Building a Longitudinal Corpus of Emails Composed by Older Adults

Krystal Klein
Department of Biomedical Engineering
Oregon Health & Science University
kleinkr@ohsu.edu

Abstract:

Previous research suggests that linguistic features associated with AD-pathology are present long before patients are diagnosed through standard practice [1–3]. The current project is aimed at longitudinal corpus building and identifying semantic and syntactic measures to detect cognitive decline over time.

Introduction:

Relative to healthy elders, people with mild cognitive impairment (MCI)---a syndrome that frequently precedes Alzheimer's Disease (AD)---display impaired lexical access[4] and favor simpler syntactic construction in their speech[1,5]. Further, epidemiological studies have found that mean differences in prodromal linguistic behavior differentiate between groups who ultimately do or do not develop symptomatic AD decades before diagnosis[2,3,6,7]. Together, these observations suggest that linguistic features associated with AD-pathology are present long before most people are diagnosed through standard practice—consistent with the long time-course of the disease's neuropathology. There is therefore reason to predict that automatic, continuous monitoring of linguistic behavior has the potential to serve as a first line of action in detecting onset of decline.

Corpus Building:

Participants: As I aspired to build a longitudinal email corpus, I recruited participants from an existing cohort participating in a longitudinal study of aging and technology through Oregon Center for Aging and Technology (ORCATECH). Qualifications were that the subject had sent at least 12 emails over the past year using an orcatech.org email account. Of 12 older adults who have consented to date, 8 have been enrolled in exchange for $20 payment, and 4 were screen failures.

***Corpus composition*:** Subject emails were composed over a one to six year period. Each sent email is considered a separate document. Currently, only messages with a plain text format are analyzed and other portions (e.g. attachments, html) are ignored.

Corpus Processing and Annotation:

Preprocessing: Any potential scientific interest in linguistic metrics derived from email is predicated on the assumption that content is user-generated. but even plain text messages frequently contain several types of content that are not user-generated or not generated at the time of email composition, for example: (a) phrases generated by websites, such as "These pictures were sent with Picasa, from Google." (b) forwarded content, (c) email signatures, which are composed at one point but automatically or manually attached to the bottom of emails, (d) content copied and pasted from the web. This last one is probably beyond scope to attempt to detect. Regarding forwards, I am detecting formatting specific to forwards and removing that content. Email signatures often have special formatting, but sometimes do not, so it may be necessary to remove identical content across emails. This has not yet been implemented in any way.

Syntactic Analysis: Sentence tokenization is achieved using the Natural Language Processing Toolkit for Python[8]. Then, the BUBS[9] parser will be used for automatic parsing (with manual annotation of a subset of the corpus to train/validate), and I will calculate syntactic complexity metrics, including: (1) Content density, calculated as the ratio of open-class (nouns, verbs, adjectives, adverbs) to closed-class (other) words (2) Yngve deviation, which describes displacement from the right-branching tree form common in English and has been found to have association with working memory. (3) Finally, we will extract average dependency distance.

Semantic Analysis: Planned semantic work includes use of topics modeling[10], analysis of age-of-acquisition of vocabulary, and type-to-token ratios.

Clinical Evaluations:

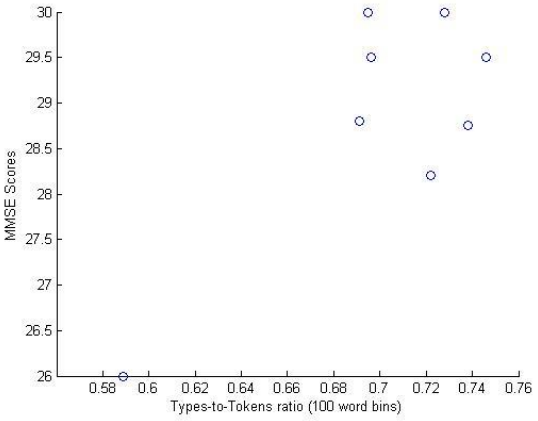
As participants in the aforementioned technology study, subjects had been given clinical neuropsychology assessments annually. Eighteen metrics were available from a cognitive battery comprising the following tests: Letter-number sequencing (1), Trail making test (2), Digit-Symbol test (1), CERAD word list (3), Visual reproduction (2), Logical Memory (2), Boston Naming (1), Digits Forward (1), Digit Span Backward (1), Block design (1), Picture completion (1) and semantic verbal fluency (2). Additional assessments from the same visit included: Mini-Mental State Examination, Functional Assessment, Geriatric Depression scale, and Cumulative illness score. Finally, the participants were assessed on the Clinical Dementia Rating (CDR) scale by a different clinical researcher who did not have test score information. The CDR is a scale commonly used to assess cognitive impairment

Results and Discussion:

Subject Characteristics:

Total Subjects (Female)	8(7)
Age (at consent), Mean (SD)	85.9 (8.4)
Clinical Visits: Mean (SD)	4.4 (0.7)
Emails sent: Mean (SD)	464.5 (243.2)
Emails sent/month: Mean (SD)	24.2 (32.6)

Data analysis is in progress; as a toy problem I have looked at average type-to-token ratio per 100 words. See figure below for a comparison of this ratio to MMSE scores.



The other analyses described are in progress, and I am actively recruiting for the study, especially from subjects who have been assessed as having mild cognitive impairment.

References

1. Kemper, S., Thompson, M. & Marquis, J. Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. *Psychol Aging* **16**, 600–614 (2001).
2. Snowdon, D. A. *et al.* Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. *Journal of the American Medical Association* **275**, 528–32 (1996).
3. Snowdon, D. A., Greiner, L. H. & Markesbery, W. R. Linguistic ability in early life and the neuropathology of Alzheimer's disease and cerebrovascular disease. Findings from the Nun Study. *Ann N Y Acad Sci* **903**, 34–8 (2000).
4. Duong, A., Whitehead, V., Hanratty, K. & Chertkow, H. The nature of lexico-semantic processing deficits in mild cognitive impairment. *Neuropsychologia* **44**, 1928–1935 (2006).
5. Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K. & Kaye, J. Spoken language derived measures for detecting Mild Cognitive Impairment. *IEEE Transactions on Audio, Speech, and Language Processing* **19**, 2081–2090 (2011).
6. Tyas, S. L., Snowdon, D. A., Desrosiers, M. F., Riley, K. P. & Markesbery, W. R. Early-life linguistic ability, late-life pathology and asymptomatic Alzheimer's disease: Findings from the Nun Study. *Alzheimer's and Dementia* **5**, P103–P104 (2009).
7. Riley, K. P., Snowdon, D. A., Desrosiers, M. F. & Markesbery, W. R. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun Study. *Neurobiol Aging* **26**, 341–7 (2005).
8. Bird, S., Loper, E. & Klein, E. *Natural Language Processing with Python*. (O'Reilly Media Inc: 2009).
9. Dunlop, A., Bodenstab, N. & Roark, B. Efficient matrix-encoded grammars and low latency parallelization strategies for CYK. 163–174 (2011).
10. Steyvers, M. & Griffiths, T. Probabilistic Topic Models. *Latent Semantic Analysis: A Road to Meaning*

Acknowledgements:

This research was supported by a pilot grant from the Oregon Center for Aging & Technology (ORCATECH, NIH #1P30AG024978-01). Thanks to my project mentor, Brian Roark, to Meg Mitchell and Steven Bedrick for helpful suggestions, and to ORCATECH investigators Jeff Kaye, Tamara Hayes, Tracy Zitzelberger, Nora Mattock, Colette Duncan, Holly Jimison, and Johanna Feuerstein for facilitating this research.