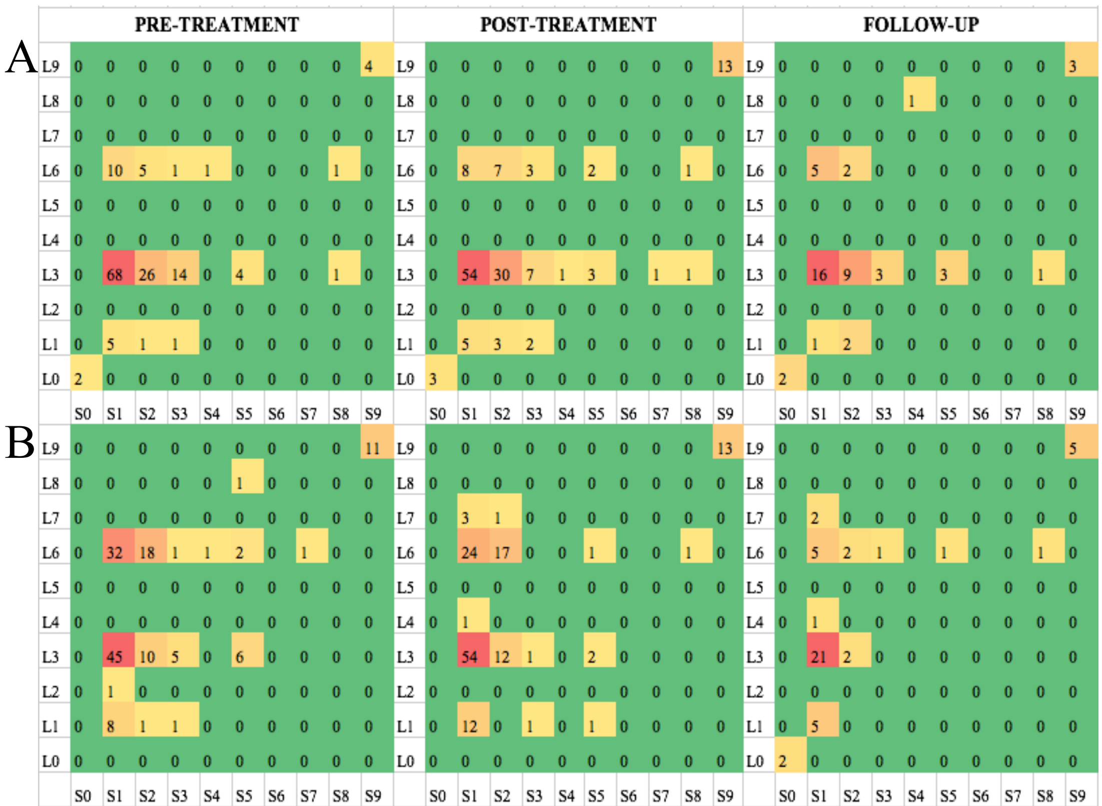


Figure 11. *P5 heat maps reflecting the trained modality of writing (A) and untrained modality of reading (B)*

Prior to treatment in writing (Figure 11A), P5 wrote a strong majority of nonword errors with minimal sublexical overlap ([S1, L3]). His remaining errors were also predominantly nonwords, some of which were accurate in length alone ([S2, L3]), and other nonwords that shared >50% of the targets' graphemes though they contained multiple errors ([S3, L3]). The same errors remained at the end of therapy, although he produced slightly more correct responses ([S9, L9], n = 13) and he demonstrated a shift from maximal sublexical difference in nonwords ([S1, L3]) to nonwords of the correct length ([S2, L3]). In follow-up testing, as well, nonwords with no sublexical accuracy and minimal sublexical accuracy were the most frequent ([S1, L3] and [S2, L3]). This participant, like P2, exhibited very weak effects in the treatment study, though they did reach significance.

Although P5 notably attempted all items in the untrained mode of reading (Figure 11B), and accurately produced 11 before the treatment phase, his most common productions were distant from the target. These included minimally overlapping nonwords ([S1, L3]) and minimally overlapping unrelated words ([S1, L6], [S2, L6]). Although he slightly increased the number of correct productions following therapy ([S9, L9], n = 13), he continued to produce sublexically inaccurate nonwords most often. In fact, he demonstrated a slight increase in nonwords with poor sublexical accuracy, as well as an increase in perseverated nonwords with poor sublexical accuracy ([S1, L1]). Finally, at the follow-up, P5 increased the number of non-responses ([S0, L0]), and maintained performance of nearly all nonwords with ≤50% overlap.

P6: 70 y.o. female, trained in writing

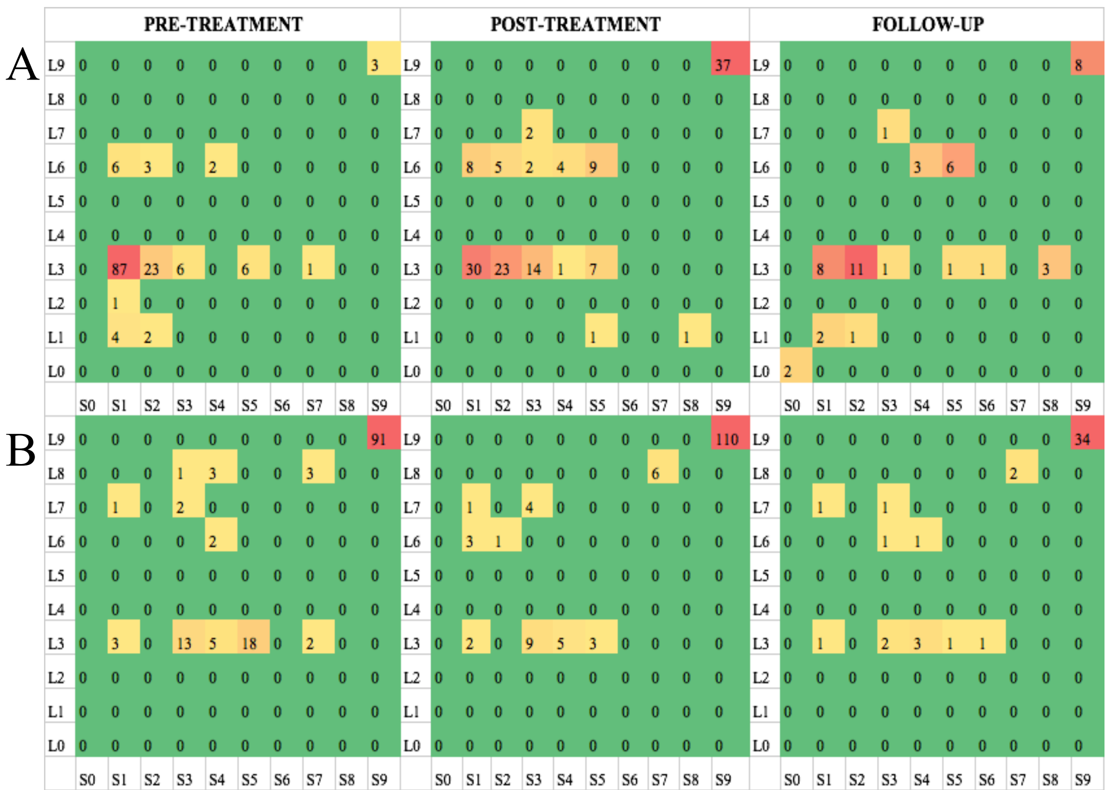


Figure 12. P6 heat maps reflecting the trained modality of writing (A) and untrained modality of reading (B)

P6 exhibited clear patterns of performance and error type in the trained modality of writing (Figure 12A), with nearly all productions scored as minimally sublexically similar nonwords prior to treatment ([S1, L3]). She also generated several nonwords with the correct number of graphemes, however considerably fewer than the former type of error ([S2, L3]).

