

Prediction of Lexical Anomalies (POLA)

Korcovelos, E.A. and McInnes, B.T., Departments of Computer Science and Bioinformatics

Introduction

Dementia is a neurodegenerative disease that results in a gradual decline in cognition and memory. Recent reports on dementia suggest 23 in 100 people aged 60-84 currently possess the disease, whereby the prevalence is expected to double every five years. Dementia is known to affect cognition and behavior in the elderly and is comprised of both behavioral and language impairments. These language impairments heavily impact communication, with many individuals experiencing declines in expressive and receptive language abilities. However, it is important to note that individuals with dementia often have difficulties with the transmission of words in written and spoken communication rather than a deficit in word knowledge and word meanings. For individuals with Mild Cognitive Impairment (MCI), a syndrome that lies between normal cognitive aging and dementia, there is a high risk of developing dementia in the following months and years. Language variants between those with MCI and dementia have not been heavily studied computationally. For this reason, there is a critical need to develop algorithms and tools that are able to assess speech in ways that would lead to earlier diagnoses and treatment. **The present study aims to elucidate the psycholinguistic properties of dementia and MCI patients, with particular focus on speech complexity through natural language processing techniques.**

Methods

DementiaBank's Cookie Theft Test to determine lexical variance

Metric of focus: type to token ratio (unigram to 11-gram models)
 In cases where the type to token is smaller, it is predicted that individuals will suffer less from dementia-like symptoms. Conversely, where the type to token ratio is greater, it is predicted individuals will encompass some dementia symptoms

Building n-gram models

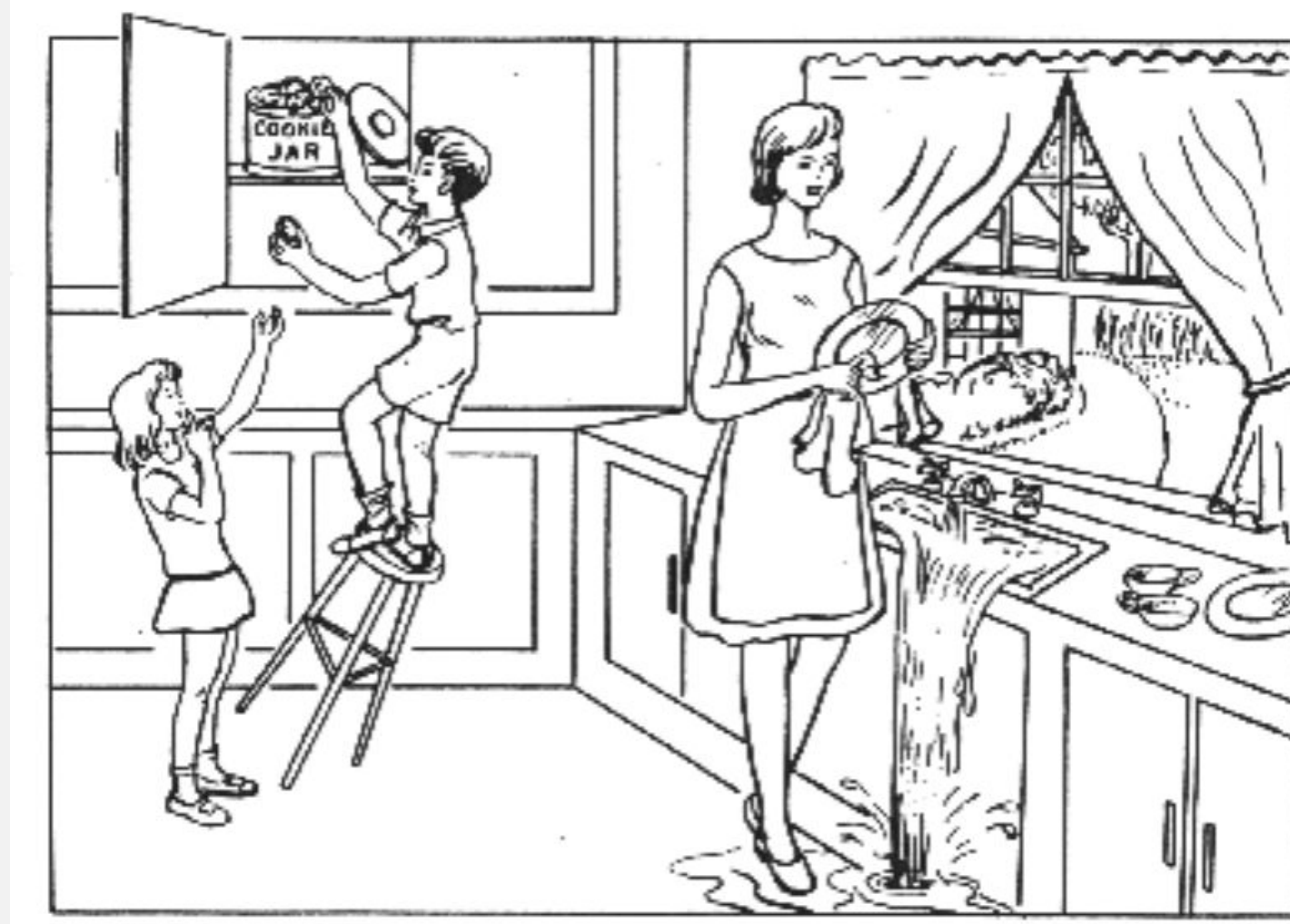
The boy is, um, the boy is in the room.

