AUTOMATIC DETECTION OF DEMENTIA IN MANDARIN CHINESE

by

Bai Li

A report submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

© Copyright 2019 by Bai Li

Abstract

Automatic Detection of Dementia in Mandarin Chinese

Bai Li
Master of Science
Graduate Department of Computer Science
University of Toronto
2019

Machine learning has shown promise for automatic detection of Alzheimer's disease (AD) through speech; however, efforts are hampered by a scarcity of data, especially in languages other than English. In this thesis, we give a description of the *Lu Corpus* of Mandarin Chinese speakers with dementia, which is part of AphasiaBank but has not been described before in the literature. Next, we propose a method to classify AD in Mandarin Chinese by combining the Lu Corpus with DementiaBank, a well-known database of AD speech in English. This method extracts lexicosyntactic features independently in the two languages, and then learns a correspondence between the features using a large parallel corpus of out-of-domain movie dialogue data. We demonstrate that our method outperforms both unilingual and machine translation-based baselines. This appears to be the first study on automatic methods of detecting dementia through speech in Mandarin Chinese, and also the first study that transfers features across domains to detect cognitive decline.

摘要

中国人痴呆症的自动检测

Bai Li
Master of Science
Graduate Department of Computer Science
University of Toronto
2019

机器学习已经显示出通过语言自动检测阿尔茨海默病（AD）的希望；然而，由于缺乏数据，特别是在英语以外的语言中，这方面的努力受到阻碍。在本文中，我们描述了痴呆症普通话的Lu Corpus，它是AphasiaBank的一部分，但在文献中没有被描述过。接下来，我们提出了一种方法，通过将Lu Corpus与DementiaBank（一个著名的英语AD语音数据库）相结合，对普通话中的AD进行分类。该方法在两种语言中独立地提取词汇句法特征，然后使用大量其他领域中英文平行的电影对话数据来学习特征之间的对应关系。我们证明我们的方法优于基于单语和基于机器翻译的基准。这应该是第一项关于通过语言自动检测痴呆症方法的研究，也是第一项跨域转移特征以检测认知能力下降的研究。

Acknowledgements

This thesis would not be possible without my advisor, Professor Frank Rudzicz, who provided guidance and support throughout my research journey. I am also thankful to my second reader, Professor Yang Xu. Thanks to Yi-Te Hsu, who helped with many experiments and results that made it into this thesis. Part of this work was inspired by an earlier collaboration with overseas colleagues Katie Fraser and Nicklas Linz. Thanks to my girlfriend Elaine, and my friends Andrei, Kevin, and Michael for their moral support during these hectic years. To my colleagues at SnapTravel, Brett, Cory, Henry, Leon, Nehil, Tommy, and Weilong: thank you for such an exciting summer internship. I have learned a lot in our weekly discussions in the CL paper reading group, with knowledgeable graduate and undergraduate students Akshay, Chloe, Eric, Jeff, John, KP, Nona, Serena, and Zining. Lastly, I would like to thank my parents and sister in Calgary for their encouragement throughout all these years of school, and for performing a lot of data annotation for this project.

Contents

Chapter 1

Introduction

Alzheimer's disease (AD) is a neurodegenerative disease affecting all aspects of cognition. The initial symptoms are mild, such as occasionally misplacing an object or forgetting names of people. As the disease progresses, dementia takes hold, with various symptoms like memory loss, language decline, and changes to personality. Eventually, basic brain functions deteriorate and death inevitably follows. In 2018, 5.7 million people suffer from AD in the US alone, with the number expected to increase in the coming decades (Alzheimer's Association, 2018).

The causes of AD are poorly understood. The primary hypothesis states that AD is caused by buildup of amyloid plaques, which is caused by abnormalities in the tau protein, but a cure has remained elusive despite decades of research (Goedert and Spillantini, 2006). AD is one of the most financially costly diseases in developed countries: in the later stages of the disease, patients require caregivers for basic functions. Over 6 billion hours a year are spent on caring for patients with dementia in the US, much of which is unpaid (Kasper et al., 2015).

Although there is no cure available for AD, early detection is important so that medical interventions can be applied to mitigate its effects (Dubois et al., 2016). Speech impairment is one of the earliest symptoms of AD, so methods in natural language processing hope to provide a cheap and easy diagnostic tool. This has achieved some success, for example, Fraser et al. (2016) classified patients with AD from healthy controls with about 82% accuracy using a variety of linguistic and acoustic signals. Surprisingly, this accuracy is achieved using only about two minutes of speech: a picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983). Further research improved and refined this result, which I will review in Chapter 2. Recently, partners in the industry have made progress towards productionizing this research to bring it to clinical practice[1].

For the most part, research on linguistic AD detection have focused on English speakers. The natural question which my thesis addresses is: can the promising results in English be generalized to other languages? Mandarin Chinese is a prime candidate, as it is the largest language by number of native speakers, and the second largest language by number of total speakers (Lewis, 2009).

Fundamentally, NLP in Chinese is not very different from NLP in English, and many techniques work well on both languages. The main difficulty towards dementia detection in Chinese is the availability of data. In English, DementiaBank (Boller and Becker, 2005) is perhaps the most studied corpus involving language and cognitive decline with several hundred participants, collected between 1983 and 1988 at the University of Pittsburgh. Unfortunately, no public corpus of similar size exists for Chinese (or any other language) – most studies consist of just a few dozen speakers. Many machine learning methods that work for DementiaBank fail to generalize when trained on a dataset of this size. Hope is not lost, however: transfer learning and few-shot learning are methods

[1] `https://winterlightlabs.com/`

to augment a small dataset by combining it with a larger dataset from a different domain, achieving performance beyond what is possible using one dataset alone.

In this thesis, I describe the *Lu corpus* of Mandarin Chinese speakers with dementia. Then, I propose a transfer learning method that learns to detect dementia in Mandarin by combining knowledge from several different sources, and demonstrate that indeed, it performs better than using the Mandarin corpus alone.

1.1 Thesis Organization

This thesis is structured as follows:

- Chapter 2 summarizes the related work on linguistic effects of AD, advances in automatic detection of AD using machine learning, and methods of combining datasets using domain adaptation.

- Chapter 3 introduces the Lu corpus, consisting of 52 speakers of Taiwanese Mandarin. Although the Lu corpus is part of AphasiaBank (MacWhinney et al., 2011), no description of it exists in the literature. We give a description of its contents, including additional annotations that were not present in the original dataset. We also perform some statistical analysis of the data and apply some transformations to prepare it for downstream analysis.

- Chapter 4 describes a novel method of detecting AD in other languages by combining DementiaBank with datasets in different languages. Specifically, our method learns a correspondence between lexicosyntactic features, using a large parallel corpus of out-of-domain movie dialogue data. We apply it to the Lu corpus and achieve promising results that significantly outperform the unilingual baseline.

1.2 Summary of contributions

My thesis makes two main research contributions:

1. It gives the first detailed description of the Lu corpus, adds additional annotations, and provides an interesting use case for this dataset. Future researchers studying dementia in Mandarin Chinese can refer to this description and contact me for annotation data that would otherwise take considerable effort to replicate.

2. It proposes a novel method of detecting dementia by combining datasets in two languages. To my knowledge, this is the first effort at using machine learning to detect dementia in Mandarin Chinese. In this thesis, we apply the method to Mandarin dementia detection, but it can be generalized to other languages, and perhaps other cognitive disorders as well. Ideally, future researchers will build upon and extend my work.

Chapter 2

Background and Related Work

In this chapter, we start by giving an overview of the language effects of AD and tests that identify them. Then, we summarize methods of automatically extracting these features and classifying AD using machine learning. Finally, we look at similar linguistic effects of AD that have been observed in other languages, and existing methods of combining datasets to train a classifier.

2.1 Alzheimer's disease and language

Alzheimer's disease is a neurodegenerative disease affecting 5.7 million people in the US in 2018, and 1 in 10 people over the age of 65. It is the most financially costly disease in the American healthcare system today; the cost is expected to further increase since the population of Americans over age 65 is projected to grow from 53 million today to 88 million in 2050 (Alzheimer's Association, 2018).

Dementia is a symptom of AD, characterized by a long-term decline in memory and cognitive ability. Although no cure currently exists for AD, early detection and intervention is crucial to mitigate the effects of dementia (Dubois et al., 2016).

Language deterioration and speech impairment are some of the earliest symptoms of AD. In particular, patients with AD have difficulty finding words for specific objects. This lends itself to reliable cognitive tests for AD detection. For example, category and letter fluency tasks, where the subject names as many objects in a certain category or starting with a certain letter in 60 seconds, reliably differentiates between AD patients and healthy controls (Monsch et al., 1992). The Mini Mental State Examination (MMSE) is a popular test in which the subject is given a battery of cognitive tests and is scored on a range from 0 to 30 (Folstein et al., 1983).