In post-treatment testing, however, P6 demonstrated a significant improvement in overall accuracy, with 37 [S9, L9] productions as compared to only 3. The most common errors remained nonwords with ≤50% sublexical accuracy ([S1, L3] and [S2, L3]), though she also demonstrated an increase in nonwords with >50% sublexical overlap

([S3, L3]), and real word productions overall ([S1–S5, L6]). This treatment effect was maintained in follow-up testing, although her most common responses at this time point were nonwords of the correct length ([S2, L3]), closely followed by accurate productions and nonwords with maximal sublexical difference from the targets ([S1, L3]).

It is important to note that P6's heat maps in the untrained modality reflect target production within phrase reading (Figure 12B; see: Methods for full description). In the pre-treatment baselines, P6 accurately read 91 items. Her remaining responses were primarily classified as nonwords with >50% sublexical accuracy though more than one error ([S3, L3]) and nonwords with a single phonemic substitution ([S5, L3]). This was also the case following treatment, though she further improved her whole-word accuracy, as well as the number of targets plus a single phoneme resulting in a morphological error ([S7, L8]). At follow-up testing, as well, the patient primarily produced the target accurately, and the relatively few errors were mostly nonwords with >50% sublexical accuracy ([S3–S4, L3]).

P7: 76 y.o. male, trained in writing

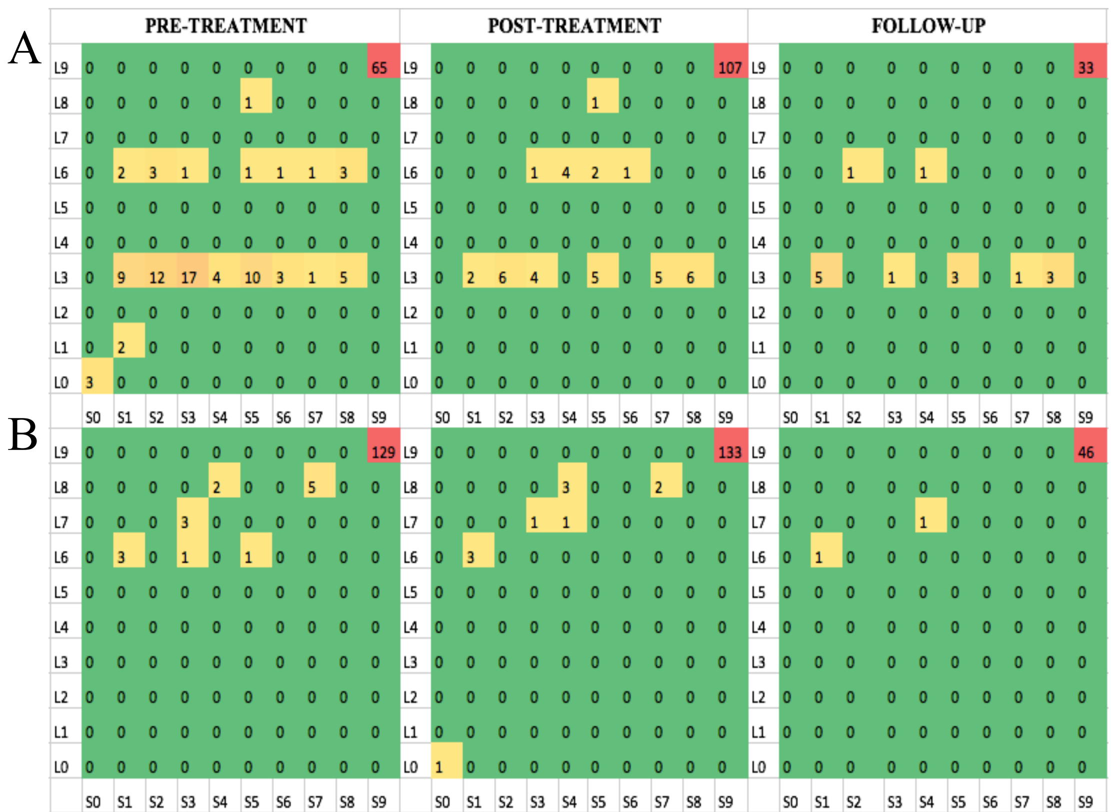


Figure 13. P7 heat maps reflecting the trained modality of writing (A) and untrained modality of reading (B)

Before initiating treatment in writing (Figure 13A), P7 produced 65 accurate responses in the trained modality. His errors were nearly all nonwords, and the most frequently occurring were nonwords with greater than half sublexical overlap with the target ([S3, L3]), nonwords with no more than half overlap though the correct number of graphemes ([S2, L3]), or nonwords with a single graphemic substitution ([S5, L3]). Once treatment was completed, his overall accuracy increased substantially ([S9, L9], n = 107). His errors, though relatively few, were nearly all nonwords with varying levels of sublexical accuracy ([S1–S2, L3], n = 21; [S3–S8, L3], n = 48). With 33 correct responses at follow-up, P7 appeared to maintain the gains he made in treatment. His errors remained consistent as well, with mostly nonword mistakes and most of those being maximally distant from the target sublexically ([S1, L3]).

Reading probes for P7 (Figure 13B), just as for P6 and P8, were comprised of short phrases containing the target. Despite the added linguistic complexity as compared to the standard single-word probes, P7 still exhibited a near-ceiling effect, with 129 accurate responses prior to initiating treatment. The most common error among the remaining responses was a single sublexical addition resulting in a morphological variation of the target ([S7, L8]). In post-treatment probes, he achieved an even higher overall accuracy (n = 133), and errors remained dispersed in the real word categories with varying degrees of sublexical accuracy (mostly [S1, L6], [S4, L8], and [S7, L8]). This mirrored his follow-up presentation.

