The Case of the Cookie Jar: Differences in Typical Language Use in Dementia

Davida Fromma

Sarah Grace Daltonb

Alexander Brickc

Gbenuola Olaiyac

Sophia Hillc

Joel Greenhousec

Brian MacWhinneya

aCarnegie Mellon University, Department of Psychology, Pittsburgh, PA

bMarquette University, Department of Speech Pathology and Audiology, Milwaukee, WI

c Carnegie Mellon University, Department of Statistics and Data Science, Pittsburgh, PA

Corresponding Author:

Davida Fromm

Department of Psychology

Carnegie Mellon University

5000 Forbes Ave.

Pittsburgh, PA 15213

fromm@andrew.cmu.edu

**Abstract** <I’ll finish this once the doc is finished>

**Purpose:** Language sample analysis tools are efficient and effective indicators of cognitive impairment in older adults. This study used automated core lexicon analyses of Cookie Theft picture descriptions from three groups: (1) adults without diagnosed cognitive impairments (CTL), (2) adults diagnosed with Alzheimer’s disease (AD), and (3) adults with mild cognitive impairments (MCI).

**Method:**

**Results:**

**Conclusion:** This standard and simple-to-administer task reveals group differences in overall core lexicon scores and the amount of time until the speaker produces the key items. Clinicians and researchers can use these tools for both early assessment and measurement of change over time.

The Case of the Cookie Jar: Differences in Typical Language Use in Dementia

Speech and language abilities are important factors in detecting cognitive impairments in older adults (Filiou et al, 2020; Mueller et al., 2018; Talor & Phillips, 2008). They are especially important given that evaluation of speech and language can be simple and non-invasive compared with many other diagnostic procedures (Vigo et al., 2022). Traditionally, using discourse-level language to investigate the lexical/semantic system in these individuals has been challenging due to the time and expertise required to collect, transcribe, and analyze large numbers of discourse samples (Bryant et al., 2017; Stark et al., 2021). However, shared databases and advances in computer technology, natural language processing, and machine learning procedures have vastly improved our ability to use connected speech as an efficient and non-invasive classification and measurement tool (Beltrami et al., 2018; Fraser et al., 2018).

Automatic detection of dementia from connected speech has achieved varying degrees of accuracy depending on which classifiers and features are used, with most studies reporting accuracies ranging from mid-70% to mid-80% (see reviews by de la Fuente Garcia et al.; Thaler & Gewald, 2021; Vigo et al., 2022). Many of these studies analyze Cookie Theft picture descriptions (Goodglass et al., 2001) from the Pitt corpus (Becker et al., 1994) in DementiaBank[[1]](#footnote-1). Given that these techniques are able to use various speech and language features from these short language samples to produce algorithms that can successfully detect dementia, it would be advantageous to find ways to translate some of that machine learning knowledge to the clinic.

One of the tools developed from large, shared databases using standard elicitation protocols, standard transcription formats (CHAT, https://talkbank.org/manuals/CHAT.pdf), and automated language analysis program (CLAN, <https://dali.talkbank.org/clan/>) is core lexicon (Dalton et al., 2020). Core lexicon analysis measures “typicality” of words used in discourse based on normative data (words produced by 50% or more of controls who did the task) (Dalton et al., 2022). It has the advantage of being straightforward and low-tech for busy clinicians, not even requiring transcription. Core lexicon may be particularly useful in dementia, since adults with AD and MCI often use less precise language than adults without cognitive impairments and would likely score lower on total core lexicon items produced. The lexical/semantic system is most vulnerable to cognitive impairment, and language is often described as “empty” (Bourgeois et al., 2010; Croisile et al., 1996; Ross et al., 1990). Fraser et al. (2018) reviewed speech and language findings for the Cookie Theft picture task in this population, focusing specifically on measures of information content. Early indications of impairment from a simple language task can facilitate early intervention such as clinical drug trials, external memory aid treatments, counselling, or lifestyle changes.

The Cookie Theft core lexicon checklist consists of 26 words, including functors (determiners, pronouns, prepositions, etc.) and content words (nouns, verbs). Croisile et al. (1996) created a list of 23 information units created for the Cookie Theft picture based on the four categories of subjects, place, objects, and actions that has been used to analyze differences between groups with and without dementia. Another list of 23 information units was derived from Swedish and English speakers’ Cookie Theft picture descriptions using cluster models (Fraser et al., 2019). The core lexicon list differs from these lists in that it contains some functor words.