Connected speech is another reliable method of detecting AD, since patients with AD have been shown to exhibit differences in word frequency, syntactic complexity, idea density, and pause frequency and duration (Taler and Phillips, 2008; Roark et al., 2011). A popular task for eliciting connected speech is the *Cookie Theft* picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983). In this task, the subject is shown the picture in Figure 2.1 and is asked to describe it in as much detail as possible.

2.2 Machine learning for detection of AD

Computerized analysis combined with machine learning is effective for working with continuous speech because computers can quickly extract a wide variety of features that would be extremely time-consuming to extract

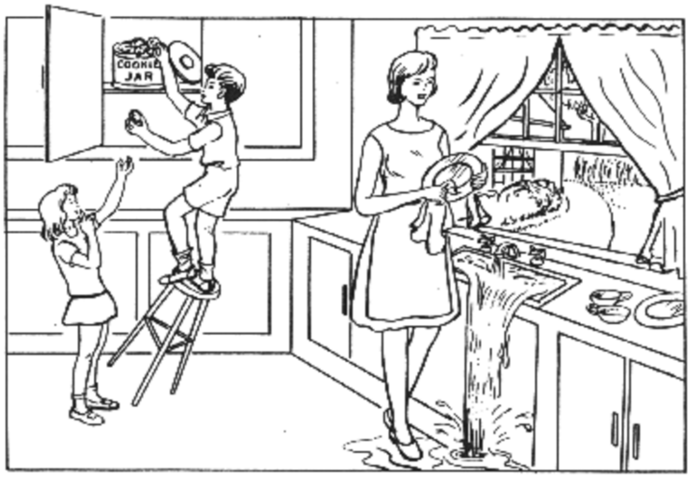

Figure 2.1: Picture used for the *Cookie Theft* picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983)

manually. Fraser et al. (2016) extracted 370 acoustic, syntactic, and semantic features such as part-of-speech ratios, vocabulary richness, and MFCC coefficients; using a logistic regression classifier, their model achieved about 82% accuracy in distinguishing between AD patients and healthy controls. It was trained on 473 samples from DementiaBank (Boller and Becker, 2005), using audio recordings and manual transcripts of the *Cookie Theft* picture description task.

These results have been further refined in subsequent work. Yancheva and Rudzicz (2016) used topic models to automatically identify clusters of content words like *cookie* and *mother*, which were specified manually by Fraser et al. (2016). Zhou et al. (2016) investigated the possibility of replacing manual transcripts with automatic speech recognition (ASR).

Karlekar et al. (2018) used deep neural networks like CNNs and LSTMs to achieve state-of-the-art classification accuracy on the DementiaBank dataset. Computerized analysis of dementia has also been applied to written text, for example, to detect signs of dementia in the works of three British novelists (Le et al., 2011).

2.3 Multilingual effects of AD

The linguistic effects of AD has been studied in numerous languages, such as Mandarin Chinese (Lai et al., 2009; Lai, 2014), Japanese (Shibata et al., 2016), Koreaen (Kim et al., 2006), Portuguese (Aluísio et al., 2016), French (Tröger et al., 2017), and Hebrew (Kavé and Levy, 2003). However, most of these studies focused on the linguistic characteristics of AD, rather than machine learning methods to detect AD through speech. Fraser et al. (2019) used multilingual topic models in English and Swedish to detect mild cognitive impairment (MCI), but so far, very little work has been done on multilingual dementia detection.

In Mandarin Chinese, Lai et al. (2009) analyzed the speech of 62 patients performing the *Cookie Theft* picture description task, and found syntactic differences among speakers with AD. A subsequent study found differences in discourse patterns (Lai, 2014). However, in both studies, features were coded manually by trained linguists. Chinese grammar is very different from most Indo-European languages: for example, it lacks verb tenses and all inflectional morphology, and has a system of noun classifiers (Chao, 1965). These differences pose unique challenges to studies of cognitive decline in Chinese. When studying grammaticality tests for aphasic patients, Lu et al. (2000) found that the free word order and simple morphology in Chinese made it difficult to construct sentences that were definitively ungrammatical.

2.4 Methods for domain adaptation

One of the biggest challenges for using machine learning to detect dementia is the sparsity of datasets. This is especially true for non-English languages, where clinical studies involving dementia consist of less than 100 people. Domain adaptation (e.g., transfer learning) methods aim to learn from a small dataset by combining it with a much larger dataset from a different domain, for example, data from English speakers with dementia, or normative speech. This is a promising direction of research for multilingual dementia detection.

Daume III (2007) proposed a simple way of combining features in different domains, assuming that the same features are extracted in each domain. In this method, given data points in domains S and T, we augment the feature space with three copies of each feature, one copy specific to domain S, one copy specific to domain T, and a third copy common to both domains. Formally, the function Φ_S transforms a vector $\mathbf{x}_S$ in the source domain S, as defined by the formula:

$$\Phi_S(\mathbf{x}_S) = \langle \mathbf{x}_S, \mathbf{x}_S, \mathbf{0} \rangle.$$

Similarly, the function Φ_T transforms a vector $\mathbf{x}_T$ in the target domain T:

$$\Phi_T(\mathbf{x}_T) = \langle \mathbf{x}_T, \mathbf{0}, \mathbf{x}_T \rangle.$$

A classifier is then trained on the combined dataset using a standard classification algorithm.

Although simple and straightforward, this method is limited to situations where the same features may be extracted in each domain. In multilingual NLP, this is unreasonably restrictive because languages may be very different from each other. Duan et al. (2012) proposed an extension to this method to train a classifier jointly on two domains with different features in each domain, by learning projections to a common subspace. That is,

$$\Phi_S(\mathbf{x}_S) = \langle P\mathbf{x}_S, \mathbf{x}_S, \mathbf{0} \rangle,$$

$$\Phi_T(\mathbf{x}_T) = \langle Q\mathbf{x}_T, \mathbf{0}, \mathbf{x}_T \rangle,$$

where P and Q are projection matrices learned from data.

Domain adaptation and multi-task learning have been applied to automatic classification of AD as well. Noorian et al. (2017) improved the classification accuracy on DementiaBank by augmenting it with a larger corpus of normative speech data. Pou-Prom and Rudzicz (2018) applied generalized canonical correlation analysis (GCCA) to combine picture description, category fluency, letter fluency, and demographic data to generate a multiview embedding for downstream classification. Zhu et al. (2018) proposed transductive consensus networks (TCNs) to generate similar interpretations of different modalities; on DementiaBank, they treated acoustic, syntactic, and lexical features as three separate modalities. Masrani et al. (2017) detected mild cognitive impairment (MCI) in DementiaBank by applying domain adaptation techniques to augment the MCI group with the much larger AD group from the same dataset.

In this thesis, we propose a new method of domain adaptation across different languages. This method learns a correspondence between feature spaces using a large bilingual parallel corpus. This appears to be the first study that transfers domains in detecting cognitive decline.

Chapter 3

The Lu Corpus: Mandarin Speakers with Dementia

In this chapter, we describe the Lu corpus[1], which is part of AphasiaBank (MacWhinney et al., 2011). From now on, we will refer to this corpus as the *Lu corpus* after its author, Ching-ching Lu.

To our knowledge, no description or analysis of this dataset exists in the literature. The dataset consists of Mandarin speakers performing a category fluency, picture description, and picture naming task; unfortunately, demographic and diagnostic information is unknown. We perform some manual annotations of the audio interviews, giving numerical scores to the category fluency and picture naming tasks. Finally, we derive a combined score using principal component analysis (PCA) that represents the degree of dementia for each patient.

3.1 Description of dataset and tasks

The dataset contains 52 audio files, each containing an interview with a participant (with a 3-digit ID) of duration about 10 minutes. Three samples were excluded, leaving with a sample size of 49 for analysis[2]. The participants spoke a Taiwanese Mandarin (a variety of Mandarin Chinese very close to Standard Mandarin Chinese). Demographic data, such as age, gender, and education levels of the participants are not given (although gender can be inferred through the audio files). It is also not known which participants (if any) are diagnosed with AD, but some participants exhibit clear signs of dementia.

The interviews consist of the following tasks:

1. **Category Fluency**: There are four category fluency tasks, where the participant is asked to name as many objects as possible of a given category:

    (a) Name as many animals as you can.

    (b) Name as many fruits as you can.

    (c) Name as many colors as you can.

    (d) Name as many places in Taiwan as you can.

[1] `https://dementia.talkbank.org/access/Mandarin/Lu.html`

[2] Subjects 002 and 045 were excluded because the interviews did not contain the category naming tasks. File 028 was excluded because the language was Taiwanese Hokkien, not Mandarin.

