Two-level SVM model with language markers for (early) detection of Alzheimer's Disease

Anonymous EMNLP submission

Abstract

This study presents a novel two-level SVM (Support Vector Machine) model for the automatic (early) detection of Alzheimer's Disease (AD) using language markers that are independent of lexical semantics. We avoid lexical semantic features because they are subject to high individual variation, thus limiting their predictive power for unseen data. Instead, we focus on morphosyntactic, syntactic, and sentence-level features, which are more stable and potentially allow for easier generalization of the model to other datasets, languages, and individuals. We constructed SVMs at both the sentence level and the subject level, applying language features extracted from automatically parsed transcriptions from the Pitt and Delaware corpora in DementiaBank. Our model demonstrated that the subject-level SVM significantly improved classification accuracy. The model yields high performance across all evaluation metrics on the test set for both AD and Mild Cognitive Impairment statuses.

1 Introduction

Language markers have been used as an inexpensive, non-invasive, accessible, and fast test for the early detection of Alzheimer's Disease (AD) (Ostrand and Gunstad 2021, Vigo et al. 2022; see Luz et al. 2021a for an overview of relevant studies). This approach enables the creation of platforms such as chatbot applications for identifying AD patients (e.g., de Arriba-Pérez et al. 2023, BT and Chen 2024), potentially leading to treatments that can preserve the cognitive functions of AD patients for a longer time (Stern 2006, Lautenschlager et al. 2008).

Previous studies have shown that integrating different language markers with machine learning leads to superior performance (Luz et al. 2021a). Recently, there has been a special focus on extracting prosodic and phonetic features for prediction purposes (Szatloczki et al. 2015, König et al. 2015), with the best models usually utilizing subject-related information (e.g., Sadeghian et al. 2021, Mahajan and Baths 2021). However, data collection methods, such as those used in chatbot applications for initial filtering purposes, may not always be able to gather subject demographic information such as age, education, gender, and language background. To make these applications most accessible for data collection and to leverage possible storage space limitations—especially when the application is used for a large population—text information may be the most accessible format (e.g., Snowdon 1997). A form of text that is readily available might be the (auto) transcription of the participants' speech.

In this study, we are searching for language markers for the detection of AD and Mild Cognitive Impairment (MCI), a major precursor of AD (Rosenberg et al. 2013). These language markers should not require extensive data collection and avoid collecting sensitive personally identifiable information. Consequently, some of the previously developed models may not be applicable to this particular requirement without significant adjustment. In this study, we will minimize the information from the potential patients to transcription of speech available in DementiaBank ((Lanzi et al., 2023)). We provide modeling results with subjective information that is available from DementiaBank for the purpose of comparison with previous studies. In addition, for models that were built off lexical-related language features, targeting the specific words that have been used (or any features that are directly determined by the word forms), will likely give rise to high variability in prediction (Antonsson et al. 2021). This pitfall can be concealed when applying classifier models, such as Support Vector Machines (SVMs), to train and test set that are not split according to the subjects but according to the data points: the model may capture some individual-level lexical use preference instead of a linguistic pattern that is generalizable to other subjects (Hoang et al., 2023).

In sum, this study intends to provide effective language markers under the restriction of data format and information type, by achieving the following three goals: (i) the language indices are exclusively extracted from parsed text transcriptions; and (ii) we focus on syntactic indices which are potentially more stable properties across languages, (iii) constructing a model with high predictive power without resorting to subject demographic information. The first two goals are challenging as previous studies have reported that speech-related features give better results than text-related features (He et al. 2023). In addition, it has been observed that lexical semantics and verbal fluency are affected in the initial stages of AD, whereas syntax is more preserved (Kempler et al. 1987). Conversely, more recent studies also reveal syntactic simplification in AD patients (Kemper et al. 2001b), which are more pronounced in their written responses (Croisile et al. 1996).Individuals who have a lower score on grammatical complexity would more likely develop dementia later in their lives (Kemper et al., 2001a). The third goal is important for promoting the widespread application of computational methods in filtering tests for AD.

