A Tale of Two Perplexities: Sensitivity of Neural Language Models to Lexical Retrieval Deficits in Dementia of the Alzheimer's Type

Trevor Cohen*
Biomedical and Health Informatics
University of Washington
Seattle
cohenta@uw.edu

Serguei Pakhomov*
Pharmaceutical Care and Health Systems
University of Minnesota
Minneapolis
pakh0002@umn.edu

Abstract

In recent years there has been a burgeoning interest in the use of computational methods to distinguish between elicited speech samples produced by patients with dementia, and those from healthy controls. The difference between perplexity estimates from two neural language models (LMs) - one trained on transcripts of speech produced by healthy participants and the other trained on transcripts from patients with dementia - as a single feature for diagnostic classification of unseen transcripts has been shown to produce state-of-the-art performance. However, little is known about *why* this approach is effective, and on account of the lack of case/control matching in the most widely-used evaluation set of transcripts (DementiaBank), it is unclear if these approaches are truly diagnostic, or are sensitive to other variables. In this paper, we interrogate neural LMs trained on participants with and without dementia using synthetic narratives previously developed to simulate progressive semantic dementia by manipulating lexical frequency. We find that perplexity of neural LMs is strongly and differentially associated with lexical frequency, and that a mixture model resulting from interpolating control and dementia LMs improves upon the current state-of-the-art for models trained on transcript text exclusively.

1 Introduction

Alzheimer's Disease (AD) is a debilitating neurodegenerative condition which currently has no cure, and Dementia of the Alzheimer's Type (DAT) is one of the most prominent manifestations of AD pathology. Prior to availability of disease-modifying therapies, it is important to focus on reducing the emotional and financial burden of this devastating disease on patients, caregivers, and the healthcare system. Recent longitudinal studies of aging show that cognitive manifestations of future dementia may appear as early as 18 years prior to clinical diagnosis - much earlier than previously believed (Rajan et al., 2015; Aguirre-Acevedo et al., 2016). With 30-40% of healthy adults subjectively reporting forgetfulness on a regular basis (Cooper et al., 2011), there is an urgent need to develop sensitive and specific, easy-to-use, safe, and cost-effective tools for monitoring AD-specific cognitive markers in individuals concerned about their cognitive function. Lack of clear diagnosis and prognosis, possibly for an extended period of time (i.e., many years), in this situation can produce uncertainty and negatively impact planning of future care (Stokes et al., 2015), and misattribution of AD symptoms to personality changes can lead to family conflict and social isolation (Boise et al., 1999; Bond et al., 2005). Delayed diagnosis also results in an estimated $7.9 trillion in medical and care costs (Association, 2018) due to high utilization of emergency care, amongst other factors, by patients with undiagnosed AD.

Cognitive status is reflected in spoken language. As manual analysis of such data is prohibitively time-consuming, the development and evaluation of computational methods through which symptoms of AD and other dementias can be identified on the basis of linguistic anomalies observed in transcripts of elicited speech samples have intensified in the last several years (Fraser et al., 2016; Yancheva and Rudzicz, 2016; Orimaye et al., 2017). This work has generally employed a supervised machine learning paradigm, in which a model is trained to distinguish between speech samples produced by patients with dementia and those from controls, using a set of deliberately engineered or computationally identified features. However, on account of the limited training data available, overfitting is a concern. This is particularly problematic in DAT, where the nature of linguistic anomalies

*denotes equal contribution

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1946–1957

varies between patients, and with AD progression (Altmann and McClung, 2008).

In the current study we take a different approach, focusing our attention on the perplexity of a speech sample as estimated by neural LMs trained on transcripts of the speech of participants completing a cognitive task. To date, the most successful approach to using LM perplexity as a sole distinguishing feature between narratives by dementia patients and controls was proposed by Fritsch et al. (2019) and replicated by Klumpp et al. (2018). The approach consists of training two recurrent neural LMs - one on transcripts from patients with dementia and the other on transcripts from controls. The difference between the perplexities estimated with these two LMs results in very high classification accuracy (AUC: 0.92) reported by both studies.

The explanation for this performance offered by Fritsch et al. (2019) relies on observations that patients with DAT describe the picture in an unforeseen way and their speech frequently diverts from the content of the picture, contains repetitions, incomplete utterances, and refers to objects in the picture using words like "thing" or "something". This explanation, however, conflicts with the findings by Klumpp et al. (2018) that demonstrate similarly high classification accuracy (AUC: 0.91) with a single hidden layer non-recurrent neural network and bag-of-words input features, suggesting that while word sequences play a role, it may not be as large as previously believed by Fritsch et al. (2019). Klumpp et al.'s (2018) explanation contrasts "local" with "global language properties" of the picture descriptions being captured by recurrent neural LMs vs. the non-recurrent bag-of-words neural network classifier, respectively. Both of these explanations are based on informal qualitative observations of the data and are not entirely satisfying because both fail to explain the fact that it is precisely the difference between the control and dementia LMs that is able to discriminate between patients and controls. The individual LMs are not nearly as good at this categorization task.

The objective of the current study is to quantify the extent to which the differences between neural LMs trained on language produced by DAT patients and controls reflect known deficits in language use in this disease - in particular the loss of access to relatively infrequent terms that occurs with disease progression (Almor et al., 1999a). We approach this objective by interrogating trained neural LMs with two methods: *interrogation by perturbation* in which we evaluate how trained neural LMs respond to text that has been deliberately perturbed to simulate AD progression; and *interrogation by interpolation* in which we develop and evaluate hybrid LMs by interpolating between neural LMs modeling language use with and without dementia. We find neural LMs are progressively more perplexed by text simulating disease of greater severity, and that this perplexity decreases with increasing contributions of a LM trained on transcripts from patients with AD, but increases again when only this LM is considered. Motivated by these observations, we modify the approach of Fritsch et al. (2019) by incorporating an interpolated model and pre-trained word embeddings, with improvements in performance over the best results reported for models trained on transcript text exclusively.