#	Chinese	Romanization	English
1	樹	shù	tree
2	蝸牛	wōniú	snail
3	筆	bǐ	pencil
4	仙人掌	xiānrénzhǎng	cactus
5	剪刀	jiǎndāo	scissors
6	聽診器	tīngzhěnqì	stethoscope
7	駱駝	luòtuó	camel
8	飛鏢	fēibiāo	dart
9	鋸子	jùzi	saw (tool)
10	口琴	kǒuqín	harmonica
11	衣架	yījià	clothes hanger
12	金字塔	jīnzìtǎ	pyramid
13	花	huā	flower
14	犀牛	xīniú	rhinoceros
15	漏斗	lòudǒu	funnel
16	夾子	jiázi	folder
17	算盤	suànpán	abacus
18	手風琴	shǒufēngqín	accordion
19	球拍	qiúpāi	tennis racket
20	量角器	liángjiǎoqì	protractor
21	樓梯	lóutī	escalator
22	三角架	sānjiǎojià	tripod
23	掃把	sàobǎ	broom
24	豎琴	shùqín	harp
25	輪椅	lúnyǐ	wheelchair
26	花架	huājià	flower stand
27	海馬	hǎimǎ	seahorse
28	蘑菇	mógū	mushroom
29	冰屋	bīngwū	igloo
30	圓規	yuánguī	compass (geometry)

Table 3.1: Sequence of 30 items used for the picture naming task, with their romanizations and English translations.

For each task, the participant is allowed a time limit of 60 seconds; if he stops before the time limit, then the interviewer encourages him to “go on” until the time limit is reached.

2. **Picture Description**: The participant is shown the *Cookie Theft* picture (Figure 2.1) and asked to describe it in as much detail as possible. There is no time limit; the task ends when the participant indicates that he has nothing more to say. Transcriptions of the picture description tasks are given in the Lu corpus.

   Next, the interviewer asks some additional questions about people and objects in the *Cookie Theft* picture. However, transcripts are not given for this section, so we ignore it in our analysis.

3. **Picture Naming**: The participant is asked to name a sequence of 30 pictures (list given in Table 3.1). If he is stuck, the interviewer gives a hint describing what the object is (e.g., for *cactus*, the hint is “a plant found in the desert”). If he is still stuck, the interviewer gives a second hint, providing the first character of a multi-character word (e.g., *xiān* for *xiānrénzhǎng* “cactus”). If he is still stuck, the interviewer moves on.

3.2 Analysis of task data

3.2.1 Annotations and task scoring

Seven volunteers were recruited to annotate the category fluency and picture naming tasks, and asked to score the tasks according to the following criteria:

- **Category Fluency**: For each of the 4 category fluency tasks, the annotator recorded the number of unique items that the participant named, and separately recorded the number of repeated items.
- **Picture Naming**: For each of the 30 picture naming tasks, the annotator scored 3 points if the participant got the correct answer without hints, 2 points the answer was given after one hint, 1 point if the answer was given after two hints, and 0 points if the participant never produced the correct answer.

This produces a list of 38 scores for each participant (8 for category fluency and 30 for picture naming). To check for accuracy, each annotator labelled participant 003, for whom the correct answers were known; if their annotations deviated from the correct answers, then their work was done again by a different annotator.

3.2.2 Data analysis

We aggregated the data from all the annotators; using R, we visualized the score distribution within each task and the relationships between scores across different tasks.

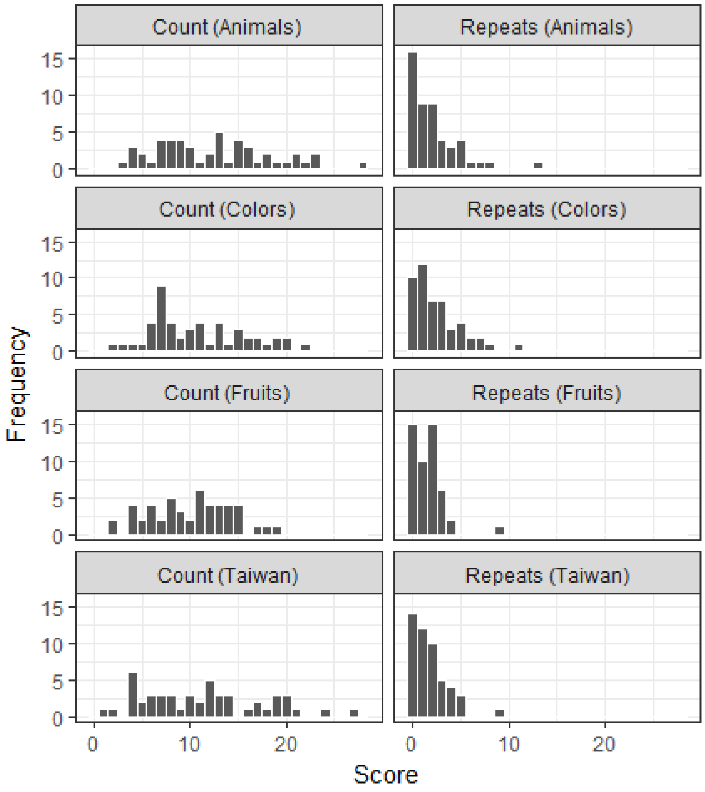

Figure 3.1: Histograms of scores (counts and repeats) for each of the 4 category fluency tasks.

Category	Mean Count	Mean Repeated
Animals	12.4	2.1
Fruits	10.0	1.5
Colors	10.9	2.5
Taiwan	11.4	1.8

Table 3.2: Mean counts and repeats for each of the 4 category fluency tasks.

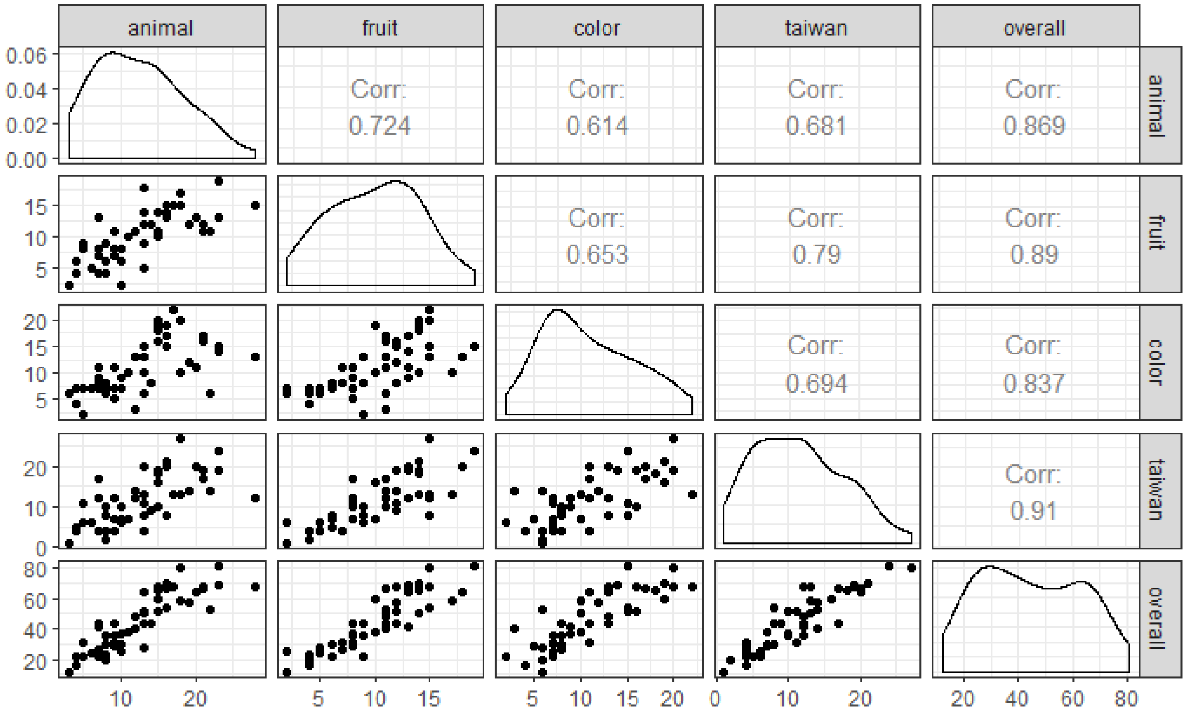

Figure 3.2: Scatterplot matrix and correlation coefficients between each of the category fluency tasks and the overall sum of their counts.

