LoSST-AD: A Longitudinal Corpus for Tracking Alzheimer's Disease Related Changes in Spontaneous Speech

Ulla Petti, Anna Korhonen
Language Technology Lab, University of Cambridge
9 West Road, CB3 9DP, UK
ump20@cam.ac.uk

Abstract
Language-based biomarkers have shown promising results in differentiating those with Alzheimer's disease (AD) diagnosis from healthy individuals, but the earliest changes in language are thought to start years or even decades before the diagnosis. Detecting these changes is critical to allow early interventions, but research into the earliest signs is challenging, as it requires large longitudinal datasets that are time-consuming and expensive to collect. There is a need for alternative methods for tracking longitudinal language change, including Natural Language Processing (NLP) and speech recognition technologies. We present a novel corpus that can enable this: a corpus of transcripts of public interviews with 20 famous figures, half of whom will eventually be diagnosed with AD, recorded over several decades. We evaluate the corpus by validating patterns of vocabulary richness changes known from literature, such as decline in noun frequency, word length, and several other features. We show that public data could be used to collect longitudinal datasets without causing extra stress for the participant, and that these data can adequately reflect longitudinal AD-related changes in vocabulary richness. Our corpus can provide a valuable starting point for the development of early detection tools and enhance our understanding of how AD affects language over time.

Keywords: Alzheimer's disease, cognitive decline, bioNLP, vocabulary richness

1. Introduction

Alzheimer's disease (AD) is the main cause for dementia – a condition which affects an estimate of 50 million people worldwide, with the number expected to triple by 2050 (WHO, 2023). While there is currently no approved cure for dementia, early diagnosis would allow to slow the progression of the disease, introduce lifestyle changes, plan for the future as a self-determining agent, as well as seek for support from relevant outside sources (Calza et al., 2015; Porteri et al., 2017; Alzheimer's Association, 2008). Considering the aging population and the developments in potential treatments (such as van Dyck et al., 2023), there is an eminent need for simple, non-invasive, and scalable tools that could aid clinicians with detecting the earliest signs of cognitive decline while causing the least stress to the patients (Fukuda et al., 2022). Such tools could be based in NLP and speech technologies because changes in language can reflect the earliest signs of cognitive decline and act as biomarkers for Alzheimer's (Mirheidari et al., 2017; Fraser et al., 2016; Satt et al., 2013). While changes in language can appear years or even decades before the diagnosis (Ringman et al., 2017; Grundman et al., 2006), there is a lack of studies focussing on the earliest AD-related language changes due to the limited availability of large-scale longitudinal datasets (Luz et al., 2021; López-de-Ipiña et al., 2018). While these datasets could be extremely valuable for both deeper linguistic understanding of Alzheimer's and the development of technologies for its early detection, they are time-consuming and expensive to collect. Also, the process has numerous ethical issues, contributing to data sparsity in this domain (Luz et al., 2021).

In this paper, we aim to tackle these issues by presenting a novel corpus of Longitudinal Spontaneous Speech Transcripts for tracking Azheimer's Disease related changes in language (LoSST-AD). The corpus consists of 135 public interviews recorded over several decades with 20 famous individuals, half of whom will eventually be diagnosed with AD. Unlike previous studies in the domain that have either focussed on a few individuals (such as Le et al., 2011; Berisha et al., 2015), used picture description task, or spanned over a shorter time period (such as Luz et al., 2021; Becker et al., 1994), LoSST-AD consists of several decades of transcribed public spontaneous speech data from a larger group of individuals, allowing us to examine longitudinal AD-related changes in language use.

We evaluate the corpus by validating the patterns of language change known from Alzheimer's literature, focussing on vocabulary richness. We show that such data can provide valuable insights into longitudinal language changes in AD, and help to develop non-invasive screening tools such as those based on NLP and speech technologies.

2. What Kind of Language Changes are we Expecting to Capture?

AD-related changes have been documented in various speech and language domains such as lexicon, semantics, syntax, discourse, and acoustics (Bayles, 1982; Lima et al., 2014; Gosztolya et al., 2019). Due to the scope of this paper and the nature of the data (varying audio quality, uncontrolled content due to secondary data, inconsistent turn-taking), we will focus on the changes in vocabulary richness and demonstrate that lexical diversity features can provide comprehensive results in tracking AD-related language change. Changes in vocabulary richness mostly fall under lexical-semantic domain, which in AD is affected by changes in cognitive function, semantic, procedural, and declarative memory (Dijkstra, 2004; Ullman, 2003;

