How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer's disease

Natasha Clarke [a, *], Peter Foltz [b] and Peter Garrard [a]

[a] Neurosciences Research Centre, Molecular & Clinical Sciences Research Institute, St George's, University of London, Cranmer Terrace, London, UK

[b] Institute of Cognitive Science, University of Colorado, Boulder, USA

Abstract

Natural Language Processing (NLP) is an ever-growing field of computational science that aims to model natural human language. Combined with advances in machine learning, which learns patterns in data, it offers practical capabilities including automated language analysis. These approaches have garnered interest from clinical researchers seeking to understand the breakdown of language due to pathological changes in the brain, offering fast, replicable and objective methods. The study of Alzheimer's disease (AD), and preclinical Mild Cognitive Impairment (MCI), suggests that changes in discourse (connected speech or writing) may be key to early detection of disease. There is currently no disease-modifying treatment for AD, the leading cause of dementia in people over the age of 65, but detection of those at risk of developing the disease could help with the identification and testing of medications which can take effect before the underlying pathology has irreversibly spread. We outline important components of natural language, as well as NLP tools and approaches with which they can be extracted, analysed and used for disease identification and risk prediction. We review literature using these tools to model discourse across the spectrum of AD, including the contribution of machine learning approaches and Automatic Speech Recognition (ASR). We conclude that NLP and machine learning techniques are starting to greatly enhance research in the field, with measurable and quantifiable language components showing promise for early detection of disease, but there remain research and practical challenges for clinical implementation of these approaches. Challenges discussed include the availability of large and diverse datasets, ethics of data collection and sharing, diagnostic specificity and clinical acceptability.

* Corresponding author. Neurosciences Research Centre, Molecular & Clinical Sciences Research Institute, St George's, University of London, Cranmer Terrace, London, SW17 0RE, UK.
E-mail addresses: p1607544@sgul.ac.uk (N. Clarke), peter.foltz@colorado.edu (P. Foltz), pgarrard@sgul.ac.uk (P. Garrard).

Key words: Alzheimer's disease, Mild Cognitive Impairment, discourse, Natural Language Processing, Machine Learning

We report no competing interests.

1. Introduction

Natural Language Processing (NLP) and machine learning have changed the way humans and computers interact, making language-processing applications a familiar part of everyday life. Alexa, Siri, and Google Translate all depend on machine learning and NLP algorithms. The growth of NLP has been attributed to recent advances in machine learning algorithms, made possible by greater distributed computing power, large amounts of data available in digital form, and a deeper understanding of the structure of human languages (Hirschberg & Manning, 2015). It is clear, however, that adoption of the technology in the clinical domain is undoubtedly beginning to transform our ability to assess neurodegenerative diseases such as Alzheimer's disease (AD).

Evidence suggests that the build-up of pathology in AD begins decades before symptoms emerge (Jack et al., 2013; Ritchie et al., 2017), so research has become focused on early detection of disease, with the aim of enrolling participants in trials of disease-modifying therapy before pathology is too advanced. Detection of mild cognitive impairment (MCI) is particularly pertinent, as MCI is associated with a ~15% annual risk of dementia compared to 1-2% in unimpaired elderly (Ritchie, 2004). MCI therefore represents an at-risk state for future AD. A reliable AD biomarker (a quantifiable change that correlates with pathological load) would undoubtedly lead to early recognition of disease, but available biomarkers (cerebrospinal fluid assays and amyloid ligand imaging) are invasive, time-consuming and expensive, and therefore not currently candidates for routine or large scale testing (Jack et al., 2013; Lovestone, 2014). Bateman et al. (2019) has argued that plasma amyloid levels could be used as a marker of AD pathology, but the diagnostic potential of this and other blood-borne biomarkers has not been fully evaluated.