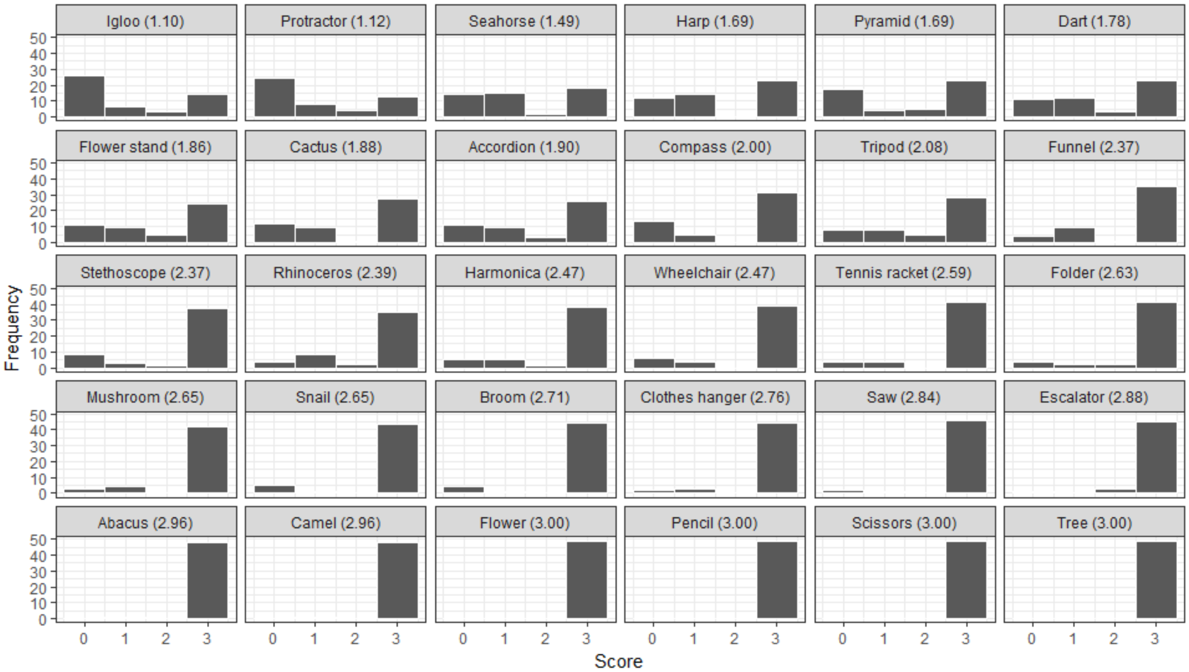

Figure 3.3: Histograms of score distributions for the 30 picture naming tasks, with mean scores in brackets (0 – didn't get the answer at all; 1 – got answer after two hints; 2 – got answer after one hint; 3 – got answer immediately).

Table 3.2 and Figure 3.1 show the score distribution for the category fluency tasks. The four tasks have comparable score distributions, with the item count having a wide dispersion and most participants having few repeats. Each of the category fluency tasks are strongly correlated, and the overall score appears to be bimodal (Figure 3.2).

Figure 3.3 shows that there is considerable variation in success rates for different pictures. For example, each participant correctly identified "tree", but less than half the participants produced the word for "igloo", even with hints. When the scores of all 30 pictures are summed, the resulting distribution appears bimodal (Figure 3.4).

Finally, Figure 3.5 shows that the category fluency and picture naming scores are correlated ($R^2 = 0.72$).

Overall, the analysis shows high levels of agreement between the different cognitive task scores. The participants in the study also exhibit a wide range of dementia severity, from serious dementia to relatively normal cognitive functioning. Therefore, it is reasonable to use a combination of the task scores as a proxy for degree of dementia. However, the question still remains of how to combine the scores from different tasks into a single unified *dementia score* for each patient. We will address this problem in the next section.

3.3 Dementia score using PCA

The simplest solution is to assign weights to different tasks in an *ad-hoc* manner, for example, 1 point for each item and picture named and -0.5 for each item repeated. This would produce a score for each patient, but requires arbitrary choices to be made. In this section, we use PCA to combine the task scores to create a *dementia score* in a more principled manner.

Principal component analysis (PCA) is a procedure that finds a linear transformation of a dataset into or-

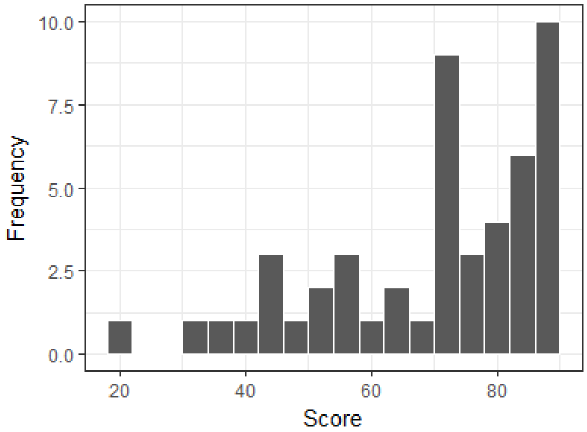

Figure 3.4: Histogram of the sum of picture naming scores (max score is 90).

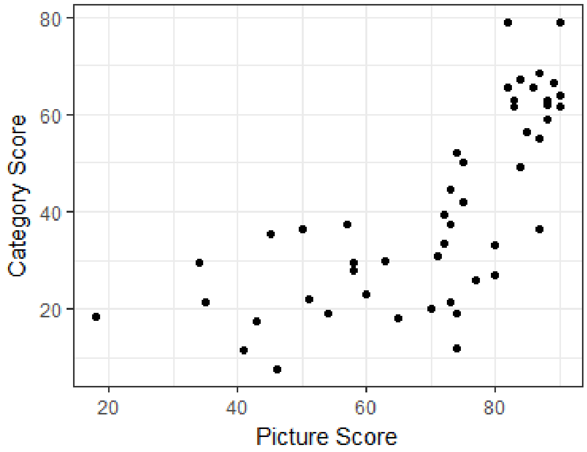

Figure 3.5: Relationship between category fluency and picture naming scores.

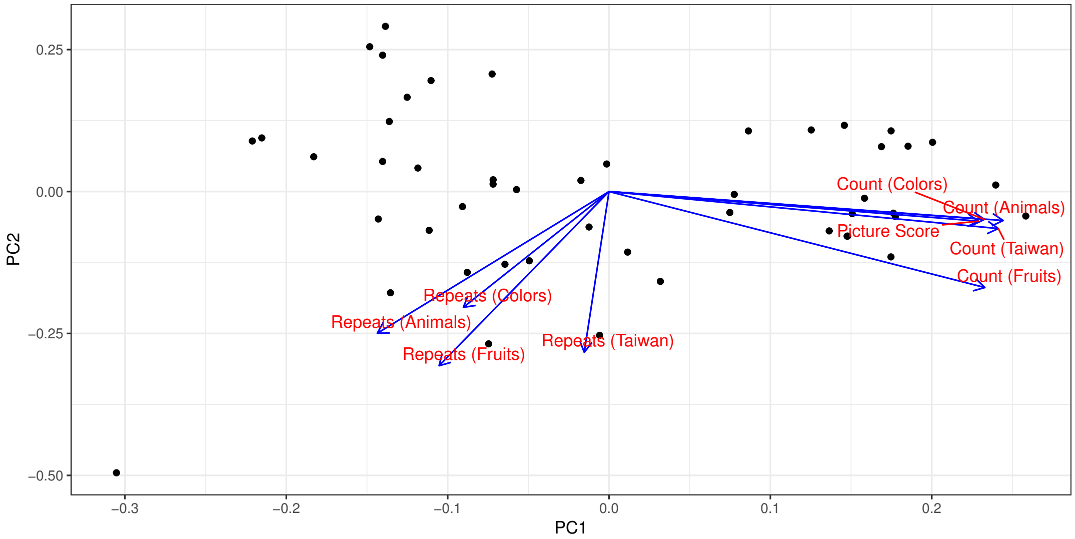

Figure 3.6: Weightings of first two principal components. The data were transformed to have zero mean and unit variance.

thogonal *principal components* (Hotelling, 1933). Furthermore, the first principal component captures the largest possible variance. Specifically, let $X \in \mathbb{R}^{n \times p}$ be a data matrix where each row is a participant and each column is a task. PCA finds a series of weight vectors $(\mathbf{w}_1, \cdots, \mathbf{w}_p)$ that maps rows of X into principal components, such that for each $\mathbf{w}_i$, $|\mathbf{w}_i| = 1$, and each successive $X\mathbf{w}_i$ takes on the maximum remaining variance of X.

This is done by transforming X such that each column has zero mean and unit variance, then taking the singular value decomposition $X = U\Sigma V^T$. Then, the leftmost column of V (corresponding to the largest singular value) gives the first principal axis $\mathbf{w}_1$ that maximizes $\mathrm{Var}(X\mathbf{w}_1)$. Further principal components can be computed by subtracting the first principal component and repeating the same procedure, but we only use the first principal component in our dementia score.

Figure 3.6 shows the dataset projected onto the first two principal components and the weightings of each of the columns. On the first principal component, the picture naming score and each of the category fluency counts have positive weighting and each of the repeat scores have negative weighting. This agrees with our intuitive understanding that healthier subjects are able to name more items and produce fewer repetitions. The first principal component accounts for significantly more of the total variance than any other principal component, further indicating its importance (Figure 3.7).