P8: 66 y.o. female, trained in writing

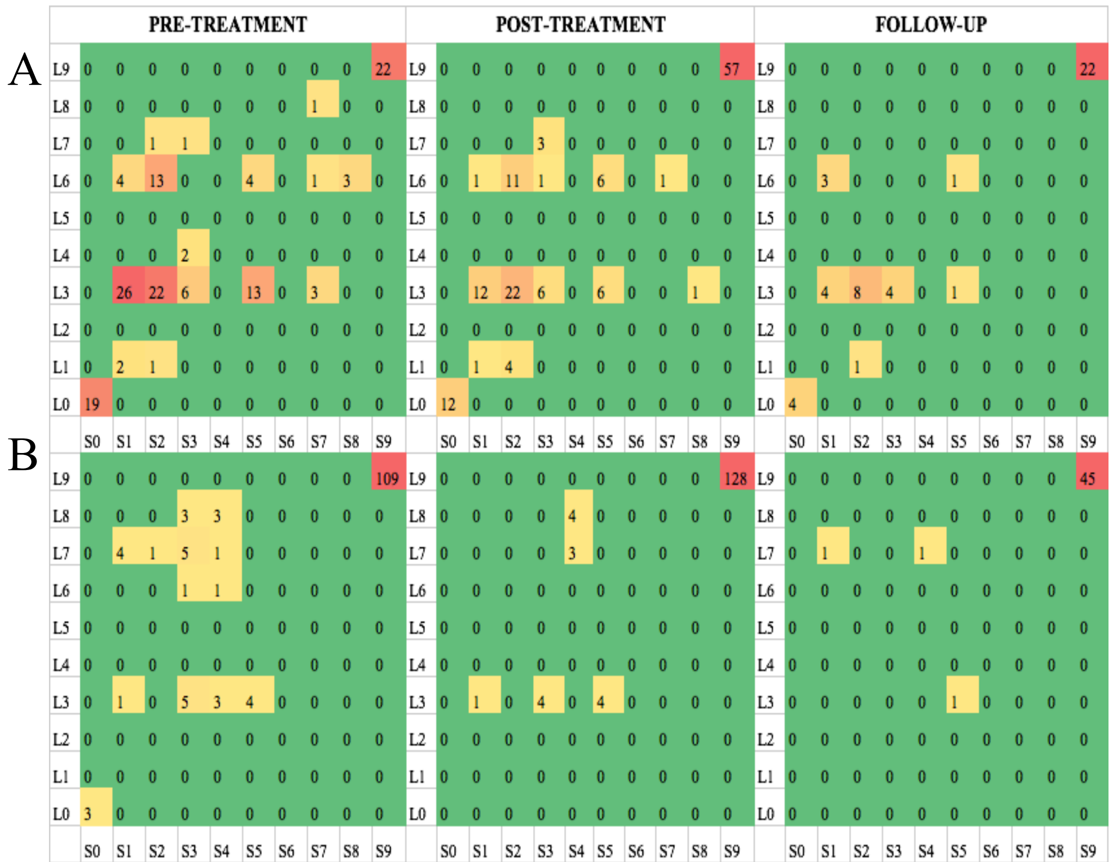


Figure 14. P8 heat maps reflecting the trained modality of writing (A) and untrained modality of reading (B)

Prior to the treatment phase in writing (Figure 14A), P8 produced a wide variety of sublexical and lexical responses. She earned virtually equal counts of accurate responses and null responses, as well as numbers of nonwords of the correct length and ≤50% overlap ([S2, L3]). Besides these error types, she also frequently produced nonwords with a single substitution ([S5, L3]) and real words with accurate graphemic length ([S2, L6]). She then produced a significant increase in accurate responses following treatment ([S9, L9], n = 57) as well as a decrease in omissions and in maximally sublexically different nonwords ([S1, L3]). The number of [S2, L3] responses, however, remained constant. In follow-up testing, as well, nonwords of the correct length with ≤50% overlap, ([S2, L3]) remained P8's second most common response after accurate production of the targets.

In the untrained modality of reading (Figure 14B), P8's scores (which reflect phrase and compound word reading, as with P6 and P7) demonstrated a high level of accuracy with 106 words produced correctly only 3 omitted. Her errors were primarily semantically or morphologically lexically related words, with ≤50% overlap though with multiple sublexical errors ([S3, L7–L8]). After the therapy program ended, P8 attempted all items in post-treatment and accurately produced 128. The remaining errors were primarily nonword responses with >50% overlap ([S3–S5, L3]). And, perhaps most remarkably, P8 produced 45 of the 48 targets accurately at follow-up, and the single error lay in a sublexical deletion resulting in an unrelated real word ([S4, L6]).

Research Question 3: Does this coding system demonstrate reliability and objectivity?

In order to answer Research Question 3, two undergraduate computer science students at Boston University generated computer codes in Python 2.7.10 (Van Rossem, 2015) for the lexical and sublexical systems. These codes were written in Python 2.7.10 to perform the same analysis as manual scoring. All 5,376 responses from the study were scored using these automated systems. In order to do this, a complete, ordered list of target items, followed by a list of corresponding responses was run through the sublexical and lexical programs developed for this study, utilizing the same hierarchical rules in both parameters. The results of these automated tests were then compared to the manually generated scores. In the lexical system, discrepancies were identified between computer and clinician scores with an overall agreement in 87% of items. However, many of these were due to an automated program error in calculating nonword perseverations. When this was controlled for, agreement reached 94%. In this controlled set, the scores identifying a clinician error were corrected in the patient data, with corrected data reported above. All remaining disagreements were between related/unrelated words and real word/nonword distinctions; these were the result of the computer dictionary over-identifying the number of lexical-semantic associates and identifying words that the clinician did not know. Further discussion of these errors follows.

In the sublexical system, a substantially greater number of discrepancies were found, amounting to an agreement in only 79% of items. This relatively low overlap is primarily due to the fact that item length in the reading modality was manually scored in phonemes rather than graphemes, while in the computerized version, due to the limitations of computer code parsing speech sounds, grapheme length was scored. Ongoing work to resolve this barrier is underway.

DISCUSSION

The aim of this study was to evaluate the quantitative and qualitative utility of a novel scoring system for print processing accuracy. Data from a recent reading and writing treatment study were used to investigate how well this two-dimensional scoring system captured change as a result of treatment; what the nature of lexical and sublexical change was; and whether or not the scoring demonstrated reliability and objectivity when compared to a computerized version of the same system. Overall, the main findings of this study were: a) significant effects of treatment and b) significant differences between patients, which is evident in their statistical change scores and in the error evolution shown by individual heat maps. Multiple patterns of change were observed in the types of error progressions made by individual subjects and by the group as a whole, and manually generated scores were in strong agreement with automated scores, indicating reliability and objectivity of the two scoring hierarchies. Each of the research questions and their results are discussed below in greater detail.

Results from repeated-measures MANOVA and post hoc analyses of averaged patient data revealed a significant treatment effect in the trained modality for both sublexical and lexical parameters, confirming hypothesis 1. These findings indicate that the subjects of this study made incremental progress towards accuracy in both dimensions as a result of treatment. That is, on the error responses for which they did not earn credit on a traditional binary correct/incorrect system, patients' productions more closely approximated targets both in lexical and sublexical qualities by the end of treatment than they did prior. Patients were also found to differ significantly from one another in the degree of treatment effects, which is reflected both in the magnitude of the change in their scores and in the spread of the data in each subject's heat maps. These results support use of this scoring system to capture treatment effects, as it captures whole-word shifts in accuracy (binary scoring changes) and those developments that occur below the whole-word level, as well as differences between individuals on the same scale.