Brief cognitive screening tests such as the Mini Mental State Examination (MMSE) (Folstein, Folstein, & McHugh, 1975) and Montreal Cognitive Assessment MoCA (Nasreddine et al., 2005) are inexpensive and quick to administer, but have low predictive value and very low specificity (Arevalo-

Rodriguez et al., 2015). They include minimal assessment of language ability, despite it long being recognised as a feature of AD: Faber-Langendoen et al. (1988) found that 48% of patients with mild AD showed evidence of aphasia on a standard language battery, and Forbes-McKay, Shanks, & Venneri (2013) documented semantic and phonological errors in the language produced by patients with mild and moderate AD, respectively. In the MCI phase, the inclusion of language-based measures in assessment improves accuracy in predicting progression to AD (Bondi et al., 2014; Laske et al., 2015; Oulhaj, Wilcock, Smith, & De Jager, 2009). Demonstration of the linguistic changes of AD could therefore be a sensitive marker of early detection of cognitive decline (Tsantali, Economidis, & Tsolaki, 2013; Bryant, Ferguson, & Spencer, 2016).

There is growing interest in naturally produced language in the form of samples of writing or speech, which are very easily collected and may be more representative of problems encountered in everyday life for individuals living with AD (López-de-Ipiña et al., 2013; Mueller et al., 2018). As manual scoring is slow and reliant on subjective judgement, a key requirement is a means of analysing and interpreting such data rapidly, reliably and at scale (Asgari, Kaye, & Dodge, 2017). NLP and machine learning approaches meet these aims, and could lead to 'flag raising' systems which identify those at risk of disease, such as those at the MCI stage. With the additional potential for remote monitoring, the nature of ongoing assessment could evolve: regular monitoring could be accomplished without the need for hospital visits and without the practice effects that can make cognitive assessment difficult to interpret. Clinical trial methodologies, which currently depend on two or more years of follow-up, could also be revolutionised, leading to shorter, more efficient testing of new dementia treatments.

Although not the focus of this review, we should mention in passing the growing interest in the application of NLP to the large-scale identification and extraction of relevant clinical data from Electronic Health Records (EHRs). In the dementia field this has been applied to identify subsets of patients already diagnosed with dementia, such as those suitable for a clinical trial (Ernecoff, Wessell, Gabriel, Carey & Hanson et al., 2017), or with agitation (Halpern et al., 2018), and to explore potential risk factors automatically (Zhou et al., 2019). Recent collaborative projects are enabling access to thousands of medical records, with overarching goals that include harnessing this complex data, along with other sources, to aid early diagnosis or increase its accuracy, such as identifying features

of misdiagnosis. Dementias Platform UK (DPUK; https://www.dementiasplatform.uk/) enables access to an online portal of rich cohort data, while project iASiS (http://project-iasis.eu/) has a particular focus on NLP techniques, mining EHRs and other records for information that will lead to better decision making at individual and policy levels (Krithara et al., 2019).

This review outlines the contributions of machine learning and NLP to the problem of dementia detection, and is structured according to discourse properties of potential importance. These are considered under the broad headings of individual words (vocabulary), and overall structure (connected language). After defining each feature we review methods and tools for their extraction (summarised in Table 1), and research into their value to predictive models of disease. This is followed by an overview of newer machine learning methods and Automatic Speech Recognition (ASR). To orient the reader to the clinical context, Table 2 lists the research studies employing one or more of these methods, grouped according to the question of clinical concern that they address.

2. Vocabulary

One of the simplest approaches to analysing language is to examine vocabulary, which provides information about the specific kinds of words people use and how those words relate to expected norms of the language being studied. A traditional starting point is the 'bag-of-words' assumption, under which the words in a discourse sample are considered without reference to the order in which they were produced, leaving their inherent lexical or grammatical properties as variables of interest (Jurafsky & Martin, 2017).

2.1. Lexical properties

The most commonly occurring words in any corpus are grammatical function words, or 'closed class', which indicate how a sentence is structured irrespective of its topic, while meaning is provided by content or 'open class' words. Content bearing nouns and verbs have a number of associated lexical properties, the analysis of which can provide information about the complexity of the vocabulary.

Lexical properties include measures of the frequency with which a word appears in discourse, its familiarity to speakers of the language (Balota & Chumbley, 1984), average age-of acquisition and

imageability (the ease with which a word's referent can be pictured or imagined). For example, the word 'ELEPHANT' tends to be acquired early in life and can be easily pictured, compared to 'LEGISLATION'. Values of lexical frequency are derived from large corpora such as the British National Corpus (BNC), a collection of contemporary samples of spoken and written British English that contains a total of 100 million words (The British National Corpus, 2007). Frequency information of content bearing nouns and verbs is derived from their 'lemma' form, which is free from inflection; for example the lemmatised form of 'BLOW', 'BLOWS', 'BLEW' and 'BLOWING' is 'BLOW'. As such researchers should convert words to their lemma form prior to calculation of these metrics so as not to, for example, underestimate the occurrence of a word in a sample.