Therefore, the first principal component is a reasonable choice for the overall dementia score, with lower scores indicating a greater severity of dementia. Figure 3.8 shows the distribution of dementia scores, which can be used in downstream regression or classification tasks. For regression, the dementia score may be used directly; for classification, an empirical threshold value of 1.5 (with scores ≥ 1.5 taken to be healthy and scores < 1.5 to have dementia) gives a reasonable separation between the two classes, with 33 participants in the dementia group and 16 in the healthy group.

Note that the sign of the weights generated by PCA are arbitrary, so it is not necessarily the case that higher values of PC1 indicate more severe dementia. In that case, we can simply re-orient the dimensions by reversing the sign of each weighting.

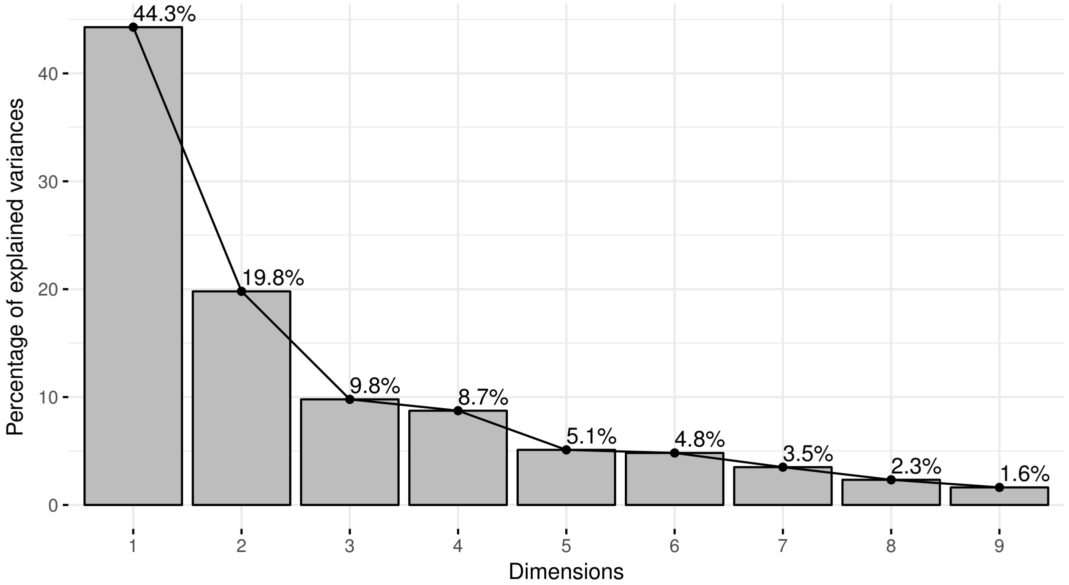

Figure 3.7: Scree plot showing the proportion of total variance explained by each principal component.

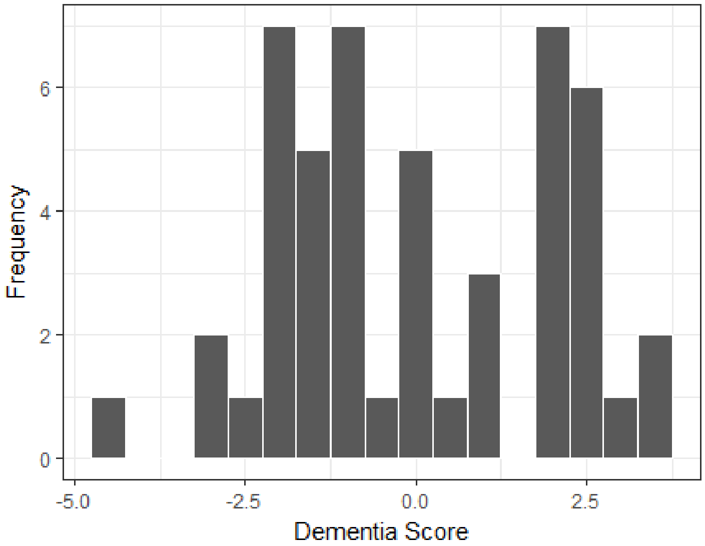

Figure 3.8: Distribution of our dementia scores (first principal component of PCA).

3.4 Conclusion and limitations

In this chapter, we described the Lu corpus and produced manual annotations for the category fluency and picture naming tasks. Then, we derived a single *dementia score* for each participant representing the severity of dementia. This score is derived independently of the *Cookie Theft* picture description task, which makes this dataset suitable for studies involving automatic detection of dementia in Mandarin Chinese using speech.

The dataset has some limitations. First, it only contains 49 participants after exclusions, which is too small for most machine learning algorithms to perform effectively. Innovative methods (e.g., few-shot or transfer learning) would likely be required to avoid overfitting. Second, it is missing demographic information, such as age, diagnosis, medical history, and education. Still, it is presently one of the only publicly available datasets studying cognitive decline in Mandarin speakers.

In the next chapter, we demonstrate that the Lu corpus can be used in combination with DementiaBank to predict dementia using Mandarin speech, using a transfer learning approach. This shows that despite its limitations, the dataset is still a valuable asset in the study of multilingual dementia detection.

Chapter 4

Transfer Learning for Multilingual Dementia Detection

In this chapter, we describe a method to detect dementia in Mandarin Chinese, using DementiaBank and a large parallel corpus of normative movie dialogue. We extract lexicosyntactic features in Mandarin and English using separate pipelines, and use the OpenSubtitles corpus of bilingual parallel movie dialogues to learn a correspondence between the different feature sets. We combine this correspondence model with a classifier trained on DB to predict dementia on Mandarin speech. To evaluate our system, we apply it to the Lu corpus, and demonstrate that our method outperforms baselines based on unilingual models and Google Translate.

Our method is successful despite the stark differences between English and Mandarin, and the fact that the parallel corpus is out-of-domain for the task. Most notably, our method is *unsupervised* in the sense that it does not require the Lu corpus for training (only for evaluation), which is important given the difficulty of acquiring sensitive clinical data.

4.1 Motivation

In Chapter 2, we reviewed two methods of domain adaptation (Daume III, 2007; Duan et al., 2012), which could be applied to transfer knowledge between languages for dementia detection. However, both of these methods have limitations that render them unsuitable for our task.

Daume III (2007) required the same features to be extracted in each domain – we might create linguistic features general enough to be applicable to any language, like average sentence length or noun ratio. Difficulties arise if we attempt this on languages as different as Mandarin and English. Consider, for example, part-of-speech (POS) tagging. Chinese does not have articles or conjugate verbs based on tense, and uses a system of noun classifiers[1] which does not exist in English (Chao, 1965). It is possible to define mappings to a broad universal POS categories common to all languages (Petrov et al., 2012), but such a mapping would necessarily be imperfect and would lose information.

Duan et al. (2012) performed heterogeneous domain adaptation, thus allowing different features to be extracted in the two languages. However, this is still difficult because the Mandarin dataset is so small ($n = 49$). A standard

[1] Classifiers (a.k.a., measure words) are required in Mandarin Chinese when a noun is quantified by a numeral. The type of classifier varies depending on the physical properties of the noun. For example, the general classifier *gè* is used in "a person" (*yí gè rén*, one-CL person), whereas the classifier for vehicles *liàng* is used to say "a car" (*yí liàng chē*, one-CL car).

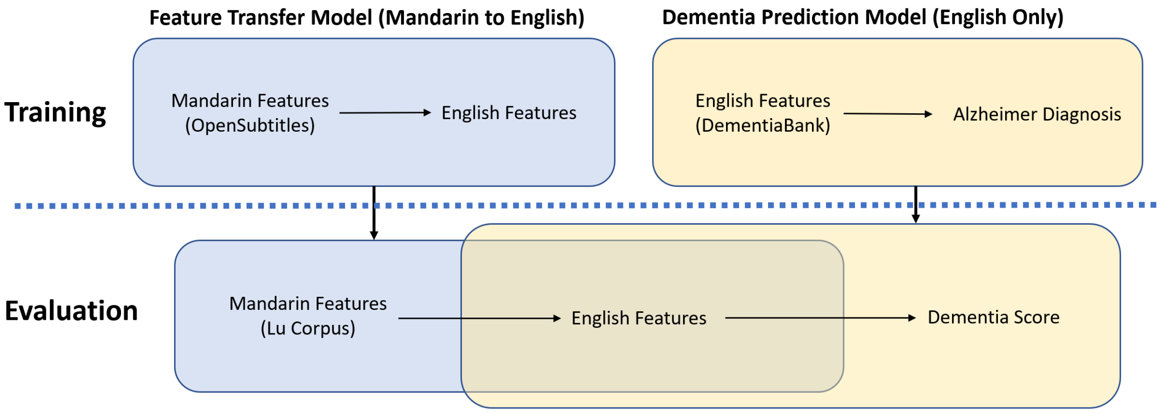
Feature Transfer Model (Mandarin to English)
Dementia Prediction Model (English Only)
Training
Mandarin Features (OpenSubtitles)
English Features
English Features (DementiaBank)
Alzheimer Diagnosis
Evaluation
Mandarin Features (Lu Corpus)
English Features
Dementia Score

Figure 4.1: Diagram of our model. We train two separate models: the first is trained on OpenSubtitles and learns to map Mandarin features to English features; the second is trained on DementiaBank and predicts dementia given English features. During evaluation, the two models are combined to predict dementia in Mandarin.