2 Background

2.1 Linguistic Anomalies in AD

AD is a progressive disease, and the linguistic impairments that manifest reflect the extent of this progression (Altmann and McClung, 2008). In its early stages, deficits in the ability to encode recent memories are most evident. As the disease progresses, it affects regions of the brain that support semantic memory (Martin and Chao, 2001) - knowledge of words and the concepts they represent - and deficits in language comprehension and production emerge (Altmann and McClung, 2008).

A widely-used diagnostic task for elicitation of abnormalities in speech is the "Cookie Theft" picture description task from the Boston Diagnostic Aphasia Examination (Goodglass, 2000), which is considered to provide an adequate approximation of spontaneous speech. In this task, participants are asked to describe a picture of a pair of children colluding in the theft of cookies from the top shelf of a raised cupboard while their mother distractedly washes dishes[1]. When used as a diagnostic instrument, the task can elicit features of AD and other dementias, such as pronoun overuse (Almor et al., 1999a), repetition (Hier et al., 1985; Pakhomov et al., 2018) and impaired recollection of key elements (or "information units") from the picture (Giles et al., 1996). Due to the human-intensive nature of the analyses to detect such anomalies, automated methods present a desirable alternative.

[1]For a contemporary edition subscribing to fewer gender stereotypes see (Berube et al., 2018).

2.2 Classification of Dementia Transcripts

A number of authors have investigated automated methods of identifying linguistic anomalies in dementia. The most widely-used data set for these studies is the DementiaBank corpus (Becker et al., 1994), which we employ for the current work. In some of the early work on this corpus, Prud'hommeaux and Roark (2015) introduced a novel graph-based content summary score to distinguish between controls and dementia cases in this corpus with an area under the receiver operating characteristic curve (AUC) of 0.83. Much of the subsequent work relied on supervised machine learning, with a progression from manually engineered features to neural models mirroring general Natural Language Processing trends. For example, Fraser and Hirst (2016) report AD classification accuracy of over 81% on 10-fold cross-validation when applying logistic regression to 370 text-derived and acoustic features. In a series of papers, Orimaye et al. (2014; 2017; 2018) report tenfold cross-validation F-measures of up to 0.73 when applying a Support Vector Machine (SVM) to 21 syntactic and lexical features; SVM AUC on leave-pair-out cross-validation (LPOCV) of 0.82 and 0.93 with the best manually-engineered feature set and the best 1,000 of 16,903 lexical, syntactic and n-gram features (with selection based on information gain) respectively; and a LPOCV AUC of 0.73-0.83 across a range of deep neural network models with high-order n-gram features. Yancheva and Rudzicz (2016) derive topic-related features from word vector clusters to obtain an F-score of 0.74 with a random forest classifier[2]. Karlekar et al. (2018) report an utterance-level accuracy of 84.9%[3] with a convolutional/recurrent neural network combination when trained on text alone. While these results are not strictly comparable as they are based on different subsets of the data, use different cross-validation strategies and report different performance metrics, they collectively show that supervised models can learn to identify patients with AD using data from elicited speech samples. However, as is generally the case with supervised learning on small data sets, overfitting is a concern.

2.3 Perplexity and Cognitive Impairment

Perplexity is used as an estimate of the fit between a probabilistic language model and a segment of previously unseen text. The notion of applying n-gram model perplexity (a derivative of cross-entropy) as a surrogate measure of syntactic complexity in spoken narratives was proposed by Roark et al. (2007) and applied to transcribed logical memory (story recall) test responses by patients with mild cognitive impairment (MCI: a frequent precursor to AD diagnosis). In this work, sequences of part-of-speech (POS) tags were used to train bi-gram models on logical memory narratives, and then cross-entropy of these models was computed on held-out cross-validation folds. They found significantly higher mean cross-entropy values in narratives of MCI patients as compared to controls. Subsequent work expanded the use of POS cross-entropy as one of the language characteristics in a predictive model for detecting MCI (Roark et al., 2011).

Perplexity can also be calculated on word tokens and serve as an indicator of an n-gram model's efficiency in predicting new utterances (Jelinek et al., 1977). Pakhomov et al (2010b) included word and POS LM perplexity amongst a set of measurements used to distinguish between speech samples elicited from healthy controls and patients with frontotemporal lobar degeneration (FTLD). A LM was trained on text from an external corpus of transcribed "Cookie Theft" picture descriptions performed by subjects without dementia from a different study. This model was then used to estimate perplexity of elicited speech samples in cases and controls, with significant differences between mean perplexity scores obtained from subjects with the semantic dementia variant of FTLD and controls. However, the authors did not attempt to use perplexity score as a variable in a diagnostic classification of FTLD or its subtypes.

Collectively, these studies suggest elevated perplexity (both at the word and POS level) may indicate the presence of dementia. A follow-up study (Pakhomov et al., 2010a) used perplexity calculated with a model trained on a corpus of conversational speech unrelated to the picture description task, as part of a factor analysis of speech and language characteristics in FTLD. Results suggested that the general English LM word- and POS-level perplexity did not discriminate between FTLD subtypes, or between cases and controls. Taken together with the prior results, these results suggest that LMs trained on transcripts elicited using a defined task (such as the "Cookie Theft" task) are better equipped to distinguish between cases and controls

[2] 0.8 with additional lexicosyntactic and acoustic features.
[3] This improved to 91.1% when incorporating POS tags.

than LM trained on a broader corpus.