NLP-based analytical platforms enable automatic extraction of these properties from a text sample. The Tool for the Automatic Analysis of Lexical Sophistication (Kyle & Crossley, 2015) outputs lexical properties across words in a sample compared to the BNC and other databases. Whilst much of the research utilising this tool has centred on evaluating the proficiency of second language acquisition, the possibilities for applying it to clinical language data are clear. Coh-Metrix (Graesser, Namara, Louwerse, & Cai, 2004) also computes lexical properties as part of a larger set of 108 features of a text; we revisit this tool in the coherence & cohesion section (3.2).

Garrard, Maloney, Hodges, & Patterson (2005) found that the mean frequency of words used by the novelist Iris Murdoch - who did not allow editing of her work - in her final novel (completed shortly before she was diagnosed with AD) was significantly higher than those of earlier works of fiction. Le, Lancashire, Hirst, & Jokel (2011) replicated these findings in 20 of Murdoch's novels, using Agatha Christie and P.D. James as comparators, also finding a 'trough' in Murdoch's vocabulary in her late forties to early fifties.

A more recent study by Masrani, Murray, Field, & Carenini (2017) attempted to classify online blog posts as having been written by a person with or without dementia, and found that frequency (estimated using the SUBTL corpus (Brysbaert & New, 2009) was the most informative marker of status. No decline over time was found, however, the sample was small with only six participants, and the dementia group was diagnostically mixed, leaving the study greatly under-powered.

There have been fewer large-scale studies of lexical properties of spoken language, which is less conducive to archiving. Moreover, people tend to produce a much wider range of vocabulary when writing (Crystal, 1987). Bird, Lambon Ralph, Patterson, & Hodges (2000) found that the representations of low imageability words are vulnerable to brain pathology in patients with semantic dementia, but Berisha, Wang, LaCross, & Liss (2015) found no evidence of this in the spoken language of President Ronald Reagan (diagnosed with AD) in the seven years before he left office, suggesting low sensitivity in isolation. The UK Prime Minister Harold Wilson was also diagnosed with AD in later life, and Garrard (2009) found that his word choices converged with those of his colleagues when looked at over a longer time period of ten years.

Two important *caveats* concerning the use of lexical properties are, first, that values (particularly for word frequency) change with fashions in word usage, and secondly that low frequency words are typically under represented (Garrard, 2009). Furthermore, while the discourse of published authors and politicians offers a unique opportunity to analyse language prior to a diagnosis, it cannot be discounted that these individuals may not be representative of the wider population. The longitudinal nature of these studies does allow characterisation of changes with respect to the individual's baseline, however more recent research has focused on applying computational techniques to new, more diverse datasets; we revisit this in section 8.1 'Availability of large and diverse datasets'.

2.2. Grammatical class

Parts of speech provide information about the relative use of grammatical word classes. Nouns, verbs and adjectives are the most familiar, but the often used Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993) includes 45 different parts of speech, including determiners (e.g. 'A' and 'THE'), conjunctions (e.g. 'AND' or 'BUT') and subcategories of nouns and verbs, which can be used as features in machine learning models (see section 6). As many as 20% of words can be assigned to more than one class, largely the highest frequency words in a language, leaving 55-67% of words in a text sample ambiguous out of context (Jurafsky & Martin, 2017).

Automatic part of speech ‘taggers’ are built on the principle of a ‘sequence labelling problem’, and are able to learn features of connected language that give rise to specific tags (classes) using different approaches. For example, using the Python Natural Language Processing Toolkit (NLTK) tags can be assigned using a machine learning algorithm (‘Perceptron tagger’), which learns the context that gives rise to a particular tag from a large corpus, and then applies this knowledge to tag words in a sample (Bird, Loper & Klein, 2009). Current tagging approaches can assign a tag to each word of a language sample with around 97% accuracy (Manning, 2011).