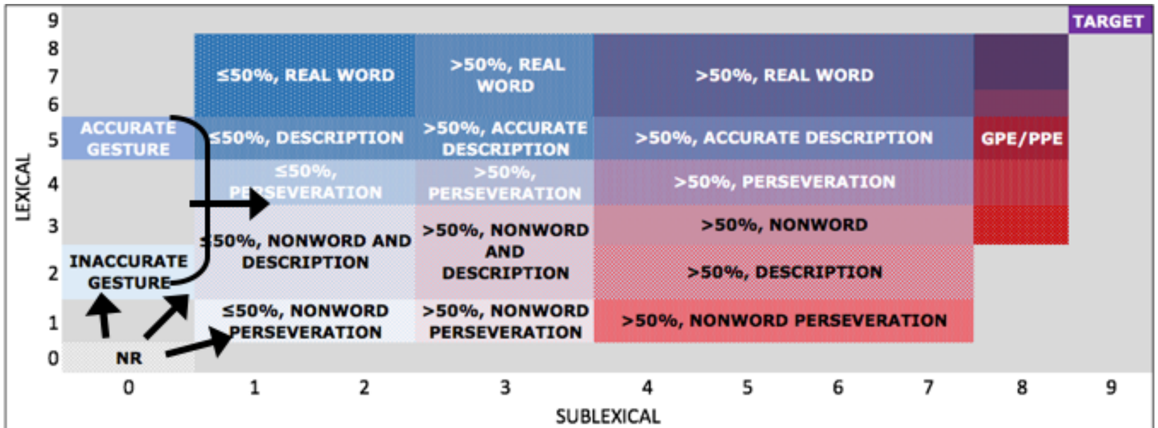
Follow-up analyses revealed that, within the overall significant findings, certain patients nevertheless demonstrated less robust differences than others. P2 and P5 in particular, who exhibited the least overall improvement in binary scoring, also showed the weakest changes lexically and sublexically. Furthermore, they differed least from one another in post hoc analyses. P3's treatment effect, although more statistically significant than P2 and P5s', was largely attributable to his extremely low initial performance. However, all three subjects performed and improved quite distinctly. P2 decreased his number of perseverative errors, and increased word-length accuracy considerably. P3 successfully stopped producing gestures and descriptions, instead attempting GPC in every trial, and he increased the number of real word responses in the process. Such changes reflect that his gains were greatest in the lexical system. Finally, P5 substantially reduced his perseverative errors, progressing lexically to real words and sublexically to items with greater overlap. These changes not only reveal that these individuals did make progress, but they also can then be used to guide intervention by targeting the weaker dimension with support to the stronger dimension.

In order to assess for generalization of treatment gains across print processing systems, treatment effects were also measured for patients' in the untrained modality. A significant main effect was once again found for both lexical and sublexical systems. This indicates that, not only did the group exhibit transfer of treatment benefits from reading to writing and vice versa, but also that this improvement occurred in both access routes for these two modalities. Furthermore, within-subject contrasts confirmed significance for the interaction effect of treatment and patient, just as it did in the trained modality. Therefore, the improvements that patients made as a result of treatment generalized to the untrained modality, and these improvements were significantly different in quality and quantity from one patient to the next.

The second time point comparison, from immediately post-treatment to a 6- to 8-week follow-up, was used to evaluate maintenance. Overall, only a few patients showed a decline in performance after treatment ended, while others maintained treatment effects and some even improved slightly. Analysis of post-treatment to follow-up effects in the untrained modality did not reveal this decline, implying that the generalized gains were maintained across the patient group once treatment ended. It should be noted that this lack of significance in the untrained modality may also be attributable to the fact that many patients were stronger in this modality at baseline.

The crux of this study, however, lies beyond even this level of analysis. Each patient in this investigation demonstrated a unique profile of sublexical and lexical strengths and weaknesses, which shifted as a result of treatment and which are revealed by their heat maps. Furthermore, these shifts are observable at the group level as well. Some subjects made particular progress in the sublexical accuracy of their responses (e.g. P3, P4), and others advanced most notably in the lexical parameter (e.g. P1 and P5), but most shifts occurred bimodally, as seen across patients, though in P4 and P7 especially.

By looking at patterns across the group as a whole, it is possible to detect integrated patterns of change in both dimensions. In order to facilitate the identification and intelligibility of these trends, the schematic representation of the two-dimensional error graph (presented above as Figure 2 in Statistical Analysis) is used. Patterns of error evolution are then indicated with arrows denoting the main shifts in each section.

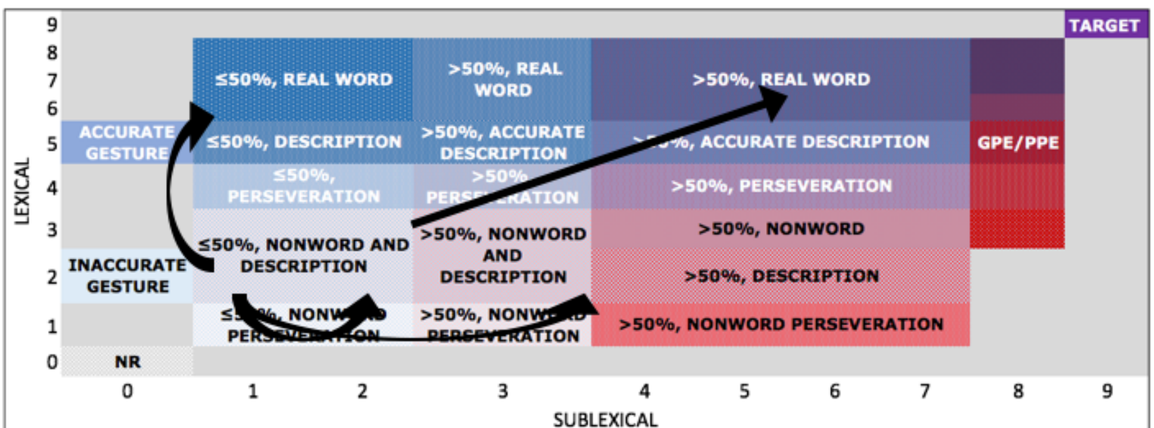


NR = no response; ≤50%/>50% indicate graphemic/phonemic overlap between the target and response; GPE/PPE = graphemically/phonemically plausible error

Figure 15. Evolution from inappropriate response types *[S0, L0], [S0, L2], [S0, L5]*

Four directional patterns of error evolution are observable in these data. The first is an evolution away from inappropriate response types, namely omissions, descriptions, and gestures (Figure 15). Omissions were quite common errors in the pre-treatment baselines (represented in the bottom left corner), especially for subjects P1, P3, P4, and P8. However, by the end of the intervention phase, these subjects attempted nearly all responses in both the trained and untrained modalities. Additionally, P2 made this shift in the untrained modality, where most of his omissions occurred. This change signifies a very meaningful qualitative shift that is not captured in traditional binary scoring. Although the patients continued to earn an "incorrect" score, this scenario involves an attempt at grapheme-to-phoneme/phoneme-to-grapheme conversion where no attempt had been made before.

Inappropriate responses also include descriptions and gestures (indicated as the far left column). Only two patients (P3 and P4) attempted descriptions and gestures, with varying degrees of lexical accuracy ([S0, L2/L5]). By the point of post-treatment, however, both were observed to eliminate this type of response in favor of single-unit attempts (e.g. words, nonwords). This change is noteworthy not only because it demonstrates adherence to the goal of grapheme-to-phoneme conversion (GPC because both P3 and P4 were trained in reading), but also because it indicates the introduction of sublexical features in the response. Gestures and descriptions are necessarily limited in sublexical accuracy, so even an increase towards a maximally dissimilar response (e.g. nonword productions with ≤50% overlap, [S1, L3]) marks stronger sublexical activation than [L2]s and [L5]s.