"[Tt]he boy is" is repeated as a possible result of cognitive impairment. The type count for this sentence is five, while the token count is ten (*pause* words such as "uh" and "um" are included in the type and token counts). These numbers would result in a token-type ratio of 6:10, whereby a normally aging individual may experience a token-type ratio of 5:6 (where the second "the boy is" and *pause* words do not appear).

Acknowledgements

I would like to thank Dr. Serguei Pakhomov Dr. Graeme Hirst for their undying support and guidance.

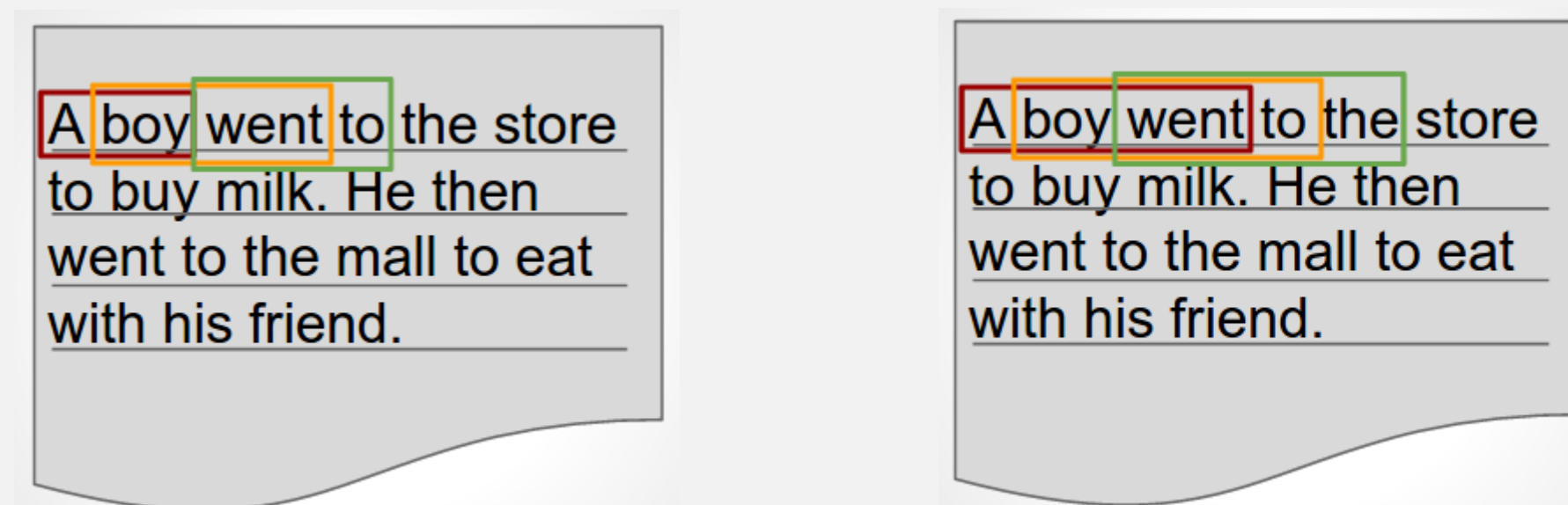
Example



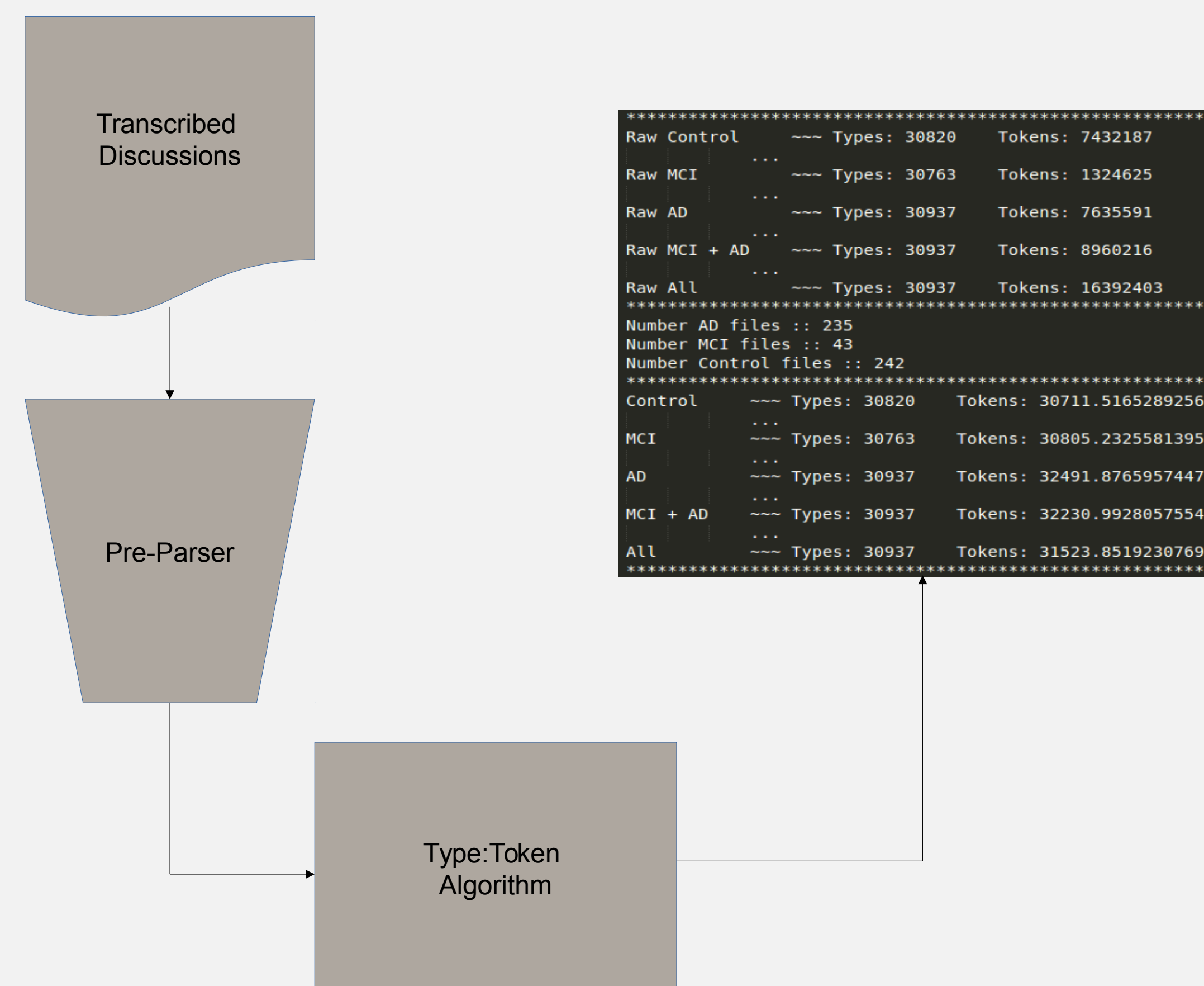
1. Transcription: Participants were asked to describe the Cookie Theft picture; descriptions were transcribed into Dementia Bank's Transcript Database

2. Pre-parsing: Transcriptions were parsed and removed of extraneous data, including part of speech tagging of words, emotion tagging ("laughs", "coughs", etc.), and removal of all investigator questions and responses.