In this study, we track lexical-independent morphosyntactic and syntactic features, along with surprisal values extracted from large language models. We extracted morphosyntactic and syntactic features from transcription texts that are parsed by a Universal Dependency parser, i.e., UDPipe (Zeman et al. 2023, Straka et al. 2016). Generally, the UDPipe parser can be applied to transcriptions and has the potential to obtain more stable language markers across speakers, including those from different language backgrounds. In past studies, researchers have suggested syntactic changes in AD and MCI, including syntactic simplification, elliptical and segmental sentences, phrase repetition, phrasing selection problems, and verb agreement errors (Crossley et al. 2007, Sajjadi et al. 2012, Eyigoz et al. 2020, Chapin et al. 2022, Varlokosta et al. 2024). Based on this foundation, we listed language features derived from syntactic-level changes in dementia after a comprehensive review of existing literature. These language features are integrated with machine learning models. With the selected morphosyntactic and syntactic features, this integration aims to analyze the diagnostic accuracy of predicting potential AD and MCI patients. The analysis can help researchers in the process of data collection and the early detection of AD.

For data structuring, our project collects sentences from the DementiaBank database, the Pitt (Becker et al., 1994) and Delaware (DE) corpora (Lanzi et al., 2023), and uses the automatic analysis tool UDPipe for universal dependencies parsing (Zeman et al. 2023, Straka et al. 2016). Our goal is to compile a set of relevant syntactic and sentence processing features and test their efficiency and accuracy in the early automatic detection of AD.

2 Data and Models

2.1 Data sets and language features extraction

We included data of 232 subjects from the Pitt corpus, with 66 healthy controls (HCs), 11 MCI patients, and 147 AD patients; with a few subjects appearing in more than one category because of their health status changes. Only subjects who have completed the Cookie Theft task were included. Given that many of the subjects are retested in one year or longer, we consider the results from each test as an independent subject. This makes the total number of subjects 400, with 149 HCs, 21 MCI patients, and 220 AD patients. Given that the amount of MCI patients is small, we excluded them from the models. For the DE corpus, we extracted data from 73 subjects, with 26 HCs and 47 MCI patients.

We removed special annotation texts and extracted all sentences produced by the subjects in Pitt and DE corpora. Specifically, we extracted sentences from the Cookie Theft task in Pitt and the multiple picture descriptions in DE. Figure 1 and 2 show the number of sentences per subject is relatively small but stable in Pitt, whereas each subject in DE has more sentences, due to the reason that the data were collected with multiple picture description tasks.

We collected a list of language features by reviewing existing literature with a focus on morphosyntactic, syntactic, and phrase/sentence-level features. Previous studies have shown that AD patients exhibit sentence processing difficulties at the syntax level as well as memory-related semantic deficits (e.g., naming difficulties) (Chapin et al. 2022, Hernández-Domínguez et al. 2018, Eyigoz et al. 2020, Ostrand and Gunstad 2021). These features are obtained by running Python scripts (see Appendix) over the transcriptions from the Pitt

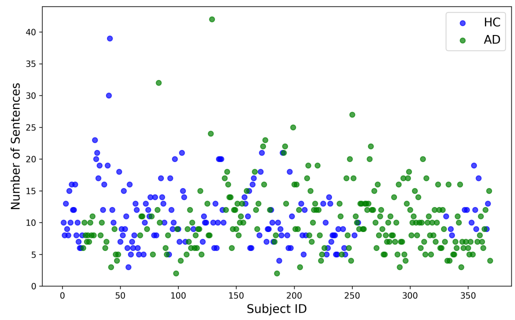

Figure 1: Number of sentences from each subject in Pitt

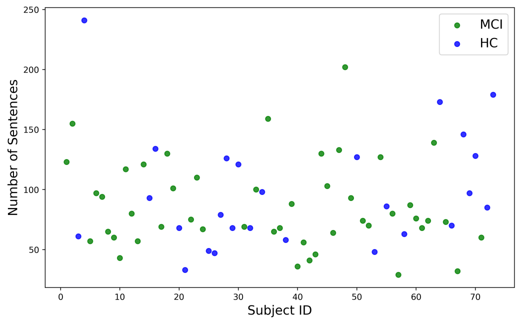

Figure 2: Number of sentences from each subject in DE

and DE corpora, which are parsed by a Universal Dependency parser provided by the UDPipe.