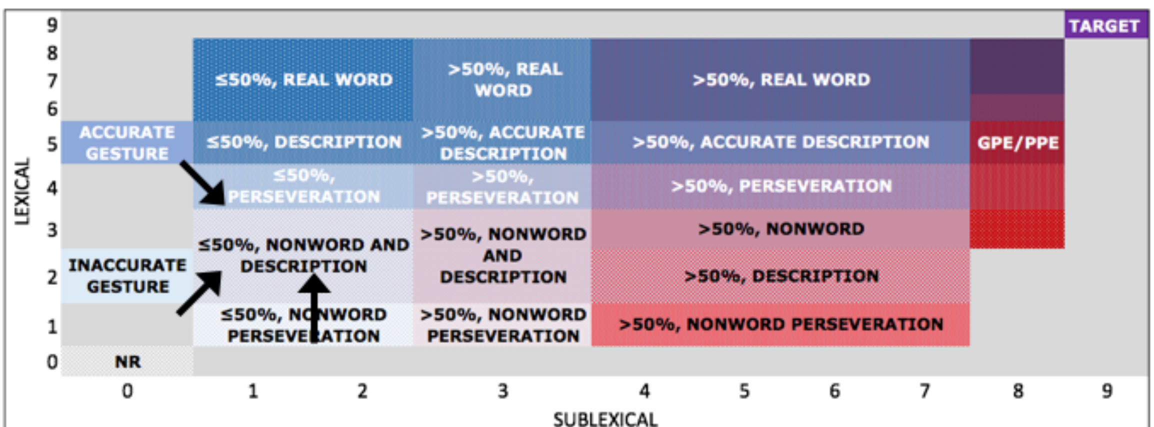


NR = no response; ≤50%/>50% indicate graphemic/phonemic overlap between the target and response; GPE/PPE = graphemically/phonemically plausible error

Figure 16. Evolution from nonword with minimal sublexical overlap [S1, L3]

The second pattern of evolution is evident in the shift from nonword productions with maximal sublexical dissimilarity ([S1, L3]; Figure 16), which is by far the most common pre-treatment error type across all patients and modalities. This evolution is a stronger improvement than the pattern previously described, because rather than establishing the appropriate activity, the participants refine it. P1 and P2 (as well as P4 in a small number of opportunities) resolved the [S1, L3] error with a slight sublexical shift, as they continued to produce nonwords following treatment, but these nonwords instead had accurate length ([S2, L3]). In other cases, the sublexical treatment effect was larger. P4, P6, P7, and P8 all shifted productions from the ≤50% overlap category to >50% ([S1, L3] to [S3–S5, L3]). P1 and P6 also shifted in the lexical parameter, from nonwords to real words ([S1, L3] to [S1, L6]). Finally, P4 improved both sublexically and lexically from [S1, L3] productions. His most common productions by the point of post-treatment testing were real words with a single sublexical substitution ([S5, L6]).

The third noteworthy pattern of evolution (made most often by subjects P2 and P3) is actually seen when patients begin producing the error that many other participants stopped producing: namely, sublexically dissimilar nonwords ([S1, L3]; Figure 17).

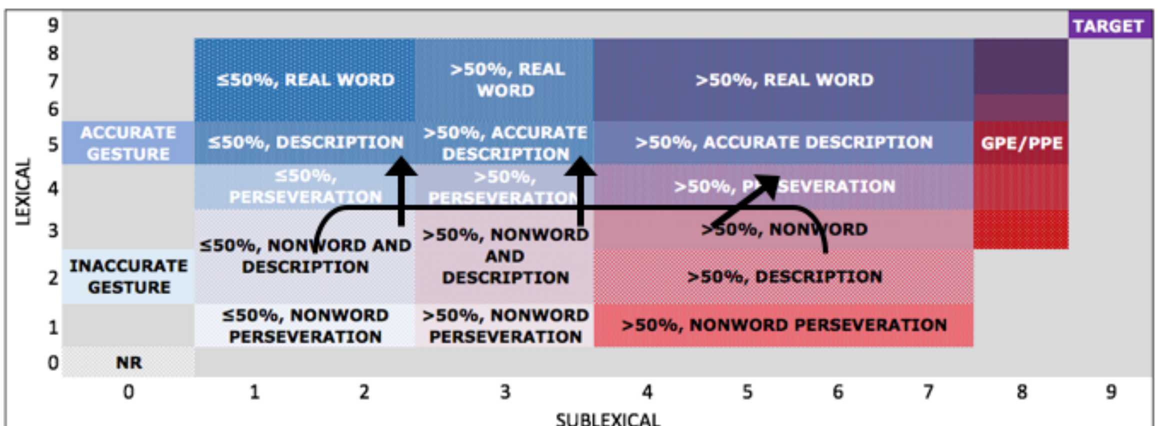


NR = no response; ≤50%/>50% indicate graphemic/phonemic overlap between the target and response; GPE/PPE = graphemically/phonemically plausible error

Figure 17. Evolution toward nonword with minimal sublexical overlap [S1, L3]

This change in performance, although it does not reflect radical movement in either parameter, is a critical shift nonetheless. As in the first pattern described, this represents the fact that patients have begun attempting grapheme-to-phoneme and phoneme-to-grapheme conversions when previously they had not, and furthermore that these attempts are specific to the stimulus. In other words, rather than adhering to a different processing system (i.e. gesture) or repeating a previous item, these individuals produced novel

GPCs/PGCs in each trial. Therefore, this shift toward is one of improved task adherence and increased competence in the activities of reading and writing. P2 produced a high volume of nonword perseverations with maximally dissimilar sublexical properties in pre-treatment ([S1, L1]), which resolved to novel nonwords; P3 shifted from descriptions and gestures, as described above, ultimately increasing his nonword count by the end ([S1, L3]). It is important to note that these were two of the three individuals who demonstrated the greatest overall deficits in both reading and writing at baseline. Although this fact may have limited their total gains, it is the reason it was possible for P2 and P3 to achieve an error type at the end of treatment that many higher-level subjects produced at baseline.



NR = no response; ≤50%/>50% indicate graphemic/phonemic overlap between the target and response; GPE/PPE = graphemically/phonemically plausible error

Figure 18. Evolution towards real words [L6-L9]

The final, most substantial evolution was toward real word productions (Figure 18). This pattern reflects the strongest treatment effect, as it demonstrates considerable improvement lexically, and almost always sublexically as well. P1, P3, P4, and P6 all exhibited this change primarily in the lexical system, by decreasing the relative number of nonwords as compared to real words. And more substantial gains were exhibited by several patients who increased both the sublexical and lexical accuracy. P4 was the only subject to significantly improve his response type to real words with a single sublexical substitution ([S5, L6]), reducing his numbers of nonwords and descriptions with ≤50% overlap. Similarly, P6, P7, and P8 made improvements in both parameters as well. In particular, these three individuals substantially increased the percentage of morphological error + single sublexical addition, thus achieving nearly the highest scores in both parameters ([S7, L8]).

In order to evaluate the reliability and objectivity of this system, manually-generated scores were compared to the scores produced by a computerized system for all patients' responses. The two methods were in high agreement, and very few errors were found in the manual system. The mistakes that were revealed were typically discrepancies between "related" and "unrelated" lexical scores. These were due either to over-identification of [L7] on the part of the computer (e.g. identifying proper nouns which are not counted in this system), or under-identification on the part of the clinician (e.g. limits of clinician vocabulary).

Other discrepancies were also originally found between manually and automatically scored GPEs/PPEs. This resolved in the computer system by outputting two scores for potential plausible errors—one S8, one other for the sublexical error(s) that caused it. It is then up to the clinician to decide which score is most appropriate.