LREC-COLING 2024, pages 10813–10821

Orange & Purves, 1996). These changes manifest in impoverished vocabulary and word finding difficulties. The two are connected – if the speaker cannot access the correct word, a simpler word, usually of higher frequency and decreased length, tends to be used instead (Saito & Takeda, 2001; Hodges et al., 1992). This results in producing less unique words, affecting type:token ratio and vocabulary richness indices (such as the Brunet index (Brunet, 1978)) (Guinn et al., 2014; Hernández-Dominiguez et al., 2016; Fang et al., 2017). The speakers with AD often struggle to find nouns, and therefore both noun and adposition frequency can be affected (Jarrold et al., 2014; Bucks et al., 2000; Ammar & Ayed, 2018). Other strategies to cope with not remembering a word include, for example, repeating the last word uttered (Croot et al., 2000), stuttering (Boyé et al., 2014), using fillers, and laughter (Sidnell & Stivers, 2012). Figure 1 illustrates the underlying AD-related problems that affect the lexical/semantic domain and vocabulary richness, as well as how these changes manifest in speech, and which respective language features could be extracted from the transcripts.

The aim of this paper is to 1) introduce and make available a new language resource – a longitudinal corpus of transcripts of interviews with public figures, half of whom will eventually be diagnosed with AD, and 2) evaluate the corpus by validating the patterns of AD-related vocabulary richness changes known from the literature. We report evaluation that shows that our new resource can provide a highly valuable starting point for the development of NLP and speech -based early detection tools for AD.

3. Corpus Creation

We collect publicly available TV and radio interviews from 20 famous individuals, half of whom will eventually be diagnosed with AD. The individuals with AD are identified based on internet searches, using websites such as the Wikipedia page 'Deaths from Alzheimer's disease'. Those with conditions potentially also affecting speech, such as Parkinson's disease, are excluded. The final selection of the 10 individuals is based on the number of publicly available interviews in English language on Youtube. Each AD participant is matched with a control participant (HC) based on demographic data, such as year of birth (within 5-year range), gender, place of residence or growing up, and occupation when possible. We identified and transcribed all publicly available interviews of the 20 individuals. A total of 135 interviews are included in the corpus. The number of interviews per speaker ranges from 3 to 25, with the earliest AD group interview being recorded 37 years before the diagnosis, and the latest 2 years after the diagnosis. See Figure 2 for the distribution of interviews over time, Table 1 for the participant group demographic information.

This research project was approved by the University of Cambridge Humanities and Social Sciences Research Ethics Committee. All transcripts and datasets – along with the datasheet with standardised data description (Gebru et al., 2021; Papakyriakopoulos et al., 2023) – are available online at: https://www.losst-ad.com/.

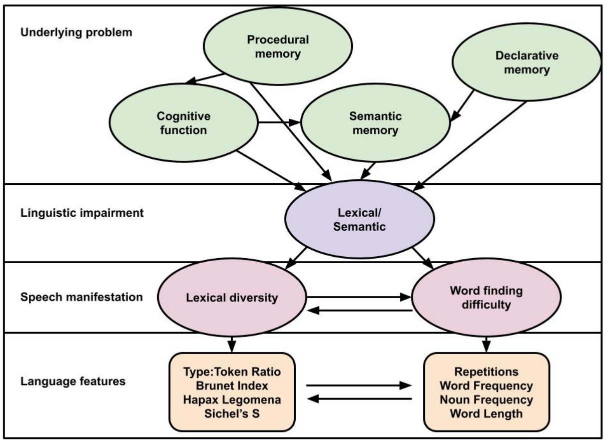

Figure 1: Changes in vocabulary richness in Alzheimer's disease.

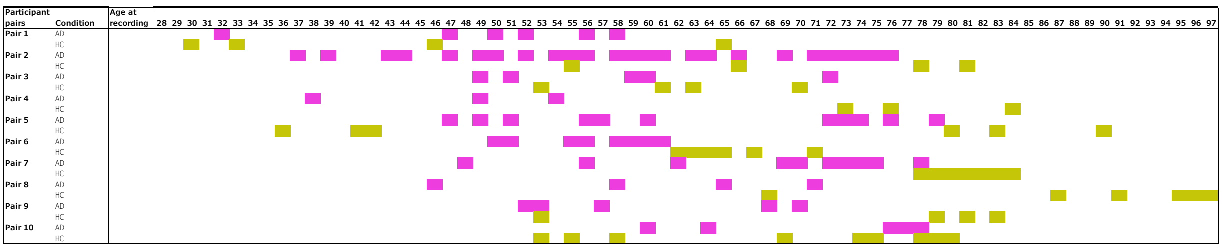

Figure 2: Distribution of interview recordings (*AD: Alzheimer's disease; HC: healthy control)*

	AD group (10 participants, 82 recordings)	**HC group** (10 participants, 53 recordings)
Average age over all recordings (years)	60 (SD: 11, range: 32-79)	70 (SD: 16, range: 30-97)
Average time before diagnosis (years)	13 (SD: 11, range: 37 years before diagnosis - 2 year after diagnosis)	-
Sex	5 female, 5 male	5 female, 5 male

Table 1: Participants' demographic information (*AD: Alzheimer's disease; HC: healthy control; SD: standard deviation)*

4. Data Processing

All interviews were manually transcribed. Speaker diarisation was applied, excluding the speech of the interviewer. Direct quotes, such as song lyrics, were also excluded from the transcripts to allow more accurate analysis of vocabulary richness in spontaneous speech. Transcripts were anonymised using the guidance from Saunders and colleagues' (2015) paper regarding people's names, places, cultural background, occupation, family relations and other identifying information.