Core lexicon scoring is straightforward, with 1 point given for each word produced from the checklist, regardless of number of times produced. Importantly, inflected forms of core lexicon items receive credit (e.g., if “wash” is a core lexicon item, “wash”, “washing”, and “washed” are acceptable forms), but not synonyms (e.g., “scrub” would not receive credit for the item “wash”, “woman” would not receive credit for the item “mother”). This is another way in which the core lexicon differs from the studies done with the information units list. It is important to note that core lexicon scores should not be interpreted as a measure of overall informativeness of a discourse sample, since synonyms do not receive credit.

Previous studies have shown that groups with and without dementia differ in the production of information units on the Cookie Theft description task, with AD groups producing fewer information units (Ahmed et al., 2018; Croisile, et al. 1996; Yancheva & Rudzicz, 2016). Fraser et al. (2019) reported that MCI group participants produced relevant information units at a significantly slower rate than controls, while not differing significantly in speaking rate. The analysis was based on computing information efficiency (number of relevant information units, divided by the total time). With newly available automated word alignment analyses it may be possible to further explore the timing of essential lexical information so that clinicians have additional tools to use when assessing individuals with complaints of cognitive impairments. The goals were to determine if core lexicon analysis could be a valid and reliable tool for use in the assessment of cognitive impairment in MCI and dementia and to better understand the linguistic behaviors that underlie the performance of the automated classification models. To accomplish this, we asked the following questions:

1. Do the groups differ in total number of core lexicon words produced?
2. Do the groups differ in the elapsed time before core lexicon content words are produced?
3. Do the groups differ in elapsed time before core lexicon content words are produced for each quadrant of the picture?
4. Do the groups differ in the order in which they look at the quadrants in the picture based on the core lexicon content words produced?
5. Do the groups still differ significantly in XXX if the analysis fixes which quadrant was looked at first?

**Method**

**Participants**

Participants with diagnoses of probableAD (n=234), MCI (n=41), and no cognitive impairment (n=243) were drawn from the Pitt corpus (<https://dementia.talkbank.org/access/English/Pitt.html>) (Becker et al., 1994) in the DementiaBank database. The Pitt corpus is the largest publicly available speech corpus including individuals with and without dementia and has been used in over 300 scientific publications. All participants received a battery of tests that included the Cookie Theft picture description task from the Boston Diagnostic Aphasia Exam (Goodglass et al., 2001). Severity of dementia was measured using the Mini Mental Status Exam (Folstein et al., 1975).

**Procedures**

Language samples were transcribed in CHAT format and analyzed using automated commands from the CLAN program. CLAN is freely downloadable software (https://dali.talkbank.org/clan/) that includes the CHAT editor and allows for automated analysis of linguistic and discourse structures (MacWhinney, 2000). Two new CLAN commands were used: 1) the basic CORELEX command, which searches for lemmas on the morphological tier of the CHAT transcript (%mor) to compute the total number of Cookie Theft core lexicon words (see Appendix) produced in the sample at least once (Dalton et al., 2022); and 2) a modified version of the basic CORELEX command that searches on the word alignment tier of the transcript (%wor) to compute the time until a given word from the core lexicon was used. The example below shows the CHAT transcript’s main speaker tier with the utterance time stamp (in msec), the %wor tier with time stamps for each word, and the %mor tier with part-of-speech and morphological tagging.

\*PAR: mom &-uh washing or drying dishes . •9650\_13350•

%wor: mom •9650\_10260• &-uh •10260\_10660• washing •11400\_11810• or

•11810\_12010• drying •12010\_12570• dishes •12570\_13350• .

%mor: n|mom part|wash-PRESP coord|or part|dry-PRESP n|dish-PL .

For this utterance, the CORELEX command would score one point each for the production of “mom”, “dry”, and “dish”, which are three of the 26 words in the Cookie Theft core lexicon checklist. As described in Dalton et al. (2022), three CLAN commands were used to acquire the basic core lexicon results for the full dataset.

1. **chstring +q1 \*.cha** was used on the original CHAT transcripts to remove revision codes and underscores (e.g., all\_of\_a\_sudden) and replace target replacements for paraphasias (e.g., chair [: stool]) with double colons (chair [:: stool]). These changes allowed the revised words and each element of the underscored words to be parsed on the %mor tier where the CORELEX command looks for lemmas. The double colon, on the other hand, prevents the target replacement word from being parsed on the %mor tier, and instead uses whatever word was spoken by the participant.
2. **mor \*.chstr.cex** was run on the new files created from step 1.
3. **corelex +lcookie +t\*par \*.chstr.cex** was run on the new files created from step 2, producing a spreadsheet showing which words from the core lexicon checklist were used.