As the vocabulary of AD patients becomes progressively constrained, one might anticipate language use becoming more predictable with disease progression. Wankerl et al. (2016) evaluate this hypothesis using the writings of Iris Murdoch who developed AD later in life - and eschewed editorial revisions. In this analysis, which was based on time-delimited train/test splits, perplexity decreased in her later output. This is consistent with recent work by Weiner et al. (2018) that found diminished perplexity was of some (albeit modest) utility in predicting transitions to AD.

The idea of combining two perplexity estimates - one from a model trained on transcripts of speech produced by healthy controls and the other from a model trained on transcripts from patients with dementia - was developed by Wankerl et al. (2017) who report an AUC of 0.83 using n-gram LMs in a participant-level leave-one-out-crossvalidation (LOOCV) evaluation across the DementiaBank dataset. Fritsch et al. (2019) further improved performance of this approach by substituting a neural LM (a LSTM model) for the n-gram LM, and report an improved AUC of 0.92. However, it is currently unclear as to whether this level of accuracy is due to dementia-specific linguistic markers, or a result of markers of other significant differences between the case and control group such as age ($\bar{x}$ = 71.4 vs. 63) and years of education ($\bar{x}$= 12.1 vs. 14.3) (Becker et al., 1994).

2.4 Neural LM perplexity

Recurrent neural network language models (RNN-LM) (Mikolov et al., 2010) are widely used in machine translation and other applications such as sequence labeling (Goldberg, 2016). Recurrent Neural Networks (RNN) (Jordan, 1986; Elman, 1990) facilitate modeling sequences of indeterminate length by maintaining a *state vector*, S_{t-1}, that is combined with a vector representing the input for the next data point in a sequence, x_t at each step of processing. Consequently, RNN-LMs have recourse to information in all words preceding the target for prediction, in contrast to n-gram models. They are also robust to previously unseen word sequences, which with naïve n-gram implementations (i.e., without smoothing or backoff) could result in an entire sequence being assigned a probability of zero. Straightforward RNN implementations are vulnerable to the so-called "vanishing" and "exploding" gradient problems (Hochreiter, 1998; Pascanu et al., 2012), which emerge on account of the numerous sequential multiplication steps that occur with backpropagation through time (time here indicating each step through the sequence to be modeled), and limit the capacity of RNNs to capture long-range dependencies. An effective way to address this problem involves leveraging Long Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), which use structures known as gates to inhibit the flow of information during training, and a mechanism using a memory cell to preserve selected information across sequential training steps. Groups of gates comprise vectors with components that have values that are forced to be close to either 1 or 0 (typically accomplished using the sigmoid function). Only values close to 1 permit transmission of information, which disrupts the sequence of multiplication steps that occurs when backpropagating through time. The three gates used with typical LSTMs are referred to as Input, Forget and Output gates, and as their names suggest they govern the flow of information from the input and past memory to the current memory state, and from the output of each LSTM unit (or cell) to the next training step. LSTM LMs have been shown to produce better perplexity estimates than n-gram models (Sundermeyer et al., 2012).

2.5 Lexical Frequency

A known distinguishing feature of the speech of AD patients is that it tends to contain higher frequency words with less specificity than that of cognitively healthy individuals (e.g., overuse of pronouns and words like "thing") (Almor et al., 1999b). Lexical frequency affects speech production; however, these effects have different origins in healthy and cognitively impaired individuals. A leading cognitive theory of speech production postulates a two-step process of lexical access in which concepts are first mapped to lemmas and, subsequently, to phonological representations prior to articulation (Levelt, 2001). In individuals without dementia, lexical frequency effects are evident only at the second step - the translation of lemmas to phonological representations and do not originate at the pre-lexical conceptual level (Jescheniak and Levelt, 1994). In contrast, in individuals with dementia, worsening word-finding difficulties are attributed to progressive degradation of semantic networks that underlie lexical access at the concep-

tual level (Astell and Harley, 1996). While lexical frequency effects are difficult to control in unconstrained purely spontaneous language production, language produced during the picture description task is much more constrained in that the picture provides a fixed set of objects, attributes, and relations that serve as referents for the the person describing the picture. Thus, in the context of the current study, we expect to find that both healthy individuals and patients with dementia describing the same picture would attempt to refer to the same set of concepts, but that patients with dementia would tend to use more frequent and less specific words due to erosion of semantic representations leading to insufficient activation of the lemmas. Changes in vocabulary have been reported in the literature as one of the most prominent linguistic manifestations of AD (Pekkala et al., 2013; Wilson et al., 1983; Rohrer et al., 2007). We do not suggest that other aspects of language such as syntactic complexity, for example, should be excluded; although, there has been some debate as to the utility of syntactic complexity specifically as a distinguishing feature (see (Fraser et al., 2015)).

3 Materials and Methods

3.1 Datasets

For *LM training and evaluation* we used transcripts of English language responses to the "Cookie Theft" component of the Boston Diagnostic Aphasia Exam (Goodglass, 2000), provided as part of the DementiaBank database (Becker et al., 1994). Transcripts (often multiple) are available for 169 subjects classified as having possible or probable DAT on the basis of clinical or pathological examination, and 99 patients classified as controls.

For *interrogation by perturbation*, we used a set of six synthetic "Cookie Theft" picture description narratives created by Bird et al. (2000) to study the impact of semantic dementia on verb and noun use in picture description tasks. While Bird et al. (2000) focused on semantic dementia, a distinct condition from DAT, these synthetic narratives were not based on patients with semantic dementia. Rather, they were created to manipulate lexical frequency by first compiling a composite baseline narrative from samples by healthy subjects, and then removing and/or replacing nouns and verbs in that baseline with words of higher lexical frequency (e.g., "mother" vs. "woman" vs. "she"). Lexical frequency was calculated using the Celex Lexical Database (LDC96L14) and words were aggregated into groups based on four log frequency bands (0.5 - 1.0, 1.0 - 1.5, 1.5 - 2.0, 2.5 - 3.0: e.g., words in the 0.5 - 1.0 band occur in Celex more than 10 times per million). These narratives are well-suited to the study of lexical retrieval deficits in DAT in which loss of access to less frequent words is observed with disease progression (Pekkala et al., 2013).