Given the variation in audio quality (some interviews were recorded decades ago, some in noisy environments, using different settings and microphones) acoustic features were not analysed in the current study. 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. While the syntactic and semantic changes could also be explored based on this data, we have focussed specifically on vocabulary richness features in the current paper.

SpaCy and NLTK were used to automatically extract language features from every transcript. One transcript refers to one interview, and all features were extracted on interview level. The extracted vocabulary richness features include Brunet index, hapax legomena frequency, type:token ratio, word frequency (referring to higher use of more frequent words), word length, the number of words used once and twice, noun and adposition frequency, and uni- and bigram repetitions. All extracted features have been proposed to be highly informative by previous studies (Ammar & Ayed, 2018; Guinn et al., 2014; Hernández-Dominiguez et al., 2016; Yeung et al., 2021).

We controlled for text-length-sensitivity of the extracted feature values by conducting Pearson correlations between the feature value and transcript length. Since the majority of the extracted features were dependent on text length, we constructed a capped sub-corpus to minimise text length impact on the analysis. Building on the methods used to tackle text-length-sensitivity in Le and colleagues (2011), we capped all transcripts at the same length (based on word count) within each participant pair (the AD speaker and their matched control), keeping at least three transcripts per participant to allow for tracking longitudinal change. We aimed to keep as much speech data as possible, resulting in excluding the shortest samples from the capped corpus. The capped corpus consists of 99 transcripts in total. The lengths of both, the transcripts in the full and in the capped corpus, are given in Table 2.

Both corpora are available online at: https://www.losst-ad.com/. We have provided only the transcripts, and do not include links to the interviews or audio for ethical reasons.

Speaker pairs (AD and HC speaker)	**The length of transcripts in the full corpus in each pair. Average (range)**	**The cut-off points in the capped corpus in each pair**
Pair 1	880 (125-2561) words	417 words
Pair 2	1723 (260-6792) words	390 words
Pair 3	553 (183-1134) words	216 words
Pair 4	1110 (350-2848) words	350 words
Pair 5	1452 (159-4424) words	251 words
Pair 6	887 (141-2170) words	740 words
Pair 7	1510 (275-6385) words	574 words
Pair 8	3702 (197-8621) words	3065 words
Pair 9	1614 (67-5653) words	116 words
Pair 10	2693 (301-6327) words	3554 words

Table 2: The length of transcripts in the full corpus and in the capped corpus (in words) *(AD: Alzheimer's disease, HC: healthy control)*

5. Tracking Changes in Vocabulary Richness

We evaluate the corpus against what is known about the development of AD in the medical literature. We conduct two experiments. First, we compare the earliest and latest recordings across the AD and HC group, hypothesising that based on previous literature, the change in the AD group should be more severe. Second, we investigate longitudinal change in vocabulary richness in relation to the time before diagnosis. In both experiments, we use the capped corpus if the language feature of interest is text-length-sensitive, and the full-length corpus if the language feature is not text-length-sensitive.

5.1 Comparison of the Earliest and the Latest Samples

We created a subset of samples, consisting of the earliest and latest available recording from each AD participant. The control group samples were chosen by matching the ages of each participant as closely as possible to the respective AD group participant. When the same age range was not available, we tried to match the recording intervals. We compared the difference in vocabulary richness features in the

earliest samples to the latest recordings between the AD and the HC group.
To compare the values of the earliest and the latest samples, we used a two-tailed paired t-test for the features that were normally distributed, and a two-tailed Wilcoxon signed rank test for those that were not. The results indicate that the average noun frequency, word length and word frequency differed significantly between the earliest and latest sample in the AD group, but not in the HC group. The statistical details are provided in Table 3.

Figure 3 visualises the data and illustrates our findings: the first column shows the difference between the feature values in the earliest and latest recordings in the AD and the HC group; the second column compares the average change in the feature values between the first and last recording, and the third column shows how the feature values of each individual change from the earliest to the latest time point.