Examples of syntactic features include the types and amount of clauses, the types and amount of adjuncts, the amount of tense and aspect markers, transitive and intransitive verbs, repeating articles, etc. The morphosyntactic features, namely, the amount of derivational and inflectional morphemes, are based on a morphological analysis using the English morpheme database from the *unimorph* package (Kirov et al. 2018). For sentence processing features, we obtained the surprisal values for each word in each sentence (as annotated by UDPipe) from the full GPT-2 model in Hugging Face. With the surprisal values, we calculated the mean, minimum, and maximum surprisal for each sentence. However, including all these surprisal values in the SVM model led to high multicollinearity (measured by the Variance Inflation Factor, VIF). Therefore, we included only the minimum and maximum surprisal values, which produce the best outcome with their VIFs kept below 5. Overall, 40 syntactic/sentence-related features are included in the models.

To ensure that the language features are extracted as expected, we constructed a gold standard file with 100 selected sentences from our data sources. Of these 100 samples, 50 are identified as AD patients, and the remaining 50 from individuals with MCI. This gold standard serves as a reference to determine the accuracy of the Python code used for extracting and counting language features from universal dependency annotations.

2.2 Data preprocessing

To preprocess the data, sentences with a total token number(including punctuation) less than or equal to 4 were excluded. Additionally, subjects who were diagnosed with MCI were excluded from the Pitt corpus due to its small sample size. Lastly, missing data in any language features, 0.4% from Pitt and 0.2% from DE, were excluded from the following modeling.

We standardized and scaled the extracted numeric features using the StandardScaler function from Scikit-learn and transformed the categorical features (subject demographic information) with one-hot encoding. The data was split into train and test sets by subject IDs to ensure the train and test sets were not from the same subjects. This method also excludes the possibility that the model is learning particular speech patterns of individual subjects (e.g., the use of particular lexical items or linguistic expressions). Some of the data collected are from individual subjects at different time periods. On average, there was at least a one-year gap between two data collection processes. In the model, we treat the data collected at different time periods as data collected from different individuals. As a result, our model will make independent predictions for a potential patient even if the subject takes the test every year. Furthermore, combining the data collected at different periods together for the subject-level SVM produces a better performance.

2.3 A two-level SVM model

Previous studies have suggested that SVM models are among the machine learning methods that yield the best evaluation matrices in predicting the diagnosis of AD (Antonsson et al. 2021, Balagopalan et al. 2021, Luz et al. 2021a; see Vigo et al. 2022 for a review). Below we explore the integration of SVM models: (1) a sentence-level SVM model that predicts the diagnostic label for each sentence based on the language features of the sentence; (2) since each subject produced multiple sentences in the corpora, a subject-level SVM model taking the percentage of a particular predicted label from the sentence-level model as input and the diagnostic label as the output to predict the final diagnostic; and (3) a subject-level SVM that has additional subject-related information such as age, gender, education years as input, whose performance will be more comparable to the previous models.

Two levels of the SVM model were built to estimate the prediction of a subject's diagnosis. The first is at the sentence level and the second is at the subject level. The separation of the sentence and subject level is quite unique among the previous models we have seen, with (Hoang et al., 2023) as a notable exception. Sentence-based organization of the data is ideal for automatic parsing with Universal Dependency parser. We extract the features from each sentence in the dialogue between the patient and the interviewer. Occasionally, the dialogue was longer than one sentence, in which case UDPipe will process them as a sentence with clauses connected with dependency relations.

For each sentence, the sentence-level SVM model outputs a diagnosis prediction. With regard to prediction, the output is expected to involve considerable noise because not every sentence is informative for the diagnostics. Both the patients and HCs may produce similar sentences. To reduce the effect of noise, we built a subject-level SVM for the diagnosis prediction of each subject by taking the percentage of a particular prediction label from the sentence-level SVM as input and their diagnostic as the target to train on. This level of SVM is important and is novel as previous studies either built a single SVM for each subject, or used a gauging technique to explore which percentage level may serve as a threshold for the final diagnostic. At the subject level, an important advantage of the SVM model is that it can identify less informative, noise sentences and exclude them when making its prediction. By adjusting the hyperparameters (C and gamma), the model determines the amount of outliers to be excluded at the sentence-level. We identify the optimal hyperparameters using grid search. We demonstrate that the subject-level SVM yields superior performance, which indicates that the SVM model has identified a refined hyperplane such that some level of the noise carried from the sentence-level SVM can be correctly identified and handled properly at the subject-level SVM model.