2.3. Richness

In addition to the lexical properties of individual words, the richness of lexical choices in a sample of discourse may also be informative. The richness of President Reagan’s discourse begins to change prior to his leaving office and being diagnosed with AD, with a decline over time in unique words used, and an increase in non-specific nouns (e.g. ‘something’) and fillers (e.g. ‘um’, ‘ah) in press conference transcripts. No change in these measures was detected in the language of his immediate successor George H.W. Bush (Berisha et al., 2015).

The type token ratio (TTR) of a text is a simple measure of the lexical diversity in a sample of text, quantifying the rate of re-use of each unique word in a sample of discourse. Types are the individual words, while tokens are the instances of types. For example, the sentence ‘I like brown dogs and big dogs’ has seven tokens, but only six types (as there are two tokens of the type ‘dogs’), giving it a TTR of 0.9. When calculated over a large window of text, TTR acts as an index of vocabulary size, while TTR at successive windows of text indicates the rate at which words tend to be re-used throughout a sample.

TTR revealed changes over time in the author Iris Murdoch’s vocabulary, which appeared to have diminished by the time she started writing her final book, in which she introduced new words at a slower rate than in earlier works (Garrard et al., 2005; Le et al, 2011). A difference between the first and second halves of the final work of the Dutch author Gerard Reve, also diagnosed with AD in late life, suggested a shrinking vocabulary over the time during which the book was being written (Van Velzen & Garrard, 2008). This contrasts with healthy ageing, in which TTR has been found to

increase with age, suggesting a more diverse vocabulary across the lifespan (Horton, Spieler, & Shriberg, 2010).

As the computation of TTR includes token counts, samples of different lengths cannot be directly compared. Many solutions have been suggested, of which the most commonly adopted has been the Moving-Average Type Token Ratio (MATTR) (Covington & McFall, 2010), which calculates the TTR for a sample of *n* words that moves, one word at a time, from the first *n* to the last *n* words of the text. Windows of varying sizes can be used, and an average calculated.

NLTK includes methods for estimating TTR and MATTR, while the Tool for the Automatic Analysis of Lexical Diversity (Kyle, n.d.) returns a range of measures relating to lexical diversity, including MATTR. When using off-the-shelf NLP tools such as this, users must ensure to understand the computational process from input to output. For example, what pre-processing is required; are features calculated using all words in a sample, or a sub-section such as content words; and what do values in the output indicate - have they been normalised by token count, for example, or if calculated for a window, as in MATTR, do all samples meet the minimum required length.

3. Connected language

Inevitably some information is lost through disregarding word order in a bag-of-words approach, but preserving word order allows investigation of syntactic complexity, coherence & cohesion, and entropy.

3.1. Syntactic complexity

Syntax refers to the rules which govern arrangement of words in a language to create sentences, such as word order. Using these rules a sentence can be ‘parsed’ according to its underlying structure, and the resulting parse visualised and analysed to investigate syntactic complexity. A syntactic parse tree is defined by clauses and sub-clauses within a sentence, and their syntactic relationships (fig. 1), while a dependency parse is defined according to the grammatical relationships between words that ‘depend’ on each other (fig. 2) (Jurafsky & Martin, 2017).

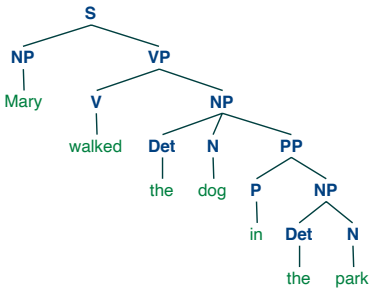

Figure 1. An example syntactic parse tree. Parsed using the NLTK Python library. Syntactic labels: Det = determiner, N = noun, NP = noun phrase, PP = prepositional phase, S = sentence, VP = verb phrase.

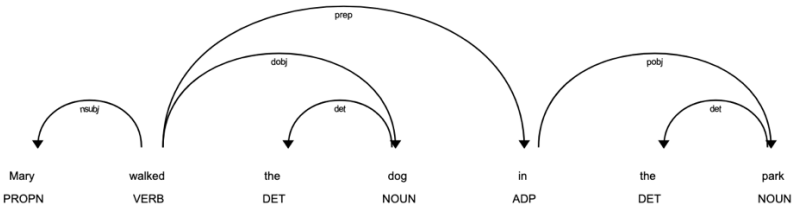

Figure 2. An example dependency parse. Parsed using the spaCy Python library. Parts of speech: ADP = adposition, DET = determiner, NOUN = noun, PROPN = proper noun, VERB = verb. Syntactic dependency labels: det = determiner, nsubj = nominal subject, pobj.