Feature	**AD early** (mean unless specified)	**AD late** (mean unless specified)	**AD group difference**	**HC early** (mean)	**HC late** (mean)	**HC group difference**
Noun frequency	0.130	0.098	Paired t-test p=0.028*	0.138	0.123	Paired t-test p=0.062
Word length	3.752	3.552	Paired t-test p=0.002**	3.852	3.833	Paired t-test p=0.832
Word frequency	12.621 (median)	12.963 (median)	Wilcoxon p=0.004**	12.828	12.759	Paired t-test p=0.466

Table 3: Statistical details of the comparison of noun frequency, word length, and word frequency values between the earliest and the latest recordings (*AD: Alzheimer's disease; HC: healthy control)*

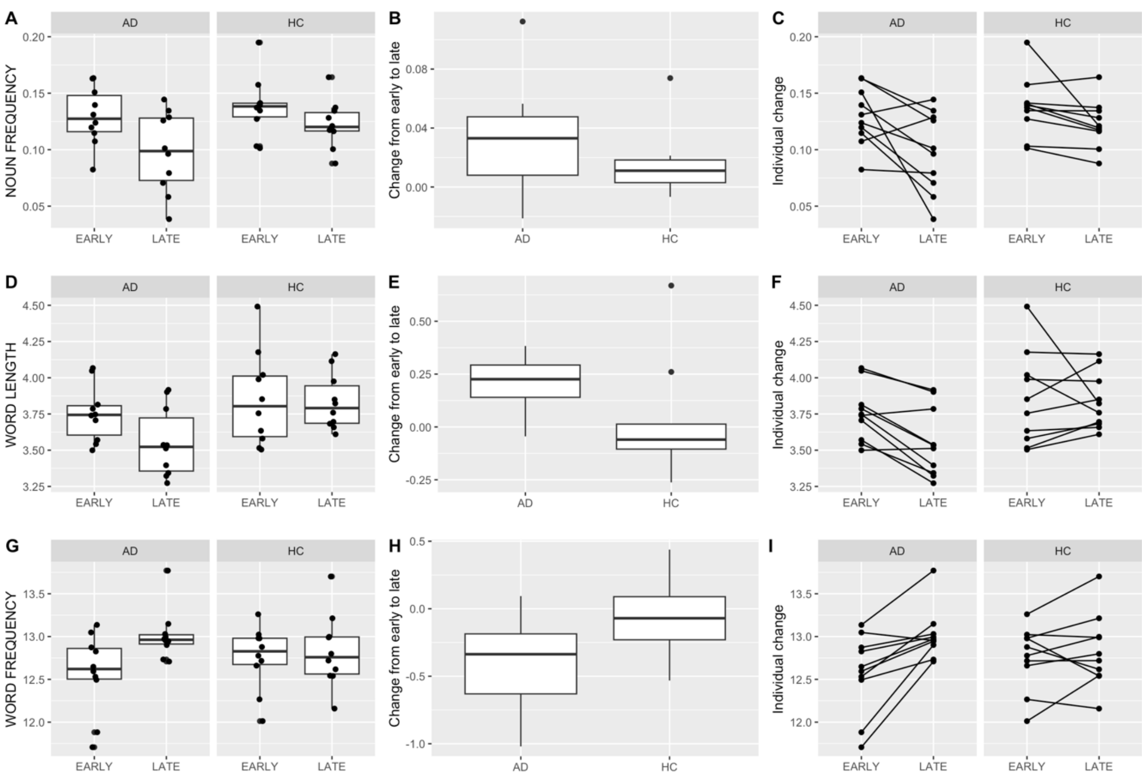

Figure 3: Comparison of noun frequency, word length, and word frequency values between the earliest and latest recordings (*AD: Alzheimer's disease; HC: healthy control)*

5.2 Longitudinal Change

In this experiment, we investigated whether longitudinal change in vocabulary richness in relation to the time before diagnosis can be detected from the corpus, using all the available recordings and time points. Time before diagnosis is established based on available public sources and media entries. As the control group participants do not have a date of diagnosis, their recordings are mapped to the respective AD participants based on age at any given time point.

A simple linear regression shows that the number of years before the diagnosis in the AD group was a significant predictor of noun frequency (p=0.022), hapax legomena (words used once) (p=0.030), words used once or twice (p=0.035), Brunet index (p=0.031), type:token ratio (p=0.023), adposition frequency (p=0.025), word frequency (Zipf p=0.0005, Subtl p=0.0004), and the interval of the uni- and bigram repetitions (p=0.003). Significant changes were not detected in the HC group.

Noun frequency, hapax legomena, words used once or twice, Brunet index, type:token ratio, and adposition frequency were extracted from the capped corpus, and word frequency measures and the interval of the uni- and bigram repetitions were extracted from the full corpus.

The change in vocabulary richness values in relation to the year before diagnosis is shown on Figure 4.