2.4 Model training

The training and testing of the model was carried out with the *Scikit-learn* package in python (Pedregosa et al. 2011). We split the data into train and test sets, with the test set comprising data from 20% of the subjects in both corpora. The split was based on subject IDs rather than individual data points to prevent the model from simply learning particular subjects' patterns of language use, and instead, focus on predicting the occurrence of AD and MCI in unseen subjects. Below, we present the performance of the SVM models on the test set.

To ensure that the model does not overfit the data, K-Fold Cross Validation, with K set to 10, was applied to the train set to generate an average evaluation metrics matrix to mitigate the effects produced by some potential variations due to sampling bias. In addition, to avoid multicollinearity, we examined the VIF of each linguistic feature and excluded linguistic features that have high VIF values. For example, mean surprisal was not included in the models because it highly correlates with both maximum and minimum surprisal. Abnormally superior performance was observed with the sentence-level SVMs if we include the mean surprisal in addition to maximum and minimum suprisals. The VIF of these three features all exceeded 5, signaling multicollinearity issues. In addition, we added minor noise to the data from a Gaussian distribution with the mean = 0 and standard deviation = 0.01 in each of the models to test whether the performance of the models will be significantly altered. If the model has a severe multicollinearity problem, small noise may change the performance significantly. No alarming signs of multicollinearity were observed.

Lastly, the model was tested with varying random state values (100 random states in total), for data splitting between the train and test set, to determine the average evaluation matrix across all random states. We report the average matrices across all random states in Table 1.

A special challenge in modeling the DE corpus is the imbalance between the data from patients and HCs. In this corpus, the number of MCI subjects is almost twice that of HCs. Additionally, the number of sentences produced by each subject varies dramatically (see Figure 2). Together with the small sample size, the imbalanced data can significantly deteriorate the model's performance, leading to a specificity value close to 0. To overcome this limita-

		Pitt Corpus		DE Corpus	
SVM level	**Score**	**w/o subj-info**	**w/ subj-info**	**w/o subj-info**	**w/ subj-info**
Sentence	**F1**	0.64 ± 0.03	0.71 ± 0.04	0.54 ± 0.02	0.61 ± 0.10
	Precision	0.65 ± 0.03	0.72 ± 0.04	0.62 ± 0.11	0.73 ± 0.14
	Recall	0.64 ± 0.03	0.71 ± 0.04	0.52 ± 0.03	0.66 ± 0.09
	Accuracy	0.64 ± 0.03	0.71 ± 0.04	0.52 ± 0.03	0.66 ± 0.09
	Specificity	0.58 ± 0.06	0.73 ± 0.10	0.57 ± 0.07	0.84 ± 0.11
Subject	**F1**	0.90 ± 0.18	0.92 ± 0.05	0.82 ± 0.10	0.68 ± 0.14
	Precision	0.91 ± 0.14	0.92 ± 0.04	0.85 ± 0.05	0.76 ± 0.14
	Recall	0.90 ± 0.16	0.92 ± 0.05	0.82 ± 0.11	0.68 ± 0.12
	Accuracy	0.90 ± 0.16	0.92 ± 0.05	0.82 ± 0.11	0.68 ± 0.12
	Specificity	0.84 ± 0.34	0.89 ± 0.09	0.80 ± 0.18	0.84 ± 0.18

Table 1: Mean and standard deviation of evaluation matrices across random states from the sentence- and subject-level SVM models with or without subject demographic information.

tion, we applied several methods to minimize the effects of an imbalanced dataset. First, we identified the subjects who produced more than one standard deviation (=41) above the mean (=91). Using this threshold, we randomly selected 132 (=41+91) sentences from sentences that each subject produced. In addition, we applied SMOTE (Synthetic Minority Over-sampling Technique, Chawla et al. 2002) as implemented in the *imbalanced-learn* package (LemaÃŽtre et al. 2017) for training to oversample the minority. Finally, we applied a balanced scoring metric for grid search. Using grid search, we determine the optimal hyperparameters that prioritize balanced evaluation metrics (accuracy, F1, precision, and recall scores) rather than individual measurements. We found that if we prioritize accuracy, the models yield a higher accuracy level, yet with a very low specificity value, indicating that the report of full evaluation metrics, including specificity, is necessary for a comprehensive assessment of the model's performance.