Limitations

The primary limitation of this project lies in its basis upon a single treatment study, essentially conducted with a convenience sample. Although the scoring hierarchies were derived from collection of clinical and theoretical frameworks of alexia and agraphia, the development and testing were all conducted in the context of a single clinical approach with a limited number of patients. For this reason, there may be response types and patterns that are not accounted for here. Alternatively, there may be specifications that these systems overemphasize. However, the subjects here represented a wide range of profiles (e.g. aphasia/alexia/agraphia type, severity, time-post-onset) and the scoring systems were used to evaluate both reading and writing modalities in an attempt to account for the small number of participants.

Another limitation is that the scores only apply to single-word responses. As such, analysis of higher order deficits can not be readily executed with the present system. This challenge can be circumvented, as it was in this project, by scoring only the most salient word in a given response; however, this is not always a reasonable solution. Finally, manual scoring presents the additional barrier of low efficiency. It is somewhat time consuming, which could be especially challenging for clinicians in treatment setting. Familiarity facilitates the process, as users may become acquainted enough with the system that they need not refer to external lists to generate a [S#, L#] score. However, the value of this system lies in the richness of its qualitative analysis. Other scoring systems, both binary and slightly more complex one-dimensional systems (e.g. % overlap, semantic relatedness, etc.), do not provide the same information that this two-dimensional hierarchy affords. Additionally, the automated system, which is in the process of being made universally available in a web-based format, will eliminate this limitation altogether.

Future Directions

This project offers a tool that may be used by clinicians and researchers to assess the two-dimensional accuracy of patient responses to print stimuli. The first and foremost future direction of this project is that the automated coding software that was used to measure reliability will be the default scoring mechanism. Work is currently underway to make this system free and accessible online to all clinicians and researchers. Other future studies may endeavor to use the novel scoring system presented here to assess what treatment approaches and elements are most effective for particular clients (e.g. degree of impairment; profiles of aphasia, alexia/ agraphia). For example, researchers may compare profiles of alexia/agraphia (i.e. surface, phonological, deep) in various treatment programs in order to analyze the types of changes that occur in the sublexical and lexical systems. Inversely, different approaches to intervention may be evaluated across patient types to assess the resultant sublexical and lexical shifts. Research may also be undertaken to evaluate the effectiveness of these two hierarchies in capturing non-print errors, specifically naming as this is most often the focus of treatment for individuals with aphasia (Beeson & Rapcsak, 2002).

CONCLUSION

This project has revealed that significant changes in print processing below the level of whole-word accuracy may be effectively captured via the scoring hierarchies presented here. Individual patient graphs tell the story of the lexical and sublexical responsiveness to treatment, and they reflect the discrete improvements that binary scoring fails to show. Collective analyses showed how these patients performed as a group, elucidating patterns and trends across the data. Furthermore, this study demonstrated that, while the two dimensions vary differentially, they do not vary independently. Thus, it is appropriate to evaluate responses, and particularly changing responses, within one integrated system

APPENDIXES

1A. Reading Treatment

	Clinician	**Patient**
*1	*Presents written target word and nonword foil.*	Identifies real word.
2	*Presents written target word.*	Attempts to read target aloud.
3	*Verbally presents target.*	Repeats target.
*4	*Presents written target and 3 pictures.*	Matches written word to picture.
5	*Facilitates Semantic Feature Analysis (based on picture).*	Analyzes features presented by clinician.
6	*Presents letter tiles, including distractors.*	Produces phoneme associated with each tile.
7	*Verbally presents target word; cues patient to spell target using letter tiles by presenting each phoneme ("which letter makes the __ sound?").*	Spells target via letter tiles with field of distractors.
8	*Prompts patient to generate phonemes for all graphemes ("what sound does this letter make?").*	Produces phonemes.
9	*Prompts patient to identify graphemes for all phonemes ("which letter makes the sound __?).*	Identifies graphemes.
10	*Covers tiles and presents written target.*	Attempts to read target aloud.
11	*If patient fails to read the target, verbally presents target.*	Repeats target.
12	*Prompts patient to count aloud 1–10.*	Counts 1–10.
13	*Prompts patient to read target word.*	Reads target word.
*Steps 1 and 4 were dropped for patients 5–8		

1B. Writing Treatment

	Clinician	Patient
*1	*Verbally presents target word and nonword foil.*	Identifies real word.
2	*Verbally presents target word.*	Attempts to write the target.
3	*Presents written word.*	Copies written target.
*4	*Verbally presents target and 3 pictures.*	Identifies corresponding picture.
5	*Facilitates Semantic Feature Analysis (based on picture).*	Analyzes features presented by clinician.
6	*Presents letter tiles, including distractors.*	Produces the phoneme associated with each tile.
7	*Verbally presents target word; cues patient to spell target using letter tiles by presenting each phoneme ("which letter makes the /p/ sound?").*	Spells target via letter tiles with field of distractors.
8	*Prompts patient to generate phonemes for all graphemes ("what sound does this letter make?")*	Produces phonemes.
9	*Prompts patient to identify graphemes for all phonemes ("which letter makes the sound /b/?).*	Identifies graphemes.
10	*Covers written examples and tiles and presents target word verbally.*	Attempts to write the target word.
11	*If patient fails to write the target, presents written target.*	Copies written target.
12	*Prompts patient to count aloud 1–10.*	Counts 1–10.
13	*Prompts patient to write target word.*	Writes target word.
*Steps 1 and 4 were dropped for patients 5–8		

References

Badecker, W., (1995) Representational properties common to phonological and orthographic output systems. *Lingua, 99,* 55–83.

Beauvois, M.F., & Dérouesné, J. (1981). Lexical or orthographic agraphia. *Brain*, *104*, 21–49.

Beeson, P.M. (1999). Treating acquired writing impairment: strengthening graphemic representations. *Aphasiology*, *13*, 767–785. doi:10.1080/026870399401867

Beeson, P.M., Magloire, J.G., & Robey, R. (2005). Letter-by-letter reading: Natural recovery and response to treatment. *Behavioural Neurology, 16,* 191–202.

Beeson, P.M. & Rapcsak, S.Z. (2002). Clinical Diagnosis and Treatment of Spelling Disorders. In A. E. Hillis (Ed.), *The Handbook of Adult Language Disorders*, 101–120. New York, NY: Psychology Press.

Beeson, P.M., Rising, K., Kim, E.S, & Rapcsak, S.Z. (2008). A novel method for examining response to spelling treatment. *Aphasiology, 22*(7–8), 707–717. doi:10.1080/02687030701800826

Beeson, P.M., Rising, K., Kim, E.S, & Rapcsak, S.Z. (2010). A treatment sequence for phonological alexia/agraphia. *Journal of Speech, Language, and Hearing Research, 53,* 450–468.

Beeson, P.M., & Robey, R.R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review, 16*, 161–169. doi:10.1007s11065-006-9013-7

Bird, S., Loper, E., & Klein, E. (2009), *Natural Language Processing with Python* [Computer Software]. O'Reilly Media Inc. Retrieved from http://www.nltk.org/

Bose, A. (2013). Phonological therapy in jargon aphasia: Effects on naming and neologisms. *International Journal of Language & Communication Disorders, 48*(5), 582–595. doi:10.1111/1460-6984.12038

Bose, A., Raza, O., & Buchanan, L. (2007). Phonological relatedness between target and error in neologistic productions. *Brain and Language, 103*(1–2), 120–121. doi:10.1016/j.bandl.2007.07.074

Buchwald, A., & Rapp, B. (2003). The orthographic representation of consonant–vowel status: Evidence from two cases of acquired dysgraphia. *Brain and Language, 87*(1), 120–121. doi:10.1016/s0093-934x(03)00228-1

Caramazza, A. (1988). Some aspects of language processing revealed through the analysis of acquired dysgraphia: The lexical system. *Annual Review of Neuroscience, 11*, 395–421.