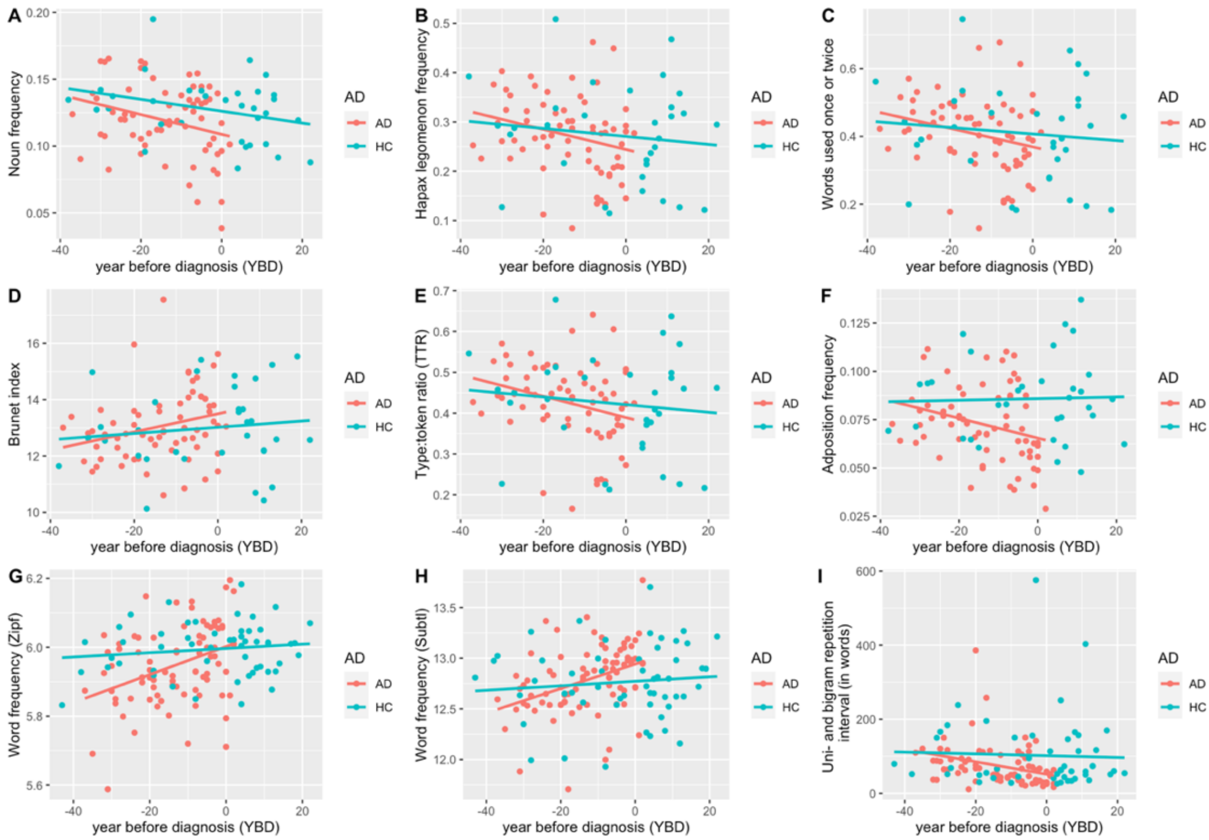

Figure 4: Change in vocabulary richness features in relation to the year before diagnosis (*AD: Alzheimer's disease; HC: healthy control)*

6. Discussion and Conclusion

We present a novel language resource - a longitudinal corpus of 135 spontaneous speech transcripts of public interviews with 20 famous individuals, half of whom will eventually be diagnosed with AD, recorded over several decades. This corpus could be highly valuable for research on AD as well as to train a system to automatically detect the risk for AD in speech. The corpus is available online at: https://www.losst-ad.com/.

We demonstrate that public data, the collection of which does not cause extra discomfort for the participants, can carry important information and has a great potential to contribute to developing language-based, fast, cheap, accessible, and non-invasive tools that could aid clinicians and help detect signs of AD early, as well as broaden our understanding of language change in AD in general.

Limited data availability is one of the most challenging issues in tracking language-based cognitive decline in AD, mostly due to data collection being time-consuming and expensive (Luz et al., 2021). We show that collecting secondary, already publicly available data can help tackle this issue and capture AD-related changes in speech comprehensively, demonstrating the potential of such approach, and encouraging the collection of more large-scale and controlled datasets to allow for more detailed analysis that could help understand the language changes in AD better.

Another aspect that contributes to data sparsity is the rise of various ethical questions when recording speakers with potential cognitive decline. One of the issues is participant burden – individuals experiencing cognitive decline can find it challenging to go through cognitive testing or to be asked to produce extensive amount of speech (López-de-Ipiña et al., 2018; Chien et al., 2019). Using secondary data as in the current study can help reduce participant burden, as no new recordings are conducted. This is not to say that ethical considerations can be overlooked when working with secondary data - researchers are still expected to handle data with care and make decisions with the best interest of the participants in mind. In the current study, we have aimed to protect the participants' identity by only providing anonymised transcripts, and not include the names of the speakers, links to the interviews or audio. Other datasets consisting of public speech and video recordings that have been used in medical research include, for example, The In-the-wild Speech Medical Corpus (Correia et al., (2021) and D-vlog (Yoon et al., 2022).