There is no single agreed measure of syntactic complexity, and tools are available that calculate a number of metrics. Lu's L2 Syntactic Complexity Analyzer (Lu, 2010) computes 23 indices, including mean sentence length of a sample and the number of clauses per sentence, also available as part of a larger set using the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (Kyle, 2016). The Computerized Linguistic Analysis System (CLAS) evaluates syntactic complexity by calculating three metrics: Yngve and Frazier scores (two approaches to calculating the depth of a syntactic parse tree) (illustrated in fig. 1), and dependency length, which measures the distance between syntactically related words in a dependency parse (see fig. 2) (Pakhomov, Chacon, Wicklund, & Gundel, 2011).

Pakhomov et al. (2011) used CLAS to analyse passages from four of Murdoch's novels and found accelerated decline while Murdoch was in her mid-thirties to late fifties. This interestingly coincides with the 'trough' in vocabulary found by Le et al. (2011), and represents an abnormally early change given that studies of syntactic complexity in healthy ageing have found it unchanged until mid-seventies (Glosser & Deser, 1992; Marini, Boewe, Caltagirone, & Carlomagno, 2005).

Investigating spoken language, Roark, Mitchell, Hosom, Hollingshead, & Kaye (2011) found that in the MCI phase syntactic complexity was reduced for stories recalled after a delay, but not immediately, suggesting it may be an informative marker when the task has higher cognitive load. Tracking syntactic complexity from the MCI stage to moderate AD, later confirmed at post-mortem, Ahmed et al. (2013) found that reduced syntactic complexity was one of the most frequently observed deficits (along with semantic content), and a linear decline was observed. Machine learning approaches that capture syntactic complexity have been found to successfully distinguish both AD and MCI groups from healthy controls (Orimaye, Wong & Golden, 2014; Orimaye, Wong, & Wong, 2018). Fraser, Meltzer, & Rudzicz (2016) assert that the utility of syntax as an early predictor of AD remains controversial due to variations in findings; this may result from different tools and methods used to quantify syntactic complexity, with not all sensitive to subtle, early change.

3.2. Coherence & cohesion

A semantically coherent piece of discourse follows a theme, or series of themes, which enables the listener (or reader) to follow along. Incoherent language places a higher cognitive load on the listener, (Graesser et al., 2004). Cohesion is an objective property of individual words; cohesive devices aid coherence by cueing the listener and helping them connect ideas, such as anaphora – words that refer back to a preceding clause. Discourse presents a unique opportunity to interrogate these properties (Glosser & Deser, 1991) and automating the analysis permits consistent approaches to characterising the flow of language across the discourse (Foltz, Kintsch & Landauer, 1998).

Traditional approaches to scoring coherence and cohesion suggest that these measures may be sensitive to cognitive decline in AD. Glosser & Deser (1991) found that there were differences between patient and control groups in global (i.e. the whole language sample) but not local (e.g. between adjacent sentences) coherence. Ripich, Carpenter, & Ziol (2000) found a longitudinal decline in cohesion measures for AD patients compared to controls. However, the use of different scoring methods across studies means that results cannot easily be compared, and may be subject to bias.

The Coh-Metrix platform (Graesser et al., 2004), previously mentioned, calculates 12 metrics of cohesion, and has been used in clinical research to explore language in psychosis (Gupta, Hespos,

Horton, & Mittal, 2016; Heidari, D'Arienzo, Crossley, & Duran, 2017), and been adapted for dementia-specific research with Portuguese speaking patients ('Coh-Metrix Dementia'). Using this tool to analyse narrations of the Cinderella story, patients with mild AD were found to have poor global coherence, while those with MCI did not differ from controls (Toledo, Aluisio, & do Santos, 2017). As Coh-Metrix was originally designed to analyse writing, researchers should take care to remove any fillers or markers in a transcript, such as laughter, prior to analysis.

3.3. Entropy & perplexity

In information theory, entropy is used as a measure of the degree of uncertainty within a random variable and is linked to the predictability of a sequence. A sequence with low entropy has high predictability; when previous values are known