Timing data were computed in msec from the beginning of the participant’s description of the picture until the time when each content word was first produced using this command: **corelex +lcookie-short +t\*par +w \*.cha**. Using the sentence in the example above, if this is the participant’s first production of “dry” (at 12010 msec) and the participant began describing the picture at 4389 msec into the transcript (after Investigator prompts and such), the CLAN program would subtract 4389 from 12010 and report that the participant first said a word containing the lemma “dry” at 7621 msec. While the full core lexicon checklist includes functors and content words, this timing analysis used only content words from the core lexicon list (the ones with an asterisk in the Appendix), as it was considered less interesting to determine when the first “a” or “and” appeared in the transcript. (Note: The %wor tier does not require any preliminary steps, as it includes revised words and excludes target replacement words from the main speaker tier.)

**Statistical Analysis**

ANOVA was used to determine if there was a difference in the means of total number of core lexicon words produced at least once by each group. Kaplan Meier Survival Curves were used to visualize the distribution of the time (t) until core lexicon words are said across the different groups of subjects. These curves show the percentage of people in each group who have not yet said a specific word of interest along the timeline. By comparing the curves, one can see how the groups differed in the time elapsed before they said the word in question. Log-Rank statistical tests were then used to compare the time distributions and determine if the groups differences in time elapsed were statistically significant.

**Results**

**Overall Core Lexicon Differences Across Groups**

Figure 1 shows a clear distinction in the cumulative frequency distribution of the total number of core lexicon words said between the Control and the ProbableAD groups, while the Control and MCI group have a more similar distribution up to a point (approximately the first 16 core lexicon words) when the MCI group’s distribution begins to be more like that of the ProbableAD group. Pairwise ANOVA tests confirmed these results, revealing that the Control and ProbableAD groups differed in the total number of unique core lexicon words that were said at least once (p < .001); the CTL and MCI groups did not. (p = 0.056). Running a 3-way ANOVA test provided enough statistical evidence to conclude that at least one of the groups differed in the mean number of core lexicon words said (p < .001), which is consistent with what is shown in Figure 1 and what was determined by the pairwise ANOVA test.

Figure 1 about here

Figure 1.

A graph of a number of words

Description automatically generated with low confidence*Cumulative frequency plot of total number of core lexicon words produced*

Figure 2 shows the mean number of times each core lexicon word was said, in order from highest frequency to lowest frequency overall. Each word has 3 bars for the mean of each group (Control, MCI, ProbableAD). The top 5 words said are “she/her”, “cookie”, “he/his”, “dish”, and “water”. The top 3 words said have a significantly higher means than the rest of the words, with the former means ranging from 2-3 and the latter from 1.5-0.5 (excluding the outlier “flow” with an average of approximately 0). The bottom 5 words are “fall”, “dry”, “window”, “run”, and “flow”. (Note: The lemma “flow” can also appear in any form of the word “overflow”, which describes the water coming from the kitchen sink onto the floor.)

For most words, the means between groups are similar. An ANOVA test of mean number of times a word was said ~ Word + Group showed that “word” was significant (p < .001) but group was not (p-value = 0.07), indicating that the means for these words was not significantly affected by the group variable. However, this does not suggest that for certain individual words, group does not contribute, like “mom” and “get”.

Figure 2 about here

Figure 2.

*Frequency of core lexicon words said across groups*

A picture containing text, plot, screenshot, diagram

Description automatically generated  
**Overall Core Lexicon Differences by Sex**

<I decided not to include this part because the analysis of core lexicon words spoken by males vs females was done for the ProbableAD group only. I don’t have good rationale to support that, and I don’t think this question is essential to the study. Then, the males vs females analysis was done for all 3 groups (Control, MCI, ProbableAD) but only for the word “mom”. Again, that won’t stand alone. But, there’s so much good material here that provides, interesting, important, and novel information. The male/female question is not irrelevant – definitely something to consider – but it was not central to the original aims.>