To compare with the results from previous studies (e.g., Luz et al. 2021b, Luz et al. 2021a) that integrated subject information from Pitt into the model, we also included modeling results with the subject demographic information in both the sentence-level and subject-level SVMs. The subject demographic information includes age, gender, race, and years of education.

2.5 Results

Table 1 displays the evaluation metrics for the sentence-level and subject-level SVM models applied to the Pitt and DE corpora. The models were evaluated both with and without the inclusion of subject demographic information, providing the models' performance under different conditions.

At the sentence level, both the Pitt and DE corpora achieve relatively lower performance. This outcome is anticipated as both patients and HC are likely to produce normal sentences that do not show any signals of morphosyntactic or syntactic deficits. The Pitt corpus yields higher performance compared to DE. The inclusion of subject demographic information resulted in an increase in evaluation metrics on the sentence-level SVMs.

Crucially, at the subject level, the SVM models demonstrate a significant improvement in their prediction performance. This underscores the effectiveness of integrating a higher-level SVM model based on the sentence-level predictions. Both the Pitt and DE corpora showed remarkably high metrics, with the Pitt corpus's SVM model accuracy scores reaching up to 90%, and the DE corpus's accuracy scores reaching 82%. In particular, the high specificity values (84% for Pitt, 80% for DE) highlight the model's ability to reduce false positives. Although the sentence-level model performed better with subject demographic information, this improvement was not as significant at the subject level. The Pitt corpus showed a higher performance with more consistency in comparison to the DE corpus.

2.6 Discussion

In this study, we track a small number of completely automatically extracted morphosyntactic, syntactic, and surprisal features. These set of language features are promising stable features applicable to different languages and different individual speakers, and they are available when data is restricted to the text format. In addition, a smaller number of features reduces the risk of overfitting and multicollinearity (Martinc and Pollak, 2020), and the current study used various methods such as keeping VIF low and adding noise to detect and prevent overfitting and multicollinearity.

With these features, we built a two-level SVM model to predict the diagnosis of MCI and AD subjects using corpora from Pitt and DE. Given its longer history, previous studies have examined Pitt to a greater extent, achieving 70-93% accuracy (e.g., Di Palo and Parde 2019, Mahajan and Baths 2021) using diagnostics based on blood-based biomarkers and imaging methods (Chen et al. 2021, Chávez-Fumagalli et al. 2021). Our results reach comparable performance with, or even outperform, these benchmarks, achieving a balanced high-profile evaluation matrix without using lexical-semantic features and subject demographic information.

The two-level SVM model is the key to improving the performance of predictions with observed significance at the subject level. On the sentence level, it is difficult to distinguish between HCs and patients due to the high level of noise in the data. Therefore, the sentence-level SVM model will not achieve a high-level performance. An SVM model that performs well solely on sentence level could be prone to overfitting. With the added layer of the subject level, the SVM model can gauge the percentage of misclassification of sentence-level models and ensure that a threshold must be met for a subject to be classified as a patient. Furthermore, with the added information from the subject level, the model can learn about this threshold according to the data.

At the subject level, we did not observe a consistent significant difference in performance with or without subject demographic information for both corpora. This result is unexpected because adding subject demographic information on sentence level improved the predictions. Since this result is observed for both corpora, it indicates that the subject-level SVM training on the features extracted is sufficient to make predictions for patients. The inclusion of subject demographic information is no longer necessary to achieve high performance.

The differences in the performance between the Pitt and DE models in both the sentence-level and subject-level SVMs are probably due to several reasons. The detection of MCI is essentially more difficult than AD (Luz et al., 2021a). In addition, the Pitt corpus includes more subjects. In the DE model, although the confusion matrix shows a high performance, the number of subjects in the test set is small (see the Appendix). A much more prominent problem of imbalance was detected with the DE corpus. If none of the previously mentioned methods were applied to properly handle the imbalance problem, the model returned a specificity approaching 0, indicating a lack of detection of true negatives and false positives. Additionally, there may be a side effect of repetitive tests for Pitt, as Pitt includes data from multiple tests for an individual subject: this could result in the model learning the patterns of language use for individual subjects.

Overall, these findings suggest that the subject-level SVM can significantly improve the performance of the model for effectively distinguishing HCs from AD and MCI patients. This is particularly valuable as it makes it possible to collect large-scale data free of privacy concerns due to the collection, storage, and use of sensitive personally identifiable information.

