LoSST-AD:

* corpus includes 135 transcripts - public interviews with 20 famous
* figures, half of whom will eventually be diagnosed with AD. They filter the corpus to be with 99 transcripts. See table 2.
* The study focuses on vocabulary richness features such as Brunet index, hapax legomena, type: token ratio, and word frequency; word length, the number of words used once and twice, noun and adposition frequency, and uni and bigram repetitions.
* Assumption: if the speaker cannot access the correct word, a simpler word, usually of higher frequency and decreased length, tends to be used instead.
* The speakers with AD often struggle to find nouns, and therefore both noun and ad-position frequency can be affected.
* Each AD participant (10 total) is matched with a control participant (HC) based on demographic data.
* Data is aimed to be balance across sex, demographics, avg recording age. See Table 1.
* All interviews were manually transcribed excluding the speech of the

Interviewer and direct quotes such as song lyrics.

* The transcripts included filled pauses (such as “uh”), false starts, and stutter, but did not include speech tempo related features, such as the length of pauses.
* SpaCy and NLTK are the major tools used to extract features from the transcripts.
* Experiments: The first experiment involves comparing the earliest and latest recordings between the AD (Alzheimer's disease) and HC (healthy control) groups. The hypothesis is that based on previous literature, the change in the AD group should be more severe. They used two-tailed paired t-test for the features that were normally distributed and two-tailed Wilcoxon signed rank test for those are not. The second experiment focuses on investigating longitudinal changes in vocabulary richness concerning the time before diagnosis, they used simple linear regression and got statistically significant results with p-value lower than 0.05 for each one of the features.
* Results for exp1 shown in Table 3. More info in figure 3 in the paper.
* Results for exp2 shown in Figure 4.
* The corpus is available online at: <https://www.losst-ad.com>
* secondary data as feature not a bug. They highly recommend use of it, as it is already publicly and avoid discomfort from collecting directly from individuals experiencing cognitive decline.
* The main limitations of this study are the small size of the dataset and the lack of medical information, such as the year of diagnosis, stage or severity of the disease, or any co-existing conditions.

תמונה שמכילה טקסט, קבלה, גופן, צילום מסך

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.תמונה שמכילה טקסט, צילום מסך, גופן

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.תמונה שמכילה טקסט, גופן, צילום מסך, מסמך

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

תמונה שמכילה טקסט, תרשים, קו, מפה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.