Caramazza, A., & Miceli, G. (1990) The Structure of graphemic representations. *Cognition, 37*, 243–297.

Caramazza, A., Miceli, G., Villa, G., & Romani, C. (1987). The role of the Graphemic Buffer in spelling: Evidence from a case of acquired dysgraphia. *Cognition, 26*(1), 59–85. doi:10.1016/0010-0277(87)90014-x

Cherney, L.R. (2004). Aphasia, alexia, and oral reading. *Topics in Stroke Rehabilitation, 11*(1), 22–36.

Chialant, D., Domoto-Reilly, K., Proios, H., & Caramazza, A. (2002). Preserved orthographic length and transitional probabilities in written spelling in a case of acquired dysgraphia. *Brain and Language, 82*, 30–46

Cohen, A. (2014). FuzzyWuzzy [Computer Software]. SeatGeek, Inc. Retrieved from https://pypi.python.org/pypi/fuzzywuzzy

Coltheart, M. (1980). Reading, phonological recoding and deep dyslexia. In M. Coltheart, K. Patterson, & J.C. Marshall (Eds.), *Deep dyslexia*. London, UK: Routledge & Kegan Paul.

Campbell, R. (1983). Writing nonwords to dictation. *Brain and Language*, *19*, 153–178.

Davies, R., Cuetos, F., & Rodriguez-Ferreiro, J. (2010). Recovery in reading: A treatment study of acquired deep dyslexia in Spanish. *Aphasiology, 24*(10), 1115–1131.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*(4), 801–838.

Ellis, A.W. (1982). Spelling and writing (and reading and speaking). In A.W. Ellis (Ed.), *Normality and pathology in cognitive functions*. New York: Academic Press.

Ellis, A.W., & Young, A.W. (1988). *Human cognitive neuropsychology.* Hove, UK: Lawrence Erlbaum Associates Ltd.

Falconer, C., & Buchwald, A. (2013). Do activated letters influence lexical selection in written word production? *Aphasiology, 27*(7), 849–866. doi:10.1080/02687038.2013.781114

Falconer, C., Miner, T., Velez, D., & Buchwald, A. (2011). Interaction between Word-Level and Letter-Level Processing in Written Language: Evidence from Acquired Dysgraphia. *Procedia - Social and Behavioral Sciences, 23*, 238–239. doi:10.1016/j.sbspro.2011.09.257

Fischer-Baum, S., & Rapp, B. (2012). Underlying cause(s) of letter perseveration errors. *Neuropsychologia, 50*(2), 305–318. doi:10.1016/j.neuropsychologia.2011.12.001

Fischer-Baum, S., Mccloskey, M., & Rapp, B. (2010). Representation of letter position in spelling: Evidence from acquired dysgraphia. *Cognition, 115*(3), 466–490. doi:10.1016/j.cognition.2010.03.013

Folk, J., & Jones, A., (2010). The purpose of lexical/sublexical interaction during spelling: Further evidence from dysgraphia and articulatory suppression. *Neurocase: The Neural Basis of Cognition, 10*(1), 65–69, doi: 10.1080/13554790490960512

Folk, J., & Rapp, B. (2004) Interaction of lexical and sublexical information in spelling: Evidence from nonword priming. *Applied Psycholinguistics, 25*, 565–585. doi: 10.1017.S0142716404001274

Folk, J., Rapp, B., & Goldrick, M. (2002). The interaction of lexical and sublexical information in spelling: What's the point? *Cognitive Neuropsychology, 19*(7), 653–671

Friedman, R.B. (1996). Phonological text alexia: poor pseudoword reading plus difficulty reading functors and affixes in text. *Cognitive Neuropsychology, 13*(6), 869–885.

Friedman, R.B. & Lott, S.N. (2002). Successful blending in a phonological reading treatment for deep alexia. *Aphasiology, 16*(3), 355–372.

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: A model of letter position coding. *Psychological Review, 115*, 577–600.

Goodman, R.A. & Caramazza, A. (1986). Aspects of the spelling process: Evidence from a case of acquired dysgraphia. *Language and Cognitive Processes, 1,* 263–296.

Goodglass, H., Kaplan, E., & Weintraub, S. (2001). *Boston Naming Test*. Philadelphia: Lea & Febiger.

Gvion, A., & Friedmann, N. (2010). Letter position dysgraphia. *Cortex, 46*(9), 1100–1113. doi:10.1016/j.cortex.2009.09.008

Henry, M., Beeson, P., Stark, A., & Rapcsak, S. (2007) The role of left perisylvian cortical regions in spelling. *Brain and Lang*uage, *100*(1), 44–52.

Hillis, A. E. (2002). Models of the Reading Process. In A. E. Hillis (Ed.), *The Handbook of Adult Language Disorders*, 3–14. New York, NY: Psychology Press.

Hillis, A. & Caramazza, A. (1991). Mechanisms for accessing lexical representations for output: Evidence from a category-specific semantic deficit. *Brain & Language, 40,* 106–144.

Hillis, A., Rapp, B., & Caramazza, A. (1999). When a rose is a rose in speech but a tulip in writing. *Cortex, 35*, 337–356

Houghton, G., & Zorzi, M. (2003). Normal and impaired spelling in a connectionist dual-route architecture. *Cognitive Neuropsychology, 20*, 115–162.

Howard, D. & Gatehouse, C. (2006). Distinguishing semantic and lexical word retrieval deficits in people with aphasia. *Aphasiology, 20*(9/10/11), 921–950. Psychology Press. doi: 10.1080/02687030600782679

Joshi, R. (2006). *Handbook of orthography and literacy*. Mahwah, N.J.: L. Erlbaum Associates.

Johnson, J., Ross, K., Kiran, S. (2015) Multifaceted Treatment for Acquired Alexia and Agraphia: Feasibility Testing and Identification of Effective Treatment Components. Doctoral Student Research Symposium Poster.

Jones, A. C., Folk, J. R., & Rapp, B. (2009). All letters are not equal: Subgraphemic texture in orthographic working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1389–1402. doi:10.1037/a0017042

Jónsdóttir, M. K., Shallice, T., & Wise, R. (1996). Phonological mediation and the graphemic buffer disorder in spelling: Cross-language differences? *Cognition, 59*(2), 169–197. doi:10.1016/0010-0277(95)00693-1

Kay, J., Lesser, R., & Coltheart, M. (1992). *Psycholinguistic Assessment of Language Processing in Aphasia (PALPA)*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Kendall, D., Conway, R., Rosenbek, J., & Gonzalez-Rothi, L. (2003). Phonological rehabilitation of acquired phonological alexia. *Aphasiology, 17,* 11. 1073–1095. doi: 10.10800268703034400355

Kertesz, A. (2006). Western Aphasia Battery-Revised. New York: Grune & Stratton.