**Elapsed Time for Core Lexicon Word Production by Group**

Figure 3 shows the length of time until the word “girl” was said by each group. The x-axis shows time (in msec); the y-axis shows the proportion of people who have not yet said “girl”. For example, at the top left, all three curves start at 1.0 at time 0 because nobody (100% of individuals in the group) has said any words yet. As time increases along the x-axis, the curves step downward, as some proportion of people in the groups say the word “girl” for the first time. In this plot, all three curves start aligned with each other, but the ProbableAD group’s curve starts to decrease its slope before that of the Control and MCI groups, indicating that individuals in the ProbableAD group are not saying this word as early in their picture descriptions as individuals in the other groups. Additionally, along the path that the MCI and Control groups take, the MCI group’s slope lowers compared to the Control group right after 50% of the MCI group participants say the word (at the dashed line, the median), indicating that they are taking longer to say “girl” than the control. A Log-Rank test validated that the differences between these three curves was not simply due to chance (p < 0.001).

Figure 3 about here

Figure 3.

*Elapsed time until “girl” was said across groups*

A picture containing text, diagram, line, plot

Description automatically generated

Figure 4 shows a very similar trend to that seen in Figure 3, except that here the Probable AD group almost immediately shows a longer time lapse before saying the word “fall” compared to the other groups. Also, like Figure 3, the MCI group once again initially continues at a similar rate as the Control group before eventually slowing down by comparison. Once again, the Log-Rank test indicated that these differences were significant (p < 0.001).

Overall, for the vast majority of content words in the dataset, the same trend occurred, with the Probable AD group differentiating itself early and the MCI and Control groups displaying varying differences of a lesser degree. <I think we need to create a table with Log-Rank test results for each of the words?> However, a different trend was observed for two words, “cookie” and “jar”. Figures 5 and 6 show that “cookie” and “jar” are different because the MCI group’s curve is almost identical to and at even slightly faster than the Control group. Additionally, particularly for the word “cookie”, the point at which the ProbableAD slope slows by comparison to the other groups is substantially lower than that on the plots for the other words. One reason we may see these differences compared to the other words in the dataset is that the words “Cookie Jar” appear in the picture. Log-Rank tests yielded a statistically significant difference between the groups for “cookie” (p < .001) but not for “jar” because of multiple comparisons (p = .02).

Figures 5 and 6 about here

Figure 5

*Elapsed time until “cookie” was said across groups*

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure 6

*Elapsed time until “jar” was said across groups*

A picture containing text, diagram, screenshot, line

Description automatically generated

**Analysis of Picture Quadrants**

The content words were broken into quadrants, such that each word unambiguously belonged to a particular quadrant, as can be seen in the Cookie Theft picture below: top left included “cookie”, “jar”, “boy”; top right included “mom”, “window”, “dish”, “dry”; bottom left included “fall”, “stool”, “girl”; and bottom right included “run”, “sink”, “water”, “flow”.

Figure 7 about here

Figure 7

*Cookie Theft Picture*

A picture containing text, linedrawing

Description automatically generated

To analyze where the individuals in each group were looking, we found the time until the first the word for each quadrant was said by members of each group and made Kaplan-Meier survival curves to visualize the distributions. (The MCI group was not included in these analyses due to the smaller sample size of that group.) While there is a visible difference in the curves in Figure 8a, it is not statistically significant after a multiple testing correction. However, for the other three figures, Log-rank tests validated that the ProbableAD group took significantly longer to produce a relevant word (p < .001), with the largest difference occurring in the bottom right quadrant.

Figures 8a, 8b, 8c, 8d about here

Figure 8a

*Elapsed time until any quadrant 1 content word is produced*

A picture containing text, screenshot, line, plot

Description automatically generated

Figure 8b

*Elapsed time until any quadrant 2 content word is produced*

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure 8c

*Elapsed time until any quadrant 3 content word is produced*

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure 8d

*Elapsed time until any quadrant 4 content word is produced*

A picture containing text, screenshot, line, diagram

Description automatically generated

The Kaplan-Meier Survival curves were also used to compare when each of the groups reached the different quadrants. Figure 9a shows that the Control group tends to do the left side quadrants together, and then the right side at a slower rate. This plot reveals a substantial difference between the two sides of the picture. Figure 9b does not show as clear a difference between the sides for the ProbableAD group, and while the order remains roughly the same as that seen in the Control group, the ordering is less clear. Log-Rank test results yield significant differences (p < .001) in the distributions of time until words in the different quadrants are said for both groups.

