Lesion mapping of syntactic and lexical features derived from Natural Language Processing of narrative speech elicited by patients with chronic post-stroke aphasia

E. Gleichgerrcht[1], J. DelGaizo[1], J. Fridriksson[2], D. den Ouden[2], A. Basilakos[2], C. Rorden[3], L. Bonilha[1]

1. Department of Neurology, Medical University of South Carolina, Charleston, SC, USA
2. Department of Communication Sciences and Disorders, University of South Carolina, Columbia, SC, USA
3. Department of Psychology, University of South Carolina, Columbia, SC, USC

Background

- Language assessment in patients with aphasia post-stroke is usually conducted with batteries of standardized tests aimed at identifying residual language capacities. Many of these tests rely on patients' responses to isolated stimuli, sometimes threatening their ecological validity in certain conditions.
- Narrative/connected speech, however, can provide a rich source of response samples to obtain insightful information about patients' language skills. The analysis of such data, however, is highly time-consuming and requires specific training.
- We sought to analyze connected speech in patients with chronic stroke using Natural Language Processing to identify measures sensitive to aphasia and to understand their neurobiological bases with VLSM.

Methods

- 65 participants (66.2 % M) who had a left-hemisphere stroke at least 6 months ago (19 without residual aphasia, 18 Broca's, 14 anomic, 9 conduction, 3 Wernicke's, and 1 global). Mean age was 58.5 (*SD* = 10.3) and mean months post-stroke was 37.2 (SD = 40.2) months.
- 10 age- and gender- comparable health controls with no history of neurological or psychiatric disease
- All participants completed the WAB-R and were asked to describe, for 2 minutes each, three scenes:
    1. Cookie Theft scene from the BDAE
    2. Picnic scene from the WAB
    3. Circus scene from the Apraxia Battery for Adults
- Discourse was transcribed verbatim

Methods & Results

- By means of sentence boundary disambiguation, parsing, and parts-of-speech tagging, we derived **37 lexical and syntactic features**
- We applied **Principal Components Analysis** (PCA) on these features and applied voxel-based lesion-symptom mapping (VLSM) to identify the neural basis associated with principal components that accounted for > 50% of the variance.

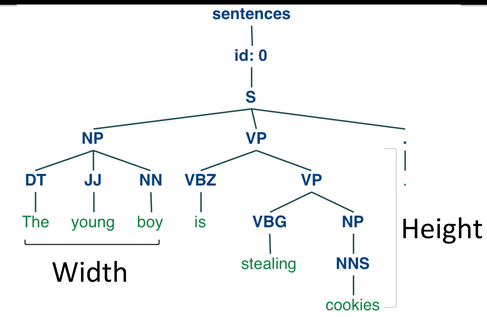
Principal Component Analysis

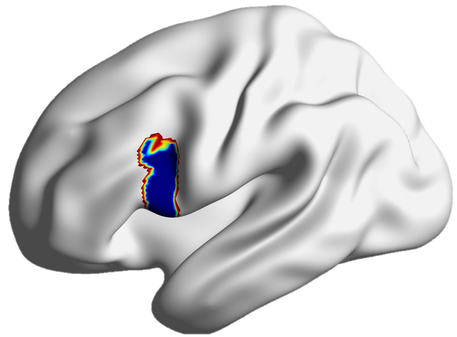
	Components 1	2
% Variance Explained	**44.8%**	**11.5%**
Clause Width	.954	-.110
VP width	.949	-.084
VP height	.926	.063
Verb Phrases	.923	.154
NP to NP distance	.919	-.133
Number of Words	.917	.258
Speech Rate	.917	.258
Clauses	.907	.123
Clasue Height	.903	-.010
NP to VP distance	.897	-.110
Numb Diff Words	.891	.012
VP to VP distance	.874	-.218
NP width	.861	.124
Length of Sentence	.854	-.227
Noun Phrases	.843	.363
Clause Per Sentence	.814	-.398
Number of Prepositions	.778	.150
Token Type Ratio	.755	.140
Dept Clause Per Clause	.718	-.299
NP height	.688	.339
Mean Imagery	-.523	.388
Lexical Density	-.471	.288
Mean Frequency	-.072	.759
Numb Nouns	-.226	.726
Numb Adverbs	-.097	-.633
Numb Adjectives	-.085	-.567
Adverb Variation	.023	-.545
Noun Variation	-.076	.149
Verb Variation	-.166	.352
Lexical Variation	-.336	.501
Word Length	.097	-.157
Mean Length Clause	.273	.287
Sentences	.421	.501
Mean AoA	.391	-.208

- **Factor 1** was strongly composed of mainly **syntactic** features and VLSM was thus controlled for **WAB fluency**
(in decreasing order of loading scores: clause width, width/height and number of verbal phrases, number of words, speech rate, number of clauses and their height, distance between noun phrase and verb phrases, etc.)

- **Factor 2** was strongly composed of mainly **lexical** features and VLSM was thus controlled for **WAB naming**
(in decreasing order of loading scores: frequency of all words, number of nouns, adverbs, and adjectives, as well as adverb, noun, verb and overall lexical variation)

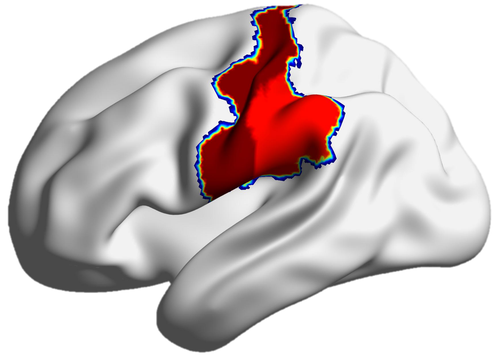
Sample parsing of the sentence "The boy is stealing cookies" demonstrating some features of syntactic complexity such as width (in this case, the NP has a width of 3) and height (in this case, the VP has a height of 2)





Factor 1 (Syntax) controlling for Fluency

Inferior frontal gyrus



Factor 2 (Lexical) controlling for Naming

Precentral gyrus
Postcentral gyrus
Pos superior temporal gyrus
Supramarginal gyrus

Conclusions

We showed that syntactic performance beyond speech fluency might rely on the inferior frontal gyrus (Broca's area), while lexical performance beyond fluency may depend on areas in the precentral and post-central gyri as well as the postero-superior temporo and inferio-parietal regions. Our findings show that NLP applied to connected speech elicited by patients with post-stroke aphasia can shed light on the organization of language in brains with vascular damage.