We conducted two experiments to evaluate the corpus and investigate its ability to track AD-related longitudinal changes in vocabulary richness. We found that significant changes in noun and adposition frequency, word length and frequency, unique words, Brunet index, and repetitions can be captured.

Noun frequency is expected to decline as AD gets more severe (Jarrold et al., 2014; Bucks et al., 2000). In line with previous literature, we found a significant difference in noun frequency between the earliest and the latest recording in the AD group – a change that does not manifest in the healthy group (Figure 3A-C). Similarly, looking at how noun frequency changes over time reveals an expected declining pattern closer to the time of diagnosis (Figure 4A). As adpositions in English tend to appear together with nouns, the decrease in adposition frequency is also detected as expected (Figure 4F). Supporting these findings, Guinn and colleagues (2014) and Ammar and Ayed (2018) conclude that noun frequency and adposition frequency are among the most informative features in distinguishing those with AD from healthy individuals. Our findings suggest that not only can these features contribute to distinguishing the speakers already diagnosed with AD from their healthy peers, but that the significant longitudinal change captured in the current study could indicate that these changes could potentially be captured years before the diagnosis, encouraging more large-scale data collection for in-depth analysis into when these changes manifest, and how reliably they can be captured.

Changes in word length have also been identified as one of the most informative manifestations of AD (Yeung et al., 2021). For example, Balagopalan and colleagues (2021) found that using shorter words can be associated with lower Mini-Mental State Examination (MMSE) (Folstein et al., 1975) scores. In line with these findings, our corpus also captures significantly shorter word length in the late recordings compared to the early ones in the speakers who will eventually be diagnosed with AD, but not in those who remain healthy (Figure 3D-F).

Previous literature also highlights the importance of word frequency (Yeung et al., 2021) and points out that people with AD tend to start using more general and frequent words instead of specific ones (for example, “animal” instead of “dog”) (Saito & Takeda, 2001; Hodges et al., 1992). The current study supports these findings: word frequency (measured using either Zipf or Subtl libraries) increases significantly closer to the time of diagnosis in the AD group (Figure 4G-H), and the average values in the early and the late samples also differ significantly (Figre 3G-I).

Hapax legomena, words used once and twice, Brunet index, and type:token ratio are all dependent on the number of unique words used. The decline in unique words, and the importance of these features in AD has been addressed by many (Guinn et al., 2014; Hernández-Dominiguez et al., 2016; Fang et al., 2017). In the current study, all four features reflect significant decrease in unique words, and therefore point to a decline in vocabulary richness (Figure 4B-E).

We also looked at how well our corpus captures repetition frequency (measured by the number of words between every occasion of a one- or two-word-repetition). These repetitions can reflect false starts

(Croot et al., 2000) or stutter (Boyé et al., 2014), both of which are common in AD. We found that the repetitions get significantly more frequent closer to the diagnosis (Figure 4I). In support, Guinn and colleagues (2014) and Ammar and Ayed (2018) propose that repetitions and related errors are among the most informative features of AD-related language changes.

The primary aim of our study was to demonstrate the usefulness of this type of corpus data. We emphasise that this study serves as an indication for future research and does not aim to generalise based on 2x10 individuals. Similarly, the type of data and the lack of medical information allows this study to only be descriptive in nature, and more in-depth medical information and expertise would be needed to argue for a causation between AD and the language changes.

Future work could collect more data with regular intervals to allow for more complex analysis and precise representation of longitudinal changes, explore the AD-related syntactic changes (such as shorter utterance length and lower depths of structure, less frequent use of reported speech and interpolated clauses, unfinished phrases closer to AD diagnosis) and semantic complexity (such as lower content density, use of information units, and cohesiveness, more frequent topic shifts, revisions, aborted phrases and indefinite words closer to AD diagnosis), and replicate the experiments with automatic transcription and anonymisation.

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. In the current study, the information related to the year of diagnosis was based on media entries, but it was not validated by a health professional, and might be untrue - for example, the public figures may not have disclosed their diagnosis immediately. Similarly, there is a lack of medical information about the control group speakers, and while their public information did not include AD-related entries, we did not have access to their validated health records. Unknown medical conditions or potential AD in the control could have had an impact on the language changes. For example, Figure 3 shows a HC group participant whose slope differs from the rest of the HC group and is more similar to the AD group participants. It is important to consider the potential impact of the lack of medical information when interpreting the results of the current study.