In order to *calculate mean log lexical frequency on the DementiaBank narratives*, we used the $\text{SUBTLEX}_{\text{us}}$ corpus shown to produce lexical frequencies more consistent with psycholinguistic measures of word processing time than those calculated from the Celex corpus (Brysbaert and New, 2009). The DementiaBank narratives were processed using NLTK's [4] implementation of the TnT part-of-speech tagger (Brants, 2000) trained on the Brown corpus (Francis and Kucera, 1979). Following Bird et al. (2000) only nouns and verbs unique within the narrative were used to calculate mean log lexical frequency. We did not stem the words in order to avoid creating potentially artificially high/low frequency items. To validate the mean log lexical frequency values obtained with the $\text{SUBTLEX}_{\text{us}}$ corpus, we compared the log lexical frequency means for the six narratives developed by Bird et al. (2000) with their frequency band values using Spearman's rank correlation and found them to be perfectly correlated ($\rho = 1.0$).

The text of DementiaBank transcripts was extracted from the original CHAT files (Macwhinney, 2000). The transcripts as well as the six synthetic narratives were lowercased and pre-processed by removing speech and non-speech noise as well as pause fillers (um's amd ah's) and punctuation (excepting the apostrophe).

3.2 Pre-trained models

Prior work with neural LMs in this context has used randomly instantiated models. We wished to evaluate the utility of pre-training for this task - both pre-training of the LSTM in its entirety and pre-training of word embeddings alone. For the former we used a LSTM trained on the WikiText-2 dataset (Merity et al., 2016) provided with the `GluonNLP` package[5]. 200-dimensional word embeddings, including embeddings augmented with subword information, (Bojanowski et al., 2017) were developed using the `Semantic Vectors` package[6] and

[4] Natural Language Toolkit: www.nltk.org
[5] https://github.com/dmlc/gluon-nlp
[6] https://github.com/semanticvectors/semanticvectors

trained using the skipgram-with-negative-sampling algorithm of Mikolov et al. (2013) for a single iteration on the English Wikipedia (10/1/2019 edition, pre-processed with `wikifl.pl`[7]) with a window radius of five[8]. We report results using skipgram embeddings augmented with subword information as these improved performance over both stochastically-initialized and WikiText-2-pretrained LSTMs in preliminary experiments.

3.3 Training

We trained two sets of dementia and control LSTM models. The first set was trained in order to replicate the findings of Fritsch et al. (2019), using the same `RWTHLM` package (Sundermeyer et al., 2014) and following their methods as closely as possible in accordance with the description provided in their paper. Each model's cross-entropy loss was optimized over 20 epochs with starting learning rate optimization performed on a heldout set of 10 transcripts. The second set was trained using the `GluonNLP` averaged stochastic gradient weight-dropped LSTM (standard-lstm-lm-200 architecture) model consisting of 2 LSTM layers with word embedding (tied at input and output) and hidden layers of 200 and 800 dimensions respectively (see Merity et al. (2017) for full details on model architecture). In training the `GluonNLP` models, the main departure from the methods used by Fritsch et al. (2019) involved not using a small heldout set of transcripts to optimize the learning rate because we observed that the `GluonNLP` models converged well prior to the 20th epoch with a starting learning rate of 20 which was used for all stochastically initialized models. With pre-trained models we used a lower starting learning rate of 5 to preserve information during subsequent training on DementiaBank. All `GluonNLP` models were trained using batch size of 20 and back propagation through time (BPTT) window size of 10. During testing, batch size was set to 1 and BPTT to the length of the transcript (tokens). Unseen transcript perplexity was calculated as e^{loss}.

3.4 Evaluation

As subjects in the DementiaBank dataset participated in multiple assessments, there are multiple transcripts for most of the subjects. In order to avoid biasing the models to individual subjects, we followed the participant-level leave-one-out cross-validation (LOOCV) evaluation protocol of Fritsch et al. (2019) whereby all of the picture description transcripts for one participant are held out in turn for testing and the LMs are trained on the remaining transcripts. Perplexities of the LMs are then obtained on the heldout transcripts, resulting in two perplexity values per transcript, one from the LM trained on the dementia (P_{dem}) and control (P_{con}) transcripts. Held-out transcripts were scored using these perplexity values, as well as by the difference ($P_{con} - P_{dem}$) between them.

3.5 Interrogation of models

For *interrogation by perturbation*, we estimated the perplexity of our models for each of the six synthetic narratives of Bird et al. (2000). We reasoned that an increase in P_{con} and a decrease in P_{dem} as words are replaced by higher-frequency alternatives to simulate progressive lexical retrieval deficits would indicate that these models were indeed capturing AD-related linguistic changes. For *interrogation by interpolation*, we extracted the parameters from all layers of paired LSTM LMs after training, and averaged these as $\alpha LM_{dem} + (1-\alpha) LM_{con}$ to create interpolated models. We hypothesized that a decrease in perplexity estimates for narratives emulating severe dementia would occur as α (the proportional contribution of LM_{dem}) increases.