3 Conclusions

In this study, we built a two-level SVM model trained on a small set of morphosyntactic, syntactic, and surprisal features extracted from transcriptions. This model achieves high performance across all evaluation metrics, especially for the Pitt corpus. The subject-level SVM has demonstrated its capacity to significantly improve the evaluation metrics for both the Pitt and DE corpora. Crucially, with the two-level SVM model, the inclusion of subject demographic information becomes unnecessary and does not contribute to further improvement of the model. This paves the way for large-scale implementation of the NLP-based model for effective automatic AD screening tests, with data collection requiring only a few dozen sentences.

Limitations

One of the limitations of the current study is that the data we are analyzing is not collected from a chatbot application, although our goal is to extend the model for the analysis of such data in the future. The transcriptions we analyzed are provided by DementiaBank, which has been manually checked. This ensures transcription's accuracy and therefore increases the model's potential to reach high performance. For chatbot applications, auto-transcription may involve more errors. It is to be evaluated how the Universal Dependency parsing and surprisal calculation will be affected by inaccurate transcriptions. Fortunately, the recent chatbot application built in ChatGPT 4.0 achieves a high-level accuracy. Our next step is to implement auto-transcription from DementiaBank audios and test the stability of our model.

Although it is one of our long-term goals, another limitation of this study is the current method has not been tested with data collected from bilingual speakers and speakers of other languages. Cross-language data is important for the robust and stability assessment of the method presented in this paper.

References

Malin Antonsson, Kristina Lundholm Fors, Marie Eckerström, and Dimitrios Kokkinakis. 2021. Using a discourse task to explore semantic ability in persons with cognitive impairment. *Frontiers in Aging Neuroscience*, 12:607449.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:635945.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594.

Balamurali BT and Jer-Ming Chen. 2024. Performance assessment of chatgpt versus bard in detecting alzheimer's dementia. *Diagnostics*, 14(8):817.

Kayla Chapin, Natasha Clarke, Peter Garrard, and Wolfram Hinzen. 2022. A finer-grained linguistic profile of alzheimer's disease and mild cognitive impairment. *Journal of Neurolinguistics*, 63:101069.

Miguel A Chávez-Fumagalli, Pallavi Shrivastava, Jorge A Aguilar-Pineda, Rita Nieto-Montesinos, Gonzalo Davila Del-Carpio, Antero Peralta-Mestas, Claudia Caracela-Zeballos, Guillermo Valdez-Lazo, Victor Fernandez-Macedo, Alejandro Pino-Figueroa, et al. 2021. Diagnosis of alzheimer's disease in developed and developing countries: systematic review and meta-analysis of diagnostic test accuracy. *Journal of Alzheimer's Disease Reports*, 5(1):15–30.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Ying-Ren Chen, Chih-Sung Liang, Hsin Chu, Joachim Voss, Xiao Linda Kang, Grant O'Connell, Hsiu-Ju Jen, Doresses Liu, Shu-Tai Shen Hsiao, and Kuei-Ru Chou. 2021. Diagnostic accuracy of blood biomarkers for alzheimer's disease and amnestic mild cognitive impairment: A meta-analysis. *Ageing research reviews*, 71:101446.

Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. 1996. Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language*, 53(1):1–19.

Scott A Crossley, Max M Louwerse, Philip M McCarthy, and Danielle S McNamara. 2007. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30.

Francisco de Arriba-Pérez, Silvia García-Méndez, Francisco J González-Castaño, and Enrique Costa-Montenegro. 2023. Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with natural language processing capabilities. *Journal of ambient intelligence and humanized computing*, 14(12):16283–16298.

Flavio Di Palo and Natalie Parde. 2019. Enriching neural models with targeted features for dementia detection. *arXiv preprint arXiv:1906.05483*.

Elif Eyigoz, Sachin Mathur, Mar Santamaria, Guillermo Cecchi, and Melissa Naylor. 2020. Linguistic markers predict onset of alzheimer's disease. *EClinicalMedicine*, 28(100583).

Rui He, Kayla Chapin, Jalal Al-Tamimi, Núria Bel, Marta Marquié, Maitee Rosende-Roca, Vanesa Pytel, Juan Pablo Tartari, Montse Alegret, Angela Sanabria, et al. 2023. Automated classification of cognitive decline and probable alzheimer's dementia across multiple speech and language domains. *American Journal of Speech-Language Pathology*, 32(5):2075–2086.