Figures 9a and 9b about here

Figure 9a

*Elapsed time until first word in each quadrant produced for Control group*

A picture containing text, screenshot, line, plot

Description automatically generated

Figure 9b

*Elapsed time until first word in each quadrant produced for ProbableAD group*

A picture containing text, screenshot, line, diagram

Description automatically generated

Lastly, the Kaplan-Meier survival analyses were used to determine if the groups still differed significantly if the analyses fixed which quadrant was looked at first. The importance of controlling for this is to ensure that the differences we see in the elapsed time until words are produced from the bottom right, for example, is not caused by individuals who start in the top left and simply spend more time there. Figure 10a shows no substantial difference in the time until words are said in the top left (Log-Rank p = .01). Both groups' curves for that section tend to fall fairly quickly compared to the other plots, indicating that people tend to look there first. Moreover, the other plots show a significant difference (Log-Rank p < .001) between times until any of the words were said on the right side, with the bottom left being insignificant after a multiple testing correction (Log-Rank p = .03), due to the low relative number of people who started in that section. However, Figures 10b, 10c, and 10d show a strong visual difference between the Control and ProbableAD groups.

Figures 10a, 10b, 10c, 10d around here

Figure 10a

*Elapsed time until any quadrant 1 word was produced*

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure 10b

*Elapsed time until any quadrant 2 word was produced*

A picture containing text, screenshot, diagram, line

Description automatically generated

Figure 10c

*Elapsed time until any quadrant 3 word was produced*

A picture containing text, screenshot, diagram, line

Description automatically generated

Figure 10d

*Elapsed time until any quadrant 4 word was produced*

A picture containing text, screenshot, diagram, line

Description automatically generated

**Discussion**

Results of this study showed that for total number of core lexicon words used at least once, the Control and Probable AD groups differed significantly, but the MCI group did not differ significantly from either group. Survival curves showed that for all but two content words in the core lexicon list, more time elapsed before the ProbableAD participants said the word when compared with the Control group. For X (number?) content words, the MCI group was also significantly slower than the Control group, though the curves show those differences usually starting after about 50% of the group said the word. The exception to the general pattern was for the words, “cookie” and “jar”. These will be discussed further later. Compared to the Control group, the ProbableAD group took significantly longer to produce any content word from the top right, bottom left, and bottom right picture quadrants of the picture. A related analysis showed that the two groups also differed significantly in the time elapsed until they produced words from the different quadrants. Finally, controlling for which quadrant participants looked at first, results showed longer elapsed times until the ProbableAD group produced any words from all quadrants except the upper left, which is where both groups tended to look first.

The findings of significantly lower core lexicon scores in the ProbableAD group compared to controls is similar to reports of fewer information units in AD groups than control groups on the same task (Ahmed et al., 2013; Croisile et al., 1996). Their information unit scoring included synonyms such as “plate” for “dish”.

Results of these analyses will provide benchmarks for performance on a picture description task commonly used to elicit discourse in these populations. The findings will shed light on the lexical deficits encountered in MCI and Alzheimer's disease. The core lexicon analysis method is a straightforward, accessible tool for clinicians and researchers to use with this population for both early assessment and measurement of change in lexical skills over time.

This suggests that those with less severe cases of dementia are able to successfully read this text of the jar, and perhaps those in the MCI group are in particular eager to use those words early because it is an easier task for them than describing the rest of the image. We do still see the Probable AD curves differing from the others, likely because those with the most severe cases are still struggling to recall the word, even when it is in front of them.

An obvious direction for future work on this topic is to use these results in a novel dataset to predict group classification. The need for large, shared datasets cannot be overstated. Fraser et al. (2019) made a strong case for data sharing, and plenty of review articles have summarized the limited publicly available resources (

**Acknowledgments**

This work was funded by ?? <Brian – which grant to acknowledge?>. Original acquisition of the Pitt corpus was supported by NIA grants AG005133 and AG003705 to the University of Pittsburgh.

**References**

Ahmed, S., de Jager, C. A., Haigh, A. M., & Garrard, P. (2013). Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer’s disease. *Neuropsychology, 27*(1), 79-85.

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, *51*(6), 585-594.

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., & Calzà, L. (2018). Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?. *Frontiers in Aging Neuroscience*, *10*, 369

Bourgeois, M., Fried-Oken, M., & Rowland, C. (2010). AAC strategies and tools for persons with dementia. *The ASHA Leader*, *15*(3), 8-11.

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology*, *31*(10), 1105-1126.