4 Results and Discussion

The results of evaluating classification accuracy of the various language models are summarized in Table 1. The 95% confidence interval for `GluonNLP` models was calculated from perplexity means obtained across ten LOOCV iterations with random model weight initialization on each iteration. The `RWTHLM` package does not provide support for GPU acceleration and requires a long time to perform a single LOOCV iteration (approximately 10 days in our case). Since the purpose of using the `RWTHLM` package was to replicate the results previously reported by Fritsch et al. (2019) that were based on a single LOOCV iteration and we obtained the exact same AUC of 0.92 on our first LOOCV iteration with this approach, we did not pursue additional LOOCV iterations. However, we should note that we obtained an AUC of 0.92 for the difference between P_{con} and P_{dem} on two of the ten LOOCV iterations with the `GluonNLP` LSTM model. Thus, we believe that the `GluonNLP`

[7] Available at https://github.com/facebookresearch/fastText
[8] Other hyperparameters per (Cohen and Widdows, 2018)

	DEMENTIA		CONTROL		CONTROL-DEMENTIA	
MODEL	AUC	95% CI	AUC	95% CI	AUC	95% CI
$\texttt{RWTHLM}_{LSTM}$	0.80	–	0.64	–	0.92	–
$\texttt{GluonNLP}_{LSTM}$	0.80	$\pm$ 0.002	0.65	$\pm$ 0.002	0.91	$\pm$ 0.004

Table 1: Classification accuracy using individual models' perplexities and their difference for various models.

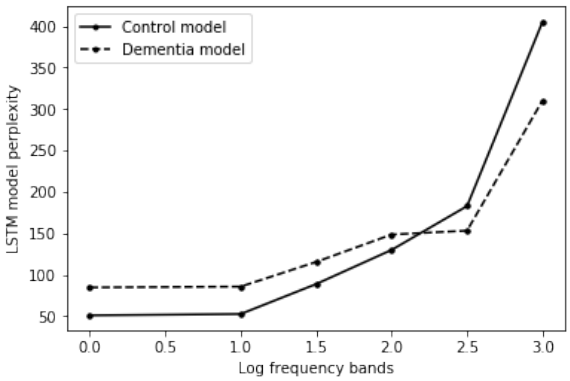

Figure 1: Relationship between log frequency bands used to replace words in synthetic Cookie Theft picture descriptions to simulate degrees of semantic dementia and perplexity of LSTM language models trained on picture descriptions by controls and dementia patients.

LSTM model has equivalent performance to the RWTHLM LSTM model.

Having replicated results of previously published studies and confirmed that using the difference in perplexities trained on narratives by controls and dementia patients is indeed the current state-of-the-art, we now turn to explaining why the difference between these LMs is much more successful than the individual models alone.

First, we used the six "Cookie Theft" narratives designed to simulate semantic dementia to examine the relationship between P_{con} and P_{dem} with GluonNLP LSTM LMs and log lexical frequency bands. The results of this analysis are illustrated in Figure 1 and show that P_{dem} is higher than P_{con} on narratives in the lower log frequency bands (less simulated impairment) and lower in the higher log frequency bands (more simulated impairment).

We confirmed these results by calculating mean log lexical frequency on all DementiaBank narratives and fitting a linear regression model to test for associations with perplexities of the two LMs. The regression model contained mean lexical frequency as the dependent variable and P_{dem} and P_{con} as independent variables, adjusted for age, education and the length of the picture description narrative. In order to avoid likely practice effects across multiple transcripts, we only used the transcript obtained on the initial baseline visit; however, we did confirm these results by using all transcripts to fit mixed effects models with random slopes and intercepts in order to account for the correlation between transcripts from the same subject (mixed effects modeling results not shown).

The results demonstrate that the association between perplexity and lexical frequency is significant and positive for the control LM (coeff: 0.563, $p < 0.001$) and negative for dementia LM (coeff: -0.543, $p < 0.001$). Age, years of education, and length of the narrative were not significantly associated with lexical frequency in this model. These associations show that the control LM and dementia LM are more "surprised" by narratives containing words of higher lexical frequency and lower lexical frequency respectively. If the use of higher lexical frequency items on a picture description task portends a semantic deficit, then this particular pattern of results explains why it is the difference between the two models that is most sensitive to manifestations of dementia and suggests that there is a point at which the two models become equally "surprised" with a difference between their perplexities close to zero. In Figure 1, that point is between log lexical frequency bands of 2.0 and 2.5 corresponding to the mild to moderate degree of semantic impairment reported by Bird et al. (2000). Notably, in the clinical setting, the mild forms of dementia such as mild cognitive impairment and mild dementia are also particularly challenging and require integration of multiple sources of evidence for accurate diagnosis (Knopman and Petersen, 2014).

The results of our interpolation studies are shown in Figure 2. Each point in the figure shows the average difference between the perplexity estimate of a perturbed transcript (Px) and the perplexity estimate for the unperturbed (Po: frequency band 0) sample for this model[9]. While all models tend

[9]We visualized this difference because perplexities at α=0.5 were generally higher, irrespective of whether component models were initialized stochastically, or had pre-trained word embeddings in common. Perplexities of α=0.75 models were slightly lower than those of their majority constituents.