Additionally, the speech data was uncontrolled: using secondary data, it was not possible to control for the content of the questions, or how scripted the interviews were, contributing great variation in the data. Similarly, it was impossible to know if any other factors, such as intoxication, were affecting the participant's speech during the interview. The variability of the data should be considered when interpreting the results, especially the linear regression in Figure 4.

All in all, even with some limitations, the constructed corpus demonstrates that changes in vocabulary richness can be comprehensively captured using public data. These findings are promising and encourage future work in collecting large-scale datasets and developing spontaneous-speech-based tools for early AD detection.

7. Acknowledgments

This work was supported by the Economic and Social Research Council Cambridge Doctoral Training Partnership (DTP) grant number ES/P000738/1. We thank Dr Simon Baker for valuable discussions.

8. Ethics Statement

All data is collected from publicly available sources, such as Youtube and Wikipedia. Even though the participants were aware of their speech being recorded and the interviews being made public while giving the TV and radio interviews, we acknowledge that they may not have been aware of the possibility of their speech later being investigated for signs of cognitive decline. Since most speakers are deceased and their families difficult to reach, it was not possible to obtain consent from the speakers. With these considerations in mind, we take extra care to protect the speakers – we have anonymised the transcripts, we do not provide speakers' names, other identifiable information, or links to the original interviews. Using spontaneous speech samples to develop tools to detect early signs of cognitive decline promotes the use of friendly conversations over stressful cognitive testing, which is less burdensome to the participants with potential AD. This data should be used for research purposes, with the best interests of the speakers in mind. This research project was approved by the University of Cambridge Humanities and Social Sciences Research Ethics Committee.

9. Bibliographical References

Alzheimer's Association. (2008). 2008 Alzheimer's disease facts and figures. *Alzheimer's & Dementia,* 4(2).

Ammar, R.B. and Ayed, Y.B. (2018). Speech processing for early Alzheimer's disease diagnosis: machine learning based approach. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1-8. Aqaba, Jordan, October.

Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F. and Novikova, J. (2021). Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13, 635945.

Becker, J.T., Boiler, F., Lopez, O.L., Saxton, J. and 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.

Berisha, V., Wang, S., LaCross, A. and Liss, J. (2015). Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of presidents

Ronald Reagan and George Herbert Walker Bush. *Journal of Alzheimer's Disease*, 45(3):959-963.

Boyé, M., Tran, T.M. and Grabar, N. (2014). NLP-oriented contrastive study of linguistic productions of Alzheimer's and control people. In *Advances in Natural Language Processing: 9th International Conference on NLP (PolTAL 2014),* pages 412-424, Warsaw, Poland, September. Springer International Publishing.

Brunet, É. (1978). *Le vocabulaire de Jean Giraudoux structure et évolution.* Slatkine.

Bucks, R.S., Singh, S., Cuerden, J.M. and Wilcock, G.K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1): 71-91.

Calzà, L., Beltrami, D., Gagliardi, G., Ghidoni, E., Marcello, N., Rossini-Favretti, R. and Tamburini, F. (2015). Should we screen for cognitive decline and dementia?. *Maturitas*, 82(1):28-35.

Chien, Y.W., Hong, S.Y., Cheah, W.T., Yao, L.H., Chang, Y.L. and Fu, L.C. (2019). An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network. *Scientific Reports*, 9(1):195-197.

Correia, J., Teixeira, F., Botelho, C., Trancoso, I., and Raj, B. (2021). The in-the-wild speech medical corpus. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6973-6977). IEEE.

Croot, K., Hodges, J.R., Xuereb, J. and Patterson, K. (2000). Phonological and articulatory impairment in Alzheimer's disease: a case series. *Brain and language*, 75(2):277-309.

Dijkstra, K., Bourgeois, M.S., Allen, R.S. and Burgio, L.D. (2004). Conversational coherence: Discourse analysis of older adults with and without dementia. *Journal of Neurolinguistics*, 17(4):263-283.

Fang, C., Janwattanapong, P., Martin, H., Cabrerizo, M., Barreto, A., Loewenstein, D., Duara, R. and Adjouadi, M. (2017). Computerized neuropsychological assessment in mild cognitive impairment based on natural language processing-oriented feature extraction. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 543-546, Kansas City, MO, USA, November. IEEE.

Folstein, M.F., Folstein, S.E., and McHugh S.C. (1975). Mini-Mental State Exam. *Journal of Psychiatric Research*, *12*(3), 196-198.

Fraser, K.C., Meltzer, J.A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407-422.

Fukuda, M., Nishimura, R., Umezawa, M., Yamamoto, K., Iribe, Y. and Kitaoka, N. (2022). Elderly Conversational Speech Corpus with Cognitive Impairment Test and Pilot Dementia Detection Experiment Using Acoustic Characteristics of Speech in Japanese Dialects. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022),* pages 1016-1022), Marseille, France, June. European Language Resource Association (ELRA).