80/20 train-test split would leave only 10 samples in the test set, which has practically no chance of reaching statistical significance. Our proposed method leverages a third dataset, a bilingual parallel corpus, for domain adaptation. This allows the entire Mandarin corpus to be set aside for evaluation, since none of it is used for training.

4.2 Datasets

We use the following datasets:

- **DementiaBank** (Boller and Becker, 2005): a corpus of *Cookie Theft* picture descriptions, containing 241 narrations from healthy controls and 310 from patients with dementia. Each narration is professionally transcribed and labelled with part-of-speech tags. In this work, we only use the narration transcripts, and neither the part-of-speech tags or raw acoustics.

- **Lu Corpus** (MacWhinney et al., 2011): contains 49 patients performing the *Cookie theft* picture description, category fluency, and picture naming tasks in Taiwanese Mandarin. The picture description narrations were human-transcribed; a *dementia score* was derived for each patient based on category fluency and picture naming scores (see Chapter 3 for details).

- **OpenSubtitles2016** (Lison and Tiedemann, 2016): a corpus of parallel dialogues extracted from the website *OpenSubtitles*[2], containing subtitles from movies and TV shows, spanning 2.6 billion sentences across 60 languages. The subtitles in different languages were preprocessed and aligned using a time-based algorithm. In this work, we use the Traditional Chinese / English language pair, which contains 3.3 million lines of dialogue.

[2] https://opensubtitles.org

4.3 Methodology

4.3.1 Feature extraction

We extract a variety of lexicosyntactic features in Mandarin and English, including type-token-ratio, the number of words per sentence, and proportions of various part-of-speech tags. A detailed description of the features is provided in Table 4.1 and 4.2. The English feature set is similar to the features extracted by Fraser et al. (2016), except that we do not use any acoustic features.

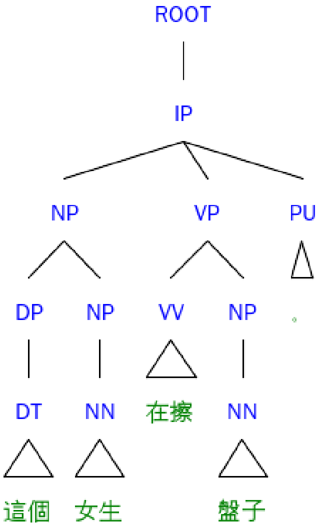

Figure 4.2: Example constituency parse tree of an utterance from the Lu corpus: "這個女生在擦盤子" *(this woman is wiping the dishes).*

Some of our features are derived from constituency parse trees (Figure 4.2); we used Stanford CoreNLP to do constituency parsing and part-of-speech tagging (Klein and Manning, 2003; Levy and Manning, 2003). We also used `wordfreq` (Speer et al., 2018) for word frequency statistics in both languages. The entire feature extraction pipeline has been made open-source[3]. In total, we extract 143 features in Mandarin and 185 in English. To reduce sparsity, we remove features in both languages that are constant for more than half of the dataset.

Feature extraction in Chinese was similar to English, despite the differences between the languages, except for two small challenges. First, extra care must be taken to ensure that Unicode is used throughout the pipeline, whereas ASCII is sufficient for English. Second, Chinese writing does not use spaces to separate words, so word segmentation must be done as an additional preprocessing step. Word segmentation is handled by Stanford CoreNLP, but still introduces a source of error.

Our feature extraction pipeline (especially the parsing stage) was computationally intensive, so it was unfeasible to run it on the entire OpenSubtitles corpus. Therefore, we randomly select 50,000 narrations from the corpus, where each narration consists of between 1 to 50 contiguous lines of dialogue (about the length of a *Cookie Theft* narration).

For English, we train a logistic regression classifier to classify between dementia and healthy controls on DB, using our features as input. Using L1 regularization and 5-fold CV, our model achieves 77% classification

[3] Code available at: `https://github.com/SPOClab-ca/COVFEFE`. The `lex` and `lex_chinese` pipelines were used for English and Chinese, respectively.

English (185 features)

Narrative length: Number of words and sentences in narration.
Vocabulary richness: Type-token ratio, moving average type-token ratio (with window sizes of 10, 20, 30, 40, and 50 words), Honoré's statistic, and Brunét's index.
Frequency metrics: Mean word frequencies for all words, nouns, and verbs.
POS counts: Counts and ratios of nouns, verbs, inflected verbs, determiners, demonstratives, adjectives, adverbs, function words, interjections, subordinate conjunctions, and coordinate conjunctions. Also includes some special ratios such as pronoun / noun and noun / verb ratios.
Syntactic complexity: Counts and mean lengths of clauses, T-units, dependent clauses, and coordinate phrases as computed by Lu's syntactic complexity analyzer (Lu, 2010).
Tree statistics: Max, median, and mean heights of all CFG parse trees in the narration.
CFG ratios: Ratio of CFG production rule count for each of the 100 most common CFG productions from the constituency parse tree.

Table 4.1: Lexicosyntactic features extracted in English.

Mandarin Chinese (143 features)

Narrative length: Number of sentences, number of characters, and mean sentence length.
Frequency metrics: Type-token ratio, mean and median word frequencies.
POS counts: For each part-of-speech category, the number of it in the utterance and ratio of it divided by the number of tokens. Also includes some special ratios such as pronoun / noun and noun / verb ratios.
Tree statistics: Max, median, and mean heights of all CFG parse trees in the narration.
CFG counts: Number of occurrences for each of the 60 most common CFG production rules from the constituency parse tree.

Table 4.2: Lexicosyntactic features extracted in Mandarin.

accuracy on DB. This is slightly lower than the 82% accuracy reported by Fraser et al. (2016), but it does not include any acoustic features as input.

4.3.2 Feature transfer

Next, we use the OpenSubtitles corpus to train a model to transform Mandarin feature vectors to English feature vectors. For each target English feature, we train a separate ElasticNet linear regression (Zou and Hastie, 2005), using the Mandarin features of the parallel text as input. We perform a hyperparameter search independently for each target feature, using 3-fold CV to minimize the MSE.

4.3.3 Regularization

Although the output of the ElasticNet regressions may be given directly to the logistic regression model to predict dementia, this method has two limitations. First, the model considers each target feature separately and cannot take advantage of correlations between target features. Second, it treats all target feature equally, even though some are noisier than others. We introduce two regularization mechanisms to address these drawbacks: reduced rank regression and joint feature selection.

Reduced rank regression

Reduced rank regression (RRR) trains a single linear model to predict all the target features: it minimizes the sum of MSE across all target features, with the constraint that the rank of the linear mapping is bounded by some given R (Izenman, 1975). More formally, let $\mathbf{X}$ be a matrix of predictors ($n \times p$) and $\mathbf{Y}$ be a matrix of responses ($n \times q$). We solve

$$\begin{aligned} \min \quad & ||\mathbf{Y} - \mathbf{XB}||^2 \\ \text{s.t.} \quad & \operatorname{rank}(\mathbf{B}) \leq R, \end{aligned}$$

where R is a hyperparameter.

If $R \geq \min(p, q)$, then the problem reduces to linear regression with ordinary least squares, and it is well known that in this case, the solution to the RRR formulation is the same as training q linear regressions separately, one for each target feature. Otherwise, if $x < \min(p, q)$, then RRR provides regularization while capturing relationships between response variables.

Following recommended procedures (Davies, 1982), we scale each target feature to have unit variance, and find the best value of R with cross validation. However, this procedure did not significantly improve results so it was not included in our best model.

Joint feature selection

A limitation of the above models is that they are not robust to noisy features. For example, if some English feature is useful for predicting dementia, but cannot be accurately predicted using the Mandarin features, then including this feature might hurt the overall performance. A desirable English feature in our pipeline needs to not only be useful for predicting dementia in English, but also be reconstructable from Mandarin features.

We modify our pipeline as follows. After training the ElasticNet regressions, we sort the target features by their R^2 (coefficient of determination) measured on the training set, where higher values indicate a better fit. Then, for each K between 1 and the number of features, we select only the top K features and re-train the DB classifier (4.3.1) to only use those features as input. The result of this experiment is shown in Figure 4.3.