	RANDOM		PRETRAINED		RANDOM		PRETRAINED	
$P_{con} - P_{\alpha}$	AUC	95% CI	AUC	95% CI	ACC_{eer}	95% CI	ACC_{eer}	95% CI CI
$\alpha = 0.25$	0.842	$\pm$ 0.008	0.838	$\pm$ 0.015	0.689	$\pm$ 0.036	0.724	$\pm$ 0.034
$\alpha = 0.5$	0.816	$\pm$ 0.009	0.813	$\pm$ 0.005	0.669	$\pm$ 0.035	0.665	$\pm$ 0.033
$\alpha = 0.75$	**0.931**	$\pm$ **0.003**	**0.941**	$\pm$ **0.006**	**0.854**	$\pm$ **0.031**	**0.872**	$\pm$ **0.010**
$\alpha = 1.0$	*0.908*	$\pm$ *0.004*	0.930	$\pm$ 0.005	*0.846*	$\pm$ *0.023*	0.839	$\pm$ 0.017

Table 2: Performance of randomly-instantiated and pre-trained (subword-based skipgram embeddings) interpolated "two perplexity" models across 10 repeated per-participant LOOCV runs. α indicates the proportional contribution of the dementia model. ACC_{eer} gives the accuracy at equal error rate. Best results are in **boldface**, and results using the approach of Fritsch et al. (2019) are in *italics*.

to find the increasingly perturbed transcripts more perplexing than their minimally perturbed counterparts, this perplexity decreases with increasing contributions of the dementia LM. However, when only this model is used, relative perplexity of the perturbed transcripts increases. This indicates that the "pure" dementia LM may be responding to linguistic anomalies other than those reflecting lack of access to infrequently occurring terms. We reasoned that on account of this, the α=0.75 model may provide a better representation of dementia-related linguistic changes. To evaluate this hypothesis, we assessed the effects on performance of replacing the dementia model with this interpolated model. The results of these experiments (Table 2) reveal improvements in performance with this approach, with best AUC (0.941) and accuracy at equal error rate (0.872) resulting from the combination of interpolation[10] with pre-trained word embeddings. That pre-trained embeddings further improve performance is consistent with the observation that the elevation in perplexity when transitioning from α=0.75 to α=1.0 is much less pronounced in these models (Figure 3). These results are significantly better than those reported by Fritsch et al (2019), and our reimplementation of their approach.

These improvements in performance appear to be attributable to a smoothing effect on the perplexity of the modified dementia models in response to unseen dementia cases. Over ten repeated LOOCV iterations, average perplexity on held-out dementia cases was significantly lower than that of the baseline 'dementia' model (51.1 $\pm$0.81) for both the α=0.75 (47.3$\pm$0.32) and pre-trained embeddings (44.8$\pm$0.53) models. This trend is further accentuated with the severity of dementia - for transcripts corresponding to a mini-mental state

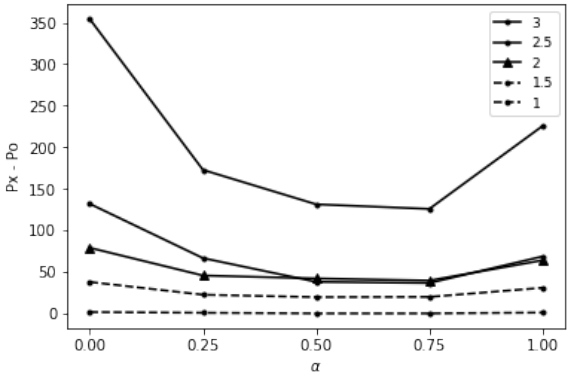

Figure 2: Stochastically initialized models. Elevation in perplexity over unperturbed transcript (Po) with the proportional contribution of a dementia model (α) to an interpolated model. Each point is the mean of 268 (held-out participants) data points. Error bars are not shown as they do not exceed the bounds of the markers.

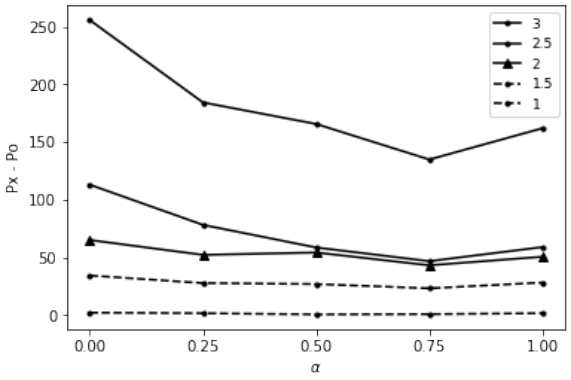

Figure 3: Pretrained word embeddings. Elevation in perplexity over unperturbed transcript (Po) with the proportional contribution of a dementia model (α) to an interpolated model. Each point is the average of 268 data points, and error bars are not shown as they do not exceed the bounds of the markers.

[10]Simply weighting the difference in model perplexities does not perform as well as interpolating model weights, with at best a 0.001 improvement in AUC over the baseline.

exam (MMSE) $\leq$ 10 (n=16), average perplexities are 148.29$\pm$7.69, 105.01$\pm$3.48 and 121.86$\pm$7.67 for baseline 'dementia', α=0.75 and pre-trained embeddings models respectively. In both cases, average perplexity of the interpolated (α=0.75) pre-trained embeddings model fell between those of the exclusively pre-trained (lowest overall) and exclusively interpolated (lowest in severe cases) models.

A practical issue for automated methods to detect dementia concerns establishing their accuracy at earlier stages of disease progression, where a readily disseminable screening tool would arguably have greatest clinical utility, especially in the presence of an effective disease-modifying therapy. To this end, Fritsch et al. (2019) defined a "screening scenario" in which evaluation was limited to participants with a last available MMSE of 21 or more, which corresponds to a range of severity encompassing mild, questionable or absent dementia (Perneczky et al., 2006). In this scenario, classification accuracy of the 'paired perplexity' LSTM based model was only slightly lower (AUC: 0.87) than the accuracy on the full range of cognitive impairment (AUC: 0.92). We found similar performance with our models. When limiting evaluation to those participants with a last-recorded MMSE $\geq$ 21, average AUCs across 10 LOOCV iterations were 0.836 $\pm$0.014, 0.879 $\pm$0.01, 0.893 $\pm$0.004, and 0.899 $\pm$0.012 for the baseline (Fritsch et al (2019)), pretrained embeddings, interpolated (α=0.75) and interpolated (α=0.75) with pretrained embeddings variants, respectively. These results support the notion that paired neural LMs can be used effectively to screen for possible dementia at earlier stages of cognitive impairment.