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D. and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86-92.

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J. and Hoffmann, I., 2019. Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language*, *53*:181-197.

Grundman, M., Petersen, R.C., Bennett, D.A., Feldman, H.H., Salloway, S., Visser, P.J., Thal, L.J., Schenk, D., Khachaturian, Z. and Thies, W. (2006). Alzheimer's association research roundtable meeting on mild cognitive impairment: what have we learned?. *Alzheimer's & Dementia*, 2(3):220-233.

Guinn, C., Singer, B. and Habash, A. (2014). A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. In *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, pages 98-103. Orlando, Florida, USA. IEEE

Hernández-Domínguez L., Ratté S., Sierra-Martínez G. and Roche-Bergua A. (2018). Computer- based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 10:260- 268.

Hodges, J.R., Patterson, K., Oxbury, S., and Funnell, E. (1992). Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain*, 115(6):1783-1806.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L. and Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality,* pages 27-37. Baltimore, Maryland, USA, June. Association for Computational Linguistics (ACL).

Le, X., Lancashire, I., Hirst, G. and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and linguistic computing*, 26(4):435-461.

Lima, T.M., Brandão, L., Parente, M.A.D.M.P., and Peña-Casanova, J. (2014). Alzheimer's disease: cognition and picture-based narrative discourse. *Revista CEFAC*, 16:1168-1177.

Lopez-de-Ipina, K., Martinez-de-Lizarduy, U., Calvo, P. M., Mekyska, J., Beitia, B., Barroso, N., Estanga, A., Tainta, M. and Ecay-Torres, M. (2018). Advances on automatic speech analysis for early detection of Alzheimer disease: a non-linear multi-task approach. *Current Alzheimer Research*, 15(2):139-148.

Luz, S., Haider, F., de la Fuente, S., Fromm, D. and MacWhinney, B. (2021). Detecting cognitive decline using speech only: The ADReSSo Challenge. In *Interspeech 2021.* Brno, Czech Republic, August. International Speech Communication Association (ISCA).

Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M. and Christensen, H. (2017). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, 58(2): 373-387.

Orange, J.B. and Purves, B. (1996). Conversational discourse and cognitive impairment: Implications for Alzheimer's disease. *Journal of Speech Language Pathology and Audiology*, *20*:139-139.

Papakyriakopoulos, O., Choi, A.S.G., Thong, W., Zhao, D., Andrews, J., Bourke, R., Xiang, A. and Koenecke, A. (2023). Augmented Datasheets for Speech Datasets and Ethical Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2023)*, pages 881-904. Chicago, USA, June. ACM.

Porteri, C., Albanese, E., Scerri, C., Carrillo, M. C., Snyder, H. M., Martensson, B., Baker, M., Giacobini, E., Boccardi, M., Winblad, B. and Frisoni, G.B. (2017). The biomarker-based diagnosis of Alzheimer's disease. 1—ethical and societal issues. *Neurobiology of aging*, 52:132-140.

Ringman, J.M. (2017). Update on Alzheimer’s and the Dementias: Introduction. *Neurologic clinics* 35(2): 171-4.

Saito, A. and Takeda, K. (2001). Semantic cueing effects on word retrieval in aphasic patients with lexical retrieval deficit. *Brain and language*, 77(1): 1-9.

Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A. and Tsolaki, M. (2013). Evaluation of speech-based protocol for detection of early-stage dementia. In *Interspeech (2013),* pages 1692-1696. Lyon, France, August. International Speech Communication Association (ISCA).

Saunders, B., Kitzinger, J., & Kitzinger, C. (2015). Anonymising interview data: Challenges and compromise in practice. *Qualitative research*, *15*(5), 616-632.

Sidnell, J. and Stivers, T. (Eds.). (2012). *The handbook of conversation analysis*. John Wiley & Sons.

Ullman, M.T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, 92(1-2):231-270.

Van Dyck, C.H., Swanson, C.J., Aisen, P., Bateman, R.J., Chen, C., Gee, M., Kanekiyo, M., Li, D., Reyderman, L., Cohen, S., and Iwatsubo, T. (2023). Lecanemab in early Alzheimer’s disease. *New England Journal of Medicine*, 388(1):9-21.

World Health Organization. (2023). Risk re- duction of cognitive decline and dementia. https://www.who.int/publications/i/item/risk-reduction-of-cognitive-decline-and-dementia. Last accessed 9 October 2023.

Yeung, A., Iaboni, A., Rochon, E., Lavoie, M., Santiago, C., Yancheva, M., Novikova, J., Xu, M., Robin, J., Kaufman, L.D. and Mostafa, F., (2021). Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer’s dementia. *Alzheimer's research & therapy*, 13(1):109.

Yoon, J., Kang, C., Kim, S., and Han, J. (2022). D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12226-12234.