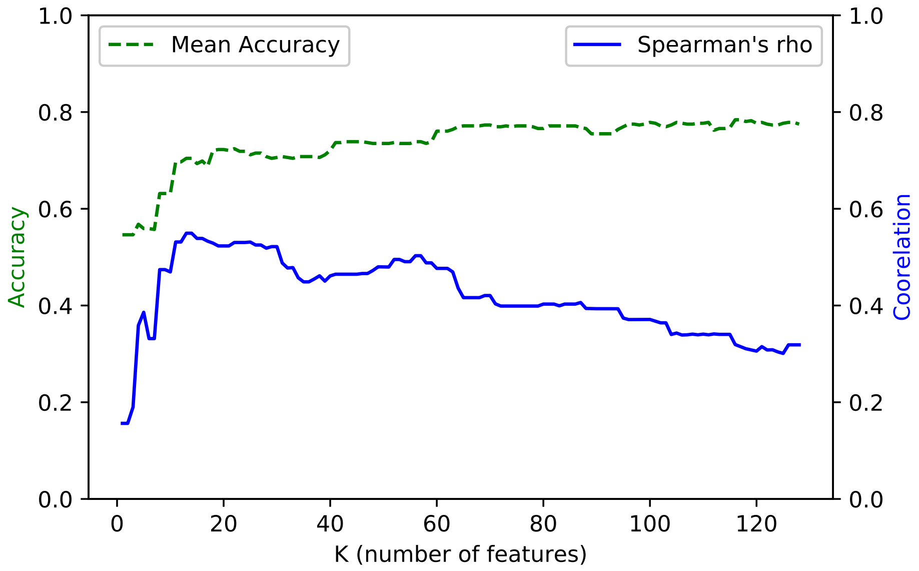

Figure 4.3: Accuracy of DementiaBank classifier model and Spearman's ρ on Lu corpus, using only the top K English features ordered by R^2 on the OpenSubtitles corpus. Spearman's ρ is maximized at $K = 13$, achieving a score of $\rho = 0.549$. DementiaBank accuracy generally increases with more features.

4.4 Experiments

4.4.1 Baseline models

We compare our system against two simple baselines:

1. **Unilingual baseline:** using the Mandarin features, we train a linear regression to predict the dementia score. We take the mean across 5 cross-validation folds.

2. **Translate baseline:** The other intuitive way to generate English features from a Mandarin corpus is by using translation. We use Google Translate[4] to translate each Mandarin transcript to English. Then, we extract features from the translated English text and feed them to the dementia classifier described in section 4.3.1.

There is reason to believe that machine translation may perform poorly on dementia-impaired speech. Koehn and Knowles (2017) showed that neural machine translation models may produce fluent but incorrect translations when given input of a different domain from the training data. Similarly, neural machine translation applied to dementia-impaired input would be expected to produce normalized and grammatically correct output, destroying useful linguistic signals in the process.

4.4.2 Evaluation metric

We evaluate each model by comparing the Spearman's rank-order correlation ρ (Spearman, 1904) between the ground truth dementia scores and the model's predictions (probability estimates from the logistic regression

[4] https://translate.google.com/

model). Spearman's ρ is defined as the Pearson correlation between the ranks of the two variables:

$$\rho(X, Y) = \frac{\mathrm{cov}(X_{\mathrm{rank}}, Y_{\mathrm{rank}})}{\sigma(X_{\mathrm{rank}})\sigma(Y_{\mathrm{rank}})},$$

where X_{rank} and Y_{rank} are the rank variables of X and Y, and σ is the standard deviation function. This measures how monotonic is the relationship between the two variables: a score of $+1$ or -1 means a perfectly monotonic relationship and a score of 0 indicates lack of any relationship. In our case, ρ measures the model's ability to rank the patients from the highest to the lowest severities of dementia.

Pearson's correlation and mean squared error (MSE) are commonly used metrics for evaluating regression models, but in this case, Spearman's ρ is the most appropriate:

- MSE is inappropriate because the model's likelihood predictions are not on the same scale as the dementia score, making direct comparison impossible. This is expected, as the dementia scores are not used during training, only for evaluation. Further, scaling the scores to be on the same scale is incorrect because doing so would introduce data leakage.

- Pearson's correlation compares the raw likelihood scores of the logistic regression model with the dementia score, rather than their relative ranks. However, it is sensitive to nonlinear transformations, thus results would differ based on whether we used the values from the logistic regression before or after the link function. In contrast, Spearman's ρ is invariant to all monotonic transformations.

4.4.3 Experimental results

Model	**Spearman ρ**
Baselines	
Unilingual	0.385
Google Translate	0.366
Our models	
Feature Transfer	0.319
+ Reduced Rank Regression (4.3.3)	0.354
+ Joint Feature Selection (4.3.3)	**0.549**

Table 4.3: Baselines compared with our models, evaluated on the Lu corpus.

Our best model achieves a Spearman's ρ of 0.549, beating the translate baseline ($n = 49$, $p = 0.06$). Hypothesis testing was performed by applying the Fisher transformation to ρ:

$$z = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right).$$

The variable z follows the normal distribution and has known standard deviation:

$$\sigma(z) = \left(\frac{1.060}{n-3} \right)^{1/2}.$$

Thus, the two-tailed z-test is applied to obtain the p-value for the difference between ρ for two models (Fieller et al., 1957).

Joint feature selection appears to be crucial, since the model performs worse than the baselines if we use all of the features. This is the case no matter if we predict each target feature independently or all at once with reduced rank regression. RRR does not outperform the baseline model, probably because it fails to account for the noisy target features in the correspondence model and considers each feature equally important. We did not attempt to use joint feature selection and RRR at the same time, because the multiplicative combination of hyperparameters K and R would produce a multiple comparisons problem using the small validation set.

Using joint feature selection, we find that the best score is achieved when we use $K = 13$ target features (Figure 4.3). With $K < 13$, performance suffers because the DementiaBank classifier is not given enough information to make accurate classifications. With $K > 13$, the accuracy for the DementiaBank classifier improves; however, the overall performance degrades because it is given noisy features with low R^2 coefficients. A list of the top features is given in Table 4.4.

#	**Feature Name**	R^2
1	Number of words	0.894
2	Number of sentences	0.828
3	Brunét's index	0.813
4	Type token ratio	0.668
5	Moving average TTR (50 word window)	0.503
6	Moving average TTR (40 word window)	0.461
7	Moving average TTR (30 word window)	0.411
8	Average word length	0.401
9	Moving average TTR (20 word window)	0.360
10	Moving average TTR (10 word window)	0.328
11	NP → PRP	0.294
12	Number of nouns	0.233
13	Mean length of clause	0.225
14	PP → IN NP	0.224
15	Total length of PP	0.222
16	Complex nominals per clause	0.220
17	Noun ratio	0.213
18	Pronoun ratio	0.208
19	Number of T-units	0.207
20	Number of PP	0.205
21	Number of function words	0.198
22	Subordinate / coordinate clauses	0.193
23	Mean word frequency	0.193
24	Number of pronouns	0.191
25	Average NP length	0.188

Table 4.4: Top English features for joint feature selection, ordered by R^2 coefficients on the OpenSubtitles corpus. The top performing model uses the first 13 features.

In our experiments, the correspondence model worked better when absolute counts were used for the Chinese CFG features (e.g., the number of $NP \rightarrow PN$ productions in the narration) rather than ratio features (e.g., the proportion of CFG productions that were $NP \rightarrow PN$). When ratios were used for source features, the R^2 coefficients for many target features decreased. A possible explanation is that the narrations have varying lengths, and dividing features by the length introduces a nonlinearity that adversely affects our linear models. However, more experimentation is required to examine this hypothesis.

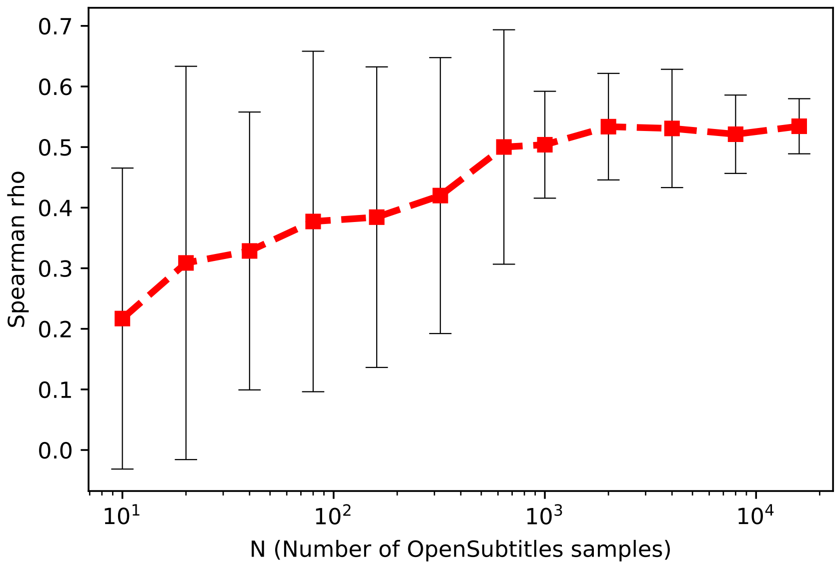

Figure 4.4: Ablation experiment where a various number of OpenSubtitles samples were used for training. The error bars indicate the two standard deviation confidence interval.