Kim, E., Rapcsak, S., Andersen, S. & Beeson, P. (2011). Multimodal alexia: Neuropsychological mechanisms and implications for treatment. *Neuropsychologia, 49*, 3551–3562

Kiran, S., Balachandran, & Lucas (2014). The nature of lexical-semantic access in bilingual aphasia. *Behavioural Neurology,* Article ID 389565, 18 pages. doi: 10.1155/2014/389565

Kiran, S. & Viswanathan, M. (2008). Effect of model-based treatment on oral reading abilities in severe alexia: a case study. *Journal of Medical Speech-Language Pathology, 16* (1), 43– 59.

Lott, S.N. & Friedman, R.B. (1996). Multimodal alexia: Neuropsychological mechanisms and implications for treatment. *Neuropsychologia, 49*, 3551–3562

Martin, D. H., & Barry, C. (2012). Writing nonsense: The interaction between lexical and sublexical knowledge in the priming of nonword spelling. *Psychonomic Bulletin & Review, 19*(4), 691–698. doi:10.3758/s13423-012-0261-7

Miozzo, M., Rawlins, K., & Rapp, B. (2014). How verbs and non-verbal categories navigate the syntax/semantics interface: Insights from cognitive neuropsychology. *Cognition, 133*(3), 621–640. doi:10.1016/j.cognition.2014.08.004

Orjada, S.A., & Beeson, P.M. (2005). Concurrent treatment for reading and spelling in aphasia. *Aphasiology, 19(3/4/5)*, 341–351.

Patterson, K. E., Marshall, J. C., & Coltheart, M. (1985) *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Perea, M., & Carreiras, M. (2006). Do Transposed-Letter Similarity Effects Occur at a Syllable Level? *Experimental Psychology, 53*(4), 308–315. doi:10.1027/1618-3169.53.4.308

Rapcsak, S. Z., Henry, M. L., Teague, S.L., Carnahan S. D, & Beeson, P. M. (2007). Do dual-route models accurately predict reading and spelling performance in individuals with acquired alexia and agraphia? *Neuropsychologia, 45*, 2519–2524.

Rapp, B. (2002). Uncovering the cognitive architecture of spelling. In A. Hillis (Ed.), *The handbook on adult language disorders*, 47–69. New York: Psychology Press.

Rapp, B. & Caramazza, A. (1989). Letter processing in reading and spelling: Some dissociations. *Reading and Writing, 1*, 3–23.

Rapp, B. & Caramazza, A. (1997). From graphemes to abstract letter shapes: Levels of representation in written spelling. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1130–1152.

Rapp, B., Epstein, C., & Tainturier, M-J. (2002). The integration of information across lexical and sublexical processes in spelling. *Cognitive Neuropsychology, 19* (1), 1–29.

Rapp, B., Folk, J.R., & Tainturier, M-J. (2001). Word reading. In B. Rapp (Ed.), *The handbook of cognitive neuropsychology,* 263–289. Ann Arbor: Edwards.

Roeltgen, D.P. & Heilman, K. (1984). Lexical agraphia: Further support for the two-system hypothesis of linguistic agraphia. *Brain, 107,* 811–827.

Seidenberg, M. S. & McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review, 96,* 523–568.

Shallice, T. (1981). Phonological agraphia and the lexical route in writing. *Brain, 104,* 413–429

Tainturier, M-J., Bosse, M-L., Roberts, D.J., Valdois, S., & Rapp, B. (2013). Lexical neighborhood effects in pseudoword spelling. *Frontiers in Psychology, 4*(862), 1–8. doi: 10.3389/fpsyg.2013.00862

Tainturier, M-J. & Rapp, B. (2001). The spelling process. In B. Rapp (Ed.), *The handbook of cognitive neuropsychology*, 263–289. Ann Arbor: Edwards

Van Rossem, G. (2015, May 23). Python (Version 2.7.10) [Computer software]. Retrieved from https://www.python.org/downloads/release/python-2710/

Weekes, B. & Coltheart, M. (1996). Surface dyslexia and surface dysgraphia: Treatment studies and their theoretical implications. *Cognitive Neuropsychology, 12*, 277–315.

Zorzi, M., Houton, G., & and Butterworth, B. (1998). Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1131–1162.

Curriculum Vitae

Katrina Ross

EDUCATION

Undergraduate Education

- Bard College – *B.A Classics, 05/2013*
- University of Washington – *B.S. Speech and Hearing Sciences, 08/2014*

Graduate Education

- Boston University, Sargent College – *M.S. Speech Language Pathology - in progress, expected 05/2016*

CLINICAL EXPERIENCE

Brighton House Rehabilitation and Skilled Nursing Facility, *Brighton, MA*
01/2016 – 05/2016

- Evaluate and treat cognitive-communication impairments, dysphagia, aphasia, and motor speech disorders in short- and long-term care units

Spaulding Rehabilitation Hospital, *Charlestown, MA*
08/2015 - 12/2015

- Evaluate and treat cognitive-communication impairments, aphasia, dysphagia, voice disorders, and motor speech disorders in the Traumatic Brain Injury and Comprehensive Rehab units

Academic Speech, Language, and Hearing Center, *Boston University - Boston, MA*
10/2014 – 05/2016

- Evaluate and treat an individual with a fluency disorder, primary therapist
- Evaluated and treated individual with aphasia, primary therapist
- Evaluated and provided phonological awareness intervention in a group setting for children ages 4 - 6
- Provided community hearing screenings, preschool to adult clients

Aphasia Research Laboratory, *Boston University - Boston, MA, 10/2014 - present*

- Reading and writing treatment study, primary therapist to 4 adults with alexia/agraphia

Early Steps Preschool at Hosmer Elementary School, *Watertown, MA*
01/2015 - 06/2015

- Assessment and intervention, preschool

CERTIFICATIONS AND RELATED EXPERIENCES

Certified in Fiberoptic Endoscopic Evaluation of Swallowing (FEES), *01/2016*

Trained in Modified Barium Swallow (MBS) Impairment Profile, *05/2015*

- Completed all MBSImP Training modules, passed final evaluation of skills and knowledge acquisition

RESEARCH

Aphasia Lab Master's Thesis - *Boston University - Boston, MA*
10/2014 - present

- Developed a coding system to compare lexical and sublexical error progression in reading and writing

Aphasia Lab Research Assistant - *Boston University - Boston, MA*
09/2014 – present

- Aided in data collection, conducting speech and language assessment, investigating nature of reading and writing treatment response

Aphasia Lab Research Assistant - *University of Washington - Seattle WA*
03/2014 – 06/2014

- Aided in data collection, conducting speech and language assessment, investigating nature of error consistency in apraxia

PUBLICATIONS

Johnson, J., **Ross, K.,** Kiran, S. (2015) Multifaceted Treatment for Acquired Alexia and Agraphia: Feasibility Testing and Identification of Effective Treatment Components. Doctoral Student Research Symposium Poster.

PROFESSIONAL VOLUNTEER SERVICES

Volunteer, *Aphasia Access - Boston, MA*
03/2015

- Volunteer assistant and facilitator at 3-day event for clinicians, researchers, people with aphasia, and caregivers
- Responsibilities: registration, transcribing issue-based lectures, posters and materials preparation/set up

Volunteer *Aphasia Day - Seattle, WA*
06/2014

- Volunteered assistant and conversation partner at University of Washington's Community Aphasia Day
- Responsibilities: communication partner with PWA, transcribing issue-based lectures, preparation/set up

MEMBERSHIPS

National Student Speech, Language, and Hearing Association
2013 - present