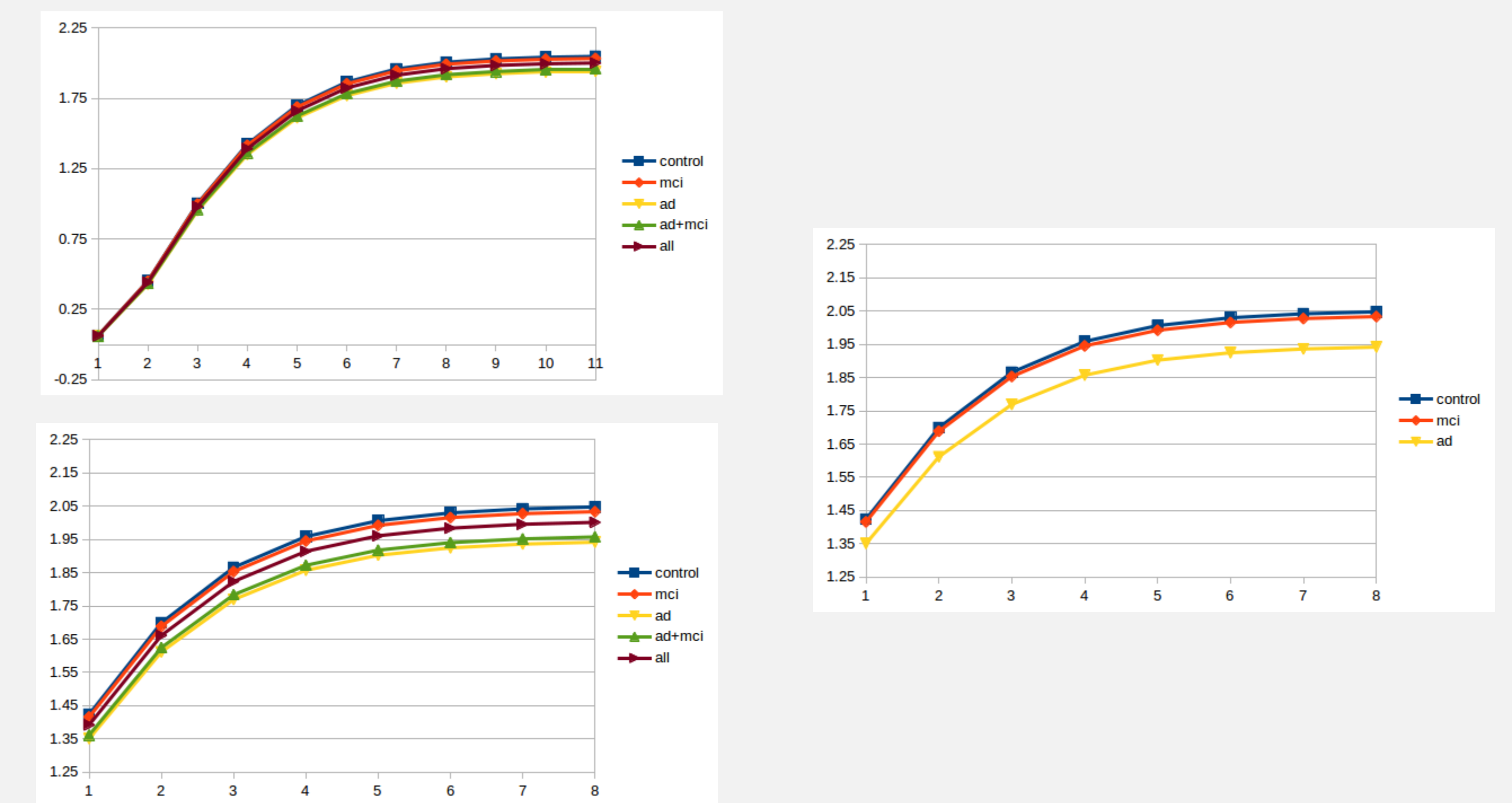
3. Main Algorithm: N-grams of size 1-11 were created and analyzed for uniqueness.



4. Type:token ratios were calculated and outputted.



Preliminary Results



Like the algorithms employed by Lancashire and Hirst (2009), our models confirm that differences persist in repeated phrase usage in dementia and control patients per the n-gram model. Moreover, the results demonstrate a similarity between the control and MCI patients, a trend that had never been seen prior to the present study.

Future Work

For future assessment, the following features will be incorporated into our algorithm to form a more robust analysis of language variation in dementia patients:
 Indefinite word usage – measured by the information content of each sentence
 Pause word usage – frequency of tokens like "uh" and "um" within text
 N-gram analysis of parts of speech – type:token extension to POS tagging

References

[1] Fratiglioni, L., De Ronchi, D., Aguero-Torres, H. (1999). Worldwide Prevalence and Incidence of Dementia. *Drugs and Aging*, 15(5), 365-375.
 [2] Potkins, D., Myint, P., Bannister, C., Tadros, G., Chithramohan, R., Swann, A., O'Brien, J., Fossey, J., George, E., Ballard, C., & Margallo-Lana, M. (2003). *International Journal of Geriatric Psychiatry*, 18, 1002-1006.
 [3] Burke, D., & Shafto, M. (2004). Aging and Language Production, *Current Directions in Psychological Science*, 13(1), 21-24.
 [4] Baiyewu, O., Unverzagt, F., Ogunniyi, A., Smith-Gamble, V., Gureje, O., Lane, K., Gao, S., Hall, K., & Hendrie, H. (2012). Behavioral Symptoms in Community-Dwelling Elderly Nigerians with Dementia, Cognitive Mild Impairment, and Normal Cognition. *Geriatric Psychiatry*, 27(9), 931-939.
 [5] Unverzagt, F., Gao, S., Baiyewu, O., Ogunniyi, A., Gureje, O., Perkins, A., Emsley, C., Dickens, J., Evans, R., Musick, B., Hall, K., Hui, S., & Hendrie, H. (2001). Prevalence of Cognitive Impairment: Data from the Indianapolis Study of Health and Aging. *Journal of the American Heart Association*, 57(9), 1655-1662.
 [6] Lancashire, I. And Hirst, G. (2009). Vocabulary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study. *Cognitive Aging: Research and Practice*, 8-10 March 2009, Toronto, Canada.