4.4.4 Ablation study

Next, we investigate how many parallel OpenSubtitles narrations were necessary to learn the correspondence model. We choose various training sample sizes from 10 to 50,000 and, for each training size, we train and evaluate the whole model from end-to-end 10 times with different random seeds (Figure 4.4). As expected, the Spearman's ρ increased as more samples were used, but only 1000-2000 samples were required to achieve comparable performance to the full model.

4.5 Summary and future work

In this chapter, we first proposed a method to use a large parallel corpus to learn mappings from engineered features in Mandarin Chinese to features in English. We combined this with a dementia classifier trained on DementiaBank; this produces a model to predict dementia in Mandarin Chinese. We evaluated it using the Lu corpus and showed that it outperforms baseline methods. Experiments showed that joint feature selection is crucial for selecting only the most informative features across both parts of the model. Models using all features or reduced-rank regression were less successful. Finally, an ablation experiment showed that the model requires a relatively small amount of data (1000-2000 normative dialogue samples) to achieve its performance. This suggests that it may be applied to low-resource languages, and not only languages for which large parallel corpora are available.

Our model takes an important step towards multilingual dementia detection, but some limitations will need to be resolved before it can be put into clinical use. It requires the narration to be manually transcribed, which is impractical in a clinical setting. This may be handled using automatic speech recognition (ASR), but errors from ASR introduce new challenges into the system. The model only makes use of lexicosyntactic features, and ignores acoustic features like pause duration and tonal contours, which are significant for dementia detection in

English. Also, we have applied our method to Mandarin Chinese, but the same method can be generalized to detect dementia in any language. Perhaps most importantly, the field suffers from poor availability of high-quality multilingual clinical data involving dementia. Good clinical data would render a much larger range of methods to be practical. There is still much to be done; resolving these issues is out of scope for this thesis, and we will leave them for future work.

Chapter 5

Conclusion

In this thesis, we investigated methods of automatically detecting dementia through speech in Mandarin speakers. In Chapter 2, we reviewed previous work on machine learning methods to detect AD and dementia in English, studies on language effects of AD in other languages, and algorithms for domain adaptation. In Chapter 3, we described and analyzed the Lu corpus containing 49 Mandarin Chinese speakers with dementia, and derived a *dementia score* for each patient, for the benefit of downstream applications that study the effects of dementia on narrative speech. Finally, in Chapter 4, we introduced a novel method of transfer learning by combining the Lu corpus with DementiaBank and a corpus of normative parallel dialogue.

One of the biggest challenges in this domain is the sparsity of clinical data, not only in Mandarin Chinese, but in all other languages as well. Despite millions of people worldwide suffering from AD, most studies of this type contain fewer than 100 participants; for the few studies that do exist, their data is private and not easily accessible. Thus, future research will greatly benefit from the collection and curation of high quality and widely accessible datasets (e.g., the AphasiaBank project). Of course, privacy of the participants must be taken into consideration when collecting such a dataset.

Overall, the goal of this research is to develop an accurate method of diagnosing AD in Mandarin Chinese speakers. Although there is no cure for AD, early detection would allow treatment to be administered earlier, reducing the negative impact of the disease on patients' lives. There is still a long way to go before this technology is ready to be deployed in a clinical setting, but this work takes a significant step towards that eventual goal.

Bibliography

Sandra Aluísio, Andre Cunha, and Carolina Scarton. 2016. Evaluating progression of Alzheimer's disease by regression and classification methods in a narrative language test in Portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 109–114. Springer.

Alzheimer's Association. 2018. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429.

Francois Boller and James Becker. 2005. Dementiabank database guide. *University of Pittsburgh*.

Yuen Ren Chao. 1965. *A grammar of spoken Chinese*. Univ of California Press.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

PT Davies. 1982. Procedures for reduced-rank regression. *Applied Statistics*, pages 244–255.

Lixin Duan, Dong Xu, and Ivor W Tsang. 2012. Learning with augmented features for heterogeneous domain adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 667–674. Omnipress.

Bruno Dubois, Harald Hampel, Howard H Feldman, Philip Scheltens, Paul Aisen, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Kaj Blennow, et al. 2016. Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 12(3):292–323.

Edgar C Fieller, Herman O Hartley, and Egon S Pearson. 1957. Tests for rank correlation coefficients. i. *Biometrika*, 44(3/4):470–481.

Marshal F Folstein, Lee N Robins, and John E Helzer. 1983. The mini-mental state examination. *Archives of general psychiatry*, 40(7):812–812.

Kathleen C Fraser, Kristina Lundholm Fors, and Dimitrios Kokkinakis. 2019. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, 53:121–139.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Michel Goedert and Maria Grazia Spillantini. 2006. A century of Alzheimer's disease. *Science*, 314(5800):777–781.

Harold Goodglass and Edith Kaplan. 1983. *Boston diagnostic examination for aphasia*, 2nd edition. Lea and Febiger, Philadelphia, Pennsylvania.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Alan Julian Izenman. 1975. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 701–707.

Judith D Kasper, Vicki A Freedman, Brenda C Spillman, and Jennifer L Wolff. 2015. The disproportionate impact of dementia on family and unpaid caregiving to older adults. *Health Affairs*, 34(10):1642–1649.

Gitit Kavé and Yonata Levy. 2003. Morphology in picture descriptions provided by persons with Alzheimer's disease. *Journal of speech, language, and hearing research*, 46(2):341–352.

Jungwan Kim, HyangHee Kim, Kee Namkoong, Sejoo Kim, and Deogyoung Kim. 2006. Spontaneous speech traits in patients with Alzheimer's disease. *Communication Sciences & Disorders*, 11(3):82–98.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Yi-hsiu Lai. 2014. Discourse features of Chinese-speaking seniors with and without Alzheimer's disease. *Language and Linguistics*, 15(3):411–434.

Yi-hsiu Lai, Hsiu-hua Pai, et al. 2009. To be semantically-impaired or to be syntactically-impaired: Linguistic patterns in Chinese-speaking persons with or without dementia. *Journal of Neurolinguistics*, 22(5):465–475.

Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 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.

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 439–446. Association for Computational Linguistics.

M. Paul Lewis. 2009. *Ethnologue: Languages of the World*, 16th edition. SIL International, Dallas, Texas.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Ching-Ching Lu, Elizabeth Bates, Ping Li, Ovid Tzeng, Daisy Hung, Chih-Hao Tsai, Shu-Er Lee, and Yu-Mei Chung. 2000. Judgements of grammaticality in aphasia: The special case of Chinese. *Aphasiology*, 14(10):1021–1054.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. AphasiaBank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Vaden Masrani, Gabriel Murray, Thalia Shoshana Field, and Giuseppe Carenini. 2017. Domain adaptation for detecting mild cognitive impairment. In *Canadian Conference on Artificial Intelligence*, pages 248–259. Springer.

Andreas U Monsch, Mark W Bondi, Nelson Butters, David P Salmon, Robert Katzman, and Leon J Thal. 1992. Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of neurology*, 49(12):1253–1258.

Zeinab Noorian, Chloé Pou-Prom, and Frank Rudzicz. 2017. On the importance of normative data in speech-based assessment. In *Proceedings of Machine Learning for Health Care Workshop (NIPS MLHC)*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*.

Chloé Pou-Prom and Frank Rudzicz. 2018. Learning multiview embeddings for assessing dementia. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2812–2817.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090.

Daisaku Shibata, Shoko Wakamiya, Ayae Kinoshita, and Eiji Aramaki. 2016. Detecting Japanese patients with Alzheimer's disease based on word category frequencies. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 78–85.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. `https://doi.org/10.5281/zenodo.1443582`. Luminosoinsight/wordfreq: v2.2.

Vanessa Taler and Natalie A Phillips. 2008. Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of clinical and experimental neuropsychology*, 30(5):501–556.

Johannes Tröger, Nicklas Linz, Jan Alexandersson, Alexandra König, and Philippe Robert. 2017. Automated speech-based screening for Alzheimer's disease in a care service scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 292–297. ACM.

Maria Yancheva and Frank Rudzicz. 2016. 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)*, volume 1, pages 2337–2346.

Luke Zhou, Kathleen C Fraser, and Frank Rudzicz. 2016. Speech recognition in Alzheimer's disease and in its assessment. In *INTERSPEECH*, pages 1948–1952.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. 2018. Semi-supervised classification by reaching consensus among modalities. *arXiv preprint arXiv:1805.09366*.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.