Croisile, B., Ska, B., Brabant, M. J., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and language*, *53*(1), 1-19.

Dalton, S. G. H., Hubbard, H. I., & Richardson, J. D. (2020, January). Moving toward non-transcription based discourse analysis in stable and progressive aphasia. In *Seminars in speech and language* (Vol. 41, No. 01, pp. 032-044). Thieme Medical Publishers.

Dalton, S. G., Stark, B. C., Fromm, D., Apple, K., MacWhinney, B., Rensch, A., & Rowedder, M. (2022). Validation of an automated procedure for calculating core lexicon from transcripts. *Journal of Speech, Language, and Hearing Research*, *65*(8), 2996-3003.

de la Fuente Garcia, S., Ritchie, C. W., & Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer’s disease: a systematic review. Journal of Alzheimer's Disease, 78(4), 1547-1574.

Filiou, R. P., Bier, N., Slegers, A., Houzé, B., Belchior, P., & Brambati, S. M. (2020). Connected speech assessment in the early detection of Alzheimer’s disease and mild cognitive impairment: a scoping review. *Aphasiology*, *34*(6), 723-755.

Folstein, M. F., Folstein, S. E., & McHugh, P.R. (1975). Mini mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 188-199.

Fraser, K. C., Fors, K. L., & Kokkinakis, D. (2019). Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, *53*, 121-139.

Fraser, K. C., Linz, N., Lindsay, H., & König, A. (2019, June). The importance of sharing patient-generated clinical speech and language data. In *CLPsyc 2019-Sixth Workshop on Computational Linguistics and Clinical Psychology* (pp. 55-61). Association for Computational Linguistics.

Fraser, K. C., Lundholm Fors, K., Eckerström, M., Themistocleous, C., & Kokkinakis, D. (2018, May). Improving the sensitivity and specificity of MCI screening with linguistic information. In *LREC workshop: RaPID-2. Miyazaki, Japan*.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *BDAE- 3: Boston Diagnostic Aphasia Examination—Third edition*. Lippincott Williams & Wilkins.

Lanzi, A. M., Saylor, A. K., Fromm, D., Liu, H., MacWhinney, B., & Cohen, M. (2023). DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*.

MacWhinney B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.).Erlbaum.

Mueller, K. D., Kosik, R. L., Hermann, B. P., Johnson, S. C., & Turkstra, L. S. (2018). Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer’s prevention. *Frontiers in Aging Neuroscience, 9*, 437.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53*(4), 695–699.

Pekkala, S., Wiener, D., Himali, J. J., Beiser, A. S., Obler, L. K., Liu, Y., ... & Au, R. (2013). Lexical retrieval in discourse: An early indicator of Alzheimer’s dementia. *Clinical Linguistics & Phonetics*, *27*(12), 905-921.

Ross, G. W., Cummings, J. L., & Benson, D. F. (1990). Speech and language alterations in dementia syndromes: Characteristics and treatment. *Aphasiology*, *4*(4), 339-352

Stark, B. C., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., ... & Roberts, A. C. (2021). Spoken discourse assessment and analysis in aphasia: An international survey of current practices. *Journal of Speech, Language, and Hearing Research*, *64*(11), 4366-4389.

Taler, V., & Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Neuropsychology*, *30*(5), 501-556.

Thaler, F. & Gewald, H. (2021). Language characteristics supporting early diagnosis through machine learning – a literature review. *Health informatics-An International Journal: HIIJ* *10*(1), 5-23.

Vigo, I., Coelho, L., & Reis, S. (2022). Speech-and language-based classification of Alzheimer’s disease: A systematic review. *Bioengineering*, *9*(1), 27.

Yancheva, M., & Rudzicz, F. (2016, August). Vector-space topic models for detecting Alzheimer’s disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2337-2346).

**Appendix --** Cookie Theft Core Lexicon

a

and

be

\*boy

\*cookie

\*dish

\*dry

\*fall

\*flow

get

\*girl

he, him, his, himself

in

\*jar

little

\*mom, mommy, mother, mama, momma, ma

on

over

\*run

she, her, herself, hers

\*sink

\*stool

the

to

\*water

\*window

\* These were included in the list used to compute how long it took the participant to first say this word.

1. DementiaBank -- <https://dementia.talkbank.org/> -- is part of the TalkBank system, currently the world’s largest repository of shared databases for spoken language and is freely open to researchers, educators, and clinicians. [↑](#footnote-ref-1)