Laura Hernández-Domínguez, Sylvie Ratté, Gerardo Sierra-Martínez, and Andrés Roche-Bergua. 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.

Bao Hoang, Yijiang Pang, Hiroko H Dodge, and Jiayu Zhou. 2023. Subject harmonization of digital biomarkers: Improved detection of mild cognitive impairment from language markers. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pages 187–200. World Scientific.

Susan Kemper, Lydia H Greiner, Janet G Marquis, Katherine Prenovost, and Tracy L Mitzner. 2001a. Language decline across the life span: findings from the nun study. *Psychology and aging*, 16(2):227–239.

Susan Kemper, Marilyn Thompson, and Janet Marquis. 2001b. Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. *Psychology and aging*, 16(4):600–614.

Daniel Kempler, Susan Curtiss, and Catherine Jackson. 1987. Syntactic preservation in alzheimer's disease. *Journal of Speech, Language, and Hearing Research*, 30(3):343–350.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H Robert, et al. 2015. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124.

Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. 2023. Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438.

Nicola T Lautenschlager, Kay L Cox, Leon Flicker, Jonathan K Foster, Frank M Van Bockxmeer, Jianguo Xiao, Kathryn R Greenop, and Osvaldo P Almeida. 2008. Effect of physical activity on cognitive function in older adults at risk for alzheimer disease: a randomized trial. *Jama*, 300(9):1027–1037.

Guillaume LemaÃŽtre, Fernando Nogueira, and Christos K Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18(17):1–5.

S Luz, F Haider, S de la Fuente Garcia, D Fromm, and B MacWhinney. 2021a. Editorial: Alzheimer's dementia recognition through spontaneous speech. *Frontiers in Computer Science*, 3.

Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. 2021b. Alzheimer's dementia recognition through spontaneous speech. *Frontiers in computer science*, 3:780169.

Pranav Mahajan and Veeky Baths. 2021. Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech. *Frontiers in Aging Neuroscience*, 13:623607.

Matej Martinc and Senja Pollak. 2020. Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia. In *Interspeech*, pages 2157–2161.

Rachel Ostrand and John Gunstad. 2021. Using automatic assessment of speech production to predict current and future cognitive function in older adults. *Journal of Geriatric Psychiatry and Neurology*, 34(5):357–369.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Paul B Rosenberg, Michelle M Mielke, Brian S Appleby, Esther S Oh, Yonas E Geda, and Constantine G Lyketsos. 2013. The association of neuropsychiatric symptoms in mci with incident dementia and alzheimer disease. *The American Journal of Geriatric Psychiatry*, 21(7):685–695.

Roozbeh Sadeghian, J David Schaffer, and Stephen A Zahorian. 2021. Towards an automatic speech-based diagnostic test for alzheimer's disease. *Frontiers in Computer Science*, 3:624594.

Seyed Ahmad Sajjadi, Karalyn Patterson, Michal Tomek, and Peter J Nestor. 2012. Abnormalities of connected speech in semantic dementia vs alzheimer's disease. *Aphasiology*, 26(6):847–866.

David A Snowdon. 1997. Aging and alzheimer's disease: lessons from the nun study. *The Gerontologist*, 37(2):150–156.

Yaakov Stern. 2006. Cognitive reserve and alzheimer disease. *Alzheimer Disease & Associated Disorders*, 20(2):112–117.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, Janos Kalman, and Magdolna Pakaski. 2015. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in aging neuroscience*, 7:195.

Spyridoula Varlokosta, Katerina Fragkopoulou, Dimitra Arfani, and Christina Manouilidou. 2024. Methodologies for assessing morphosyntactic ability in people with alzheimer's disease. *International journal of language & communication disorders*, 59(1):38–57.

Ines Vigo, Luis Coelho, and Sara Reis. 2022. Speech- and language-based classification of alzheimer's disease: a systematic review. *Bioengineering*, 9(1):27.

Daniel Zeman, Joakim Nivre, and Mitchell... Abrams. 2023. Universal dependencies 2.12. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Appendix

The OSF link includes the python code to extract the language features from Universal Dependency annotations, and the list of language features that are used for modeling in this study, as well as the code for SVM modeling (for the Pitt corpus).