The contributions of our work can be summarized as follows. First, our results demonstrate that the relationship between LM perplexity and lexical frequency is consistent with the phenomenology of DAT and its deleterious effects on patients' vocabulary. We show that the "two perplexities" approach is successful at distinguishing between cases and controls in the DementiaBank corpus *because of* its ability to capture specifically linguistic manifestations of the disease. Second, we observe that interpolating between dementia and control LMs mitigates the tendency of dementia-based LMs to be "surprised" by transcripts indicating severe dementia, which is detrimental to performance when the difference between these LMs is used as a basis for classification. In addition, we find a similar smoothing effect when using pre-trained word embeddings in place of a randomly instantiated word embedding layer. Finally, we develop a modification of Fritsch et al's "two perplexity" approach that is consistent with these observations - replacing the dementia model with an interpolated variant, and introducing pre-trained word embeddings at the embedding layer. Both modifications exhibit significant improvements in performance, with best results obtained by using them in tandem. Though not strictly comparable on account of differences in segmentation of the corpus amongst others, we note the performance obtained also exceeds that reported with models trained on text alone in prior research. Code to reproduce the results of our experiments is available on GitHub[11].

While using transcript text directly is appealing in its simplicity, others have reported substantial improvements in performance when POS tags and paralinguistic features are incorporated, suggesting fruitful directions for future research. Furthermore, prior work on using acoustic features shows that they can contribute to discriminative models (König et al., 2015); however, Dementia Bank audio is challenging for acoustic analysis due to poor quality and background noise. Lastly, while our results do support the claim that classification occurs on the basis of dementia-specific linguistic anomalies, we also acknowledge that DementiaBank remains a relatively small corpus by machine learning standards, and that more robust validation would require additional datasets.

5 Conclusion

We offer an empirical explanation for the success of the difference between neural LM perplexities in discriminating between DAT patients and controls, involving lexical frequency effects. Interrogation of control- and dementia-based LMs using synthetic transcripts and interpolation of parameters reveals inconsistencies harmful to model performance that can be remediated by incorporating interpolated models and pre-trained embeddings, with significant performance improvements.

Acknowledgments

This research was supported by Administrative Supplement R01 LM011563 S1 from the National Library of Medicine.

[11] https://github.com/treversec/tale_of_two_perplexities

References

Daniel C Aguirre-Acevedo, Francisco Lopera, Eliana Henao, Victoria Tirado, Claudia Muñoz, Margarita Giraldo, Shrikant I Bangdiwala, Eric M Reiman, Pierre N Tariot, Jessica B Langbaum, et al. 2016. Cognitive decline in a colombian kindred with autosomal dominant alzheimer disease: a retrospective cohort study. *JAMA neurology*, 73(4):431–438.

Amit Almor, Daniel Kempler, Maryellen C. MacDonald, Elaine S. Andersen, and Lorraine K. Tyler. 1999a. Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease. *Brain and language*, 67(3):202–227.

Amit Almor, Daniel Kempler, Maryellen C. MacDonald, Elaine S. Andersen, and Lorraine K. Tyler. 1999b. Why do alzheimer patients have difficulty with pronouns? working memory, semantics, and reference in comprehension and production in alzheimer's disease. *Brain and Language*, 67(3):202 – 227.

Lori JP Altmann and Jill S McClung. 2008. Effects of semantic impairment on language use in alzheimer's disease. In *Seminars in Speech and Language*, 01, pages 018–031. © Thieme Medical Publishers.

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

Arlene J. Astell and Trevor A. Harley. 1996. Tip-of-the-tongue states and lexical access in dementia. *Brain and Language*, 54(2):196 – 215.

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.

Shauna Berube, Jodi Nonnemacher, Cornelia Demsky, Shenly Glenn, Sadhvi Saxena, Amy Wright, Donna C Tippett, and Argye E Hillis. 2018. Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with aphasia. *American journal of speech-language pathology*, 28(1S):321–329.

H Bird, MA Lambon Ralph, K Patterson, and JR Hodges. 2000. The rise and fall of frequency and imageability: how the progression of semantic dementia impacts on noun and verb production in the cookie theft description. *Brain and Language*, 73.:17 – 49.

Linda Boise, Richard Camicioli, David L Morgan, Julia H Rose, and Leslie Congleton. 1999. Diagnosing dementia: perspectives of primary care physicians. *The Gerontologist*, 39(4):457–464.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

John Bond, C Stave, A Sganga, O Vincenzino, B O'connell, and RL Stanley. 2005. Inequalities in dementia care across europe: key findings of the facing dementia survey. *International Journal of Clinical Practice*, 59:8–14.

Thorsten Brants. 2000. Tnt - a statistical part-of-speech tagger.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, 41(4):977–990.

Trevor Cohen and Dominic Widdows. 2018. Bringing order to neural word embeddings with embeddings augmented by random permutations (earp). In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 465–475.

Claudia Cooper, Paul Bebbington, James Lindesay, Howard Meltzer, Sally McManus, Rachel Jenkins, and Gill Livingston. 2011. The meaning of reporting forgetfulness: a cross-sectional study of adults in the English 2007 Adult Psychiatric Morbidity Survey. *Age and ageing*, 40(6):711–717.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

W. N. Francis and H. Kucera. 1979. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Kathleen Fraser, Jed Meltzer, and Frank Rudzicz. 2015. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease : JAD*, 49.

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.

Julian Fritsch, Sebastian Wankerl, and Elmar Nöth. 2019. Automatic diagnosis of alzheimer's disease using neural network language models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5841–5845. IEEE.

Elaine Giles, Karalyn Patterson, and John R. Hodges. 1996. Performance on the Boston Cookie theft picture description task in patients with early dementia of the Alzheimer's type: Missing information. *Aphasiology*, 10(4):395–408.

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420.

Harold Goodglass. 2000. *Boston diagnostic aphasia examination: Short form record booklet*. Lippincott Williams & Wilkins.

Daniel B. Hier, Karen Hagenlocker, and Andrea Gellin Shindler. 1985. Language disintegration in dementia: Effects of etiology and severity. *Brain and Language*, 25(1):117–133.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Frederick Jelinek, Robert Mercer, L R Bahl, and J K Baker. 1977. Perplexity - a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62:S63.

Jörg D. Jescheniak and Willem J. M. Levelt. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4):824–843.

Michael I Jordan. 1986. Serial order: A parallel distributed processing approach. Technical report, CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. *arXiv preprint arXiv:1804.06440*.

Philipp Klumpp, Julian Fritsch, and Elmar Nöth. 2018. Ann-based alzheimer's disease classification from bag of words. In *Speech Communication; 13th ITG-Symposium*, pages 1–4. VDE.

David S. Knopman and Ronald C. Petersen. 2014. Mild cognitive impairment and mild dementia: A clinical perspective. *Mayo Clinic Proceedings*, 89(10):1452 – 1459.

Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H. Robert, and Renaud David. 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.

Willem J. M. Levelt. 2001. Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, 98(23):13464–13471.

Brian Macwhinney. 2000. The childes project: Tools for analyzing talk (third edition): Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics - COLI*, 26:657–657.

Alex Martin and Linda L Chao. 2001. Semantic memory and the brain: structure and processes. *Current opinion in neurobiology*, 11(2):194–201.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Sylvester O Orimaye, Jojo SM Wong, Karen J Golden, Chee P Wong, and Ireneous N Soyiri. 2017. Predicting probable alzheimer's disease using linguistic deficits and biomarkers. *BMC bioinformatics*, 18(1):34.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 78–87.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Chee Piau Wong. 2018. Deep language space neural network for classifying mild cognitive impairment and alzheimer-type dementia. *PloS one*, 13(11):e0205636.

Serguei V. S. Pakhomov, Lynn E. Eberly, and David S. Knopman. 2018. Recurrent perseverations on semantic verbal fluency tasks as an early marker of cognitive impairment. *Journal of Clinical and Experimental Neuropsychology*, 40(8):832–840.

Serguei V S Pakhomov, Glenn E Smith, Dustin Chacon, Yara Feliciano, Neill Graff-Radford, Richard Caselli, and David S Knopman. 2010a. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and behavioral neurology : official journal of the Society for Behavioral and Cognitive Neurology*, 23(3):165–177.

Serguei VS Pakhomov, Glenn E Smith, Susan Marino, Angela Birnbaum, Neill Graff-Radford, Richard Caselli, Bradley Boeve, and David S Knopman. 2010b. A computerized technique to assess language use patterns in patients with frontotemporal dementia. *Journal of neurolinguistics*, 23(2):127–144.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2.

Seija Pekkala, Debra Wiener, Jayandra J. Himali, Alexa S. Beiser, Loraine K. Obler, Yulin Liu, Ann McKee, Sanford Auerbach, Sudha Seshadri, Philip A. Wolf, and Rhoda Au. 2013. Lexical retrieval in discourse: An early indicator of alzheimer's dementia. *Clinical Linguistics & Phonetics*, 27(12):905–921. PMID: 23985011.

Robert Perneczky, Stefan Wagenpfeil, Katja Komossa, Timo Grimmer, Janine Diehl, and Alexander Kurz. 2006. Mapping scores onto stages: mini-mental state examination and clinical dementia rating. *The American journal of geriatric psychiatry*, 14(2):139–144.

Emily Prud'hommeaux and Brian Roark. 2015. Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549–578.

Kumar B Rajan, Robert S Wilson, Jennifer Weuve, Lisa L Barnes, and Denis A Evans. 2015. Cognitive impairment 18 years before clinical diagnosis of alzheimer disease dementia. *Neurology*, 85(10):898–904.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Biological, translational, and clinical language processing*, pages 1–8.

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.

Jonathan D. Rohrer, William D. Knight, Jane E. Warren, Nick C. Fox, Martin N. Rossor, and Jason D. Warren. 2007. Word-finding difficulty: a clinical analysis of the progressive aphasias. *Brain*, 131(1):8–38.

Laura Stokes, Helen Combes, and Graham Stokes. 2015. The dementia diagnosis: a literature review of information, understanding, and attributions. *Psychogeriatrics*, 15(3):218–225.

M. Sundermeyer, R. Schlüter, and Hermann Ney. 2014. Rwthlm - the rwth aachen university neural network language modeling toolkit. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2093–2097.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Sebastian Wankerl, Elmar Nöth, and Stefan Evert. 2016. An analysis of perplexity to reveal the effects of alzheimer's disease on language. In *Speech Communication; 12. ITG Symposium; Proceedings of*, pages 1–5. VDE.

Sebastian Wankerl, Elmar Nöth, and Stefan Evert. 2017. An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language. In *INTERSPEECH*, pages 3162–3166.

Jochen Weiner and Tanja Schultz. 2018. Automatic screening for transition into dementia using speech. In *Speech Communication; 13th ITG-Symposium*, pages 1–5. VDE.

Robert S. Wilson, Lynd D. Bacon, Jacob H. Fox, Richard L. Kramer, and Alfred W. Kaszniak. 1983. Word frequency effect and recognition memory in dementia of the alzheimer type. *Journal of Clinical Neuropsychology*, 5(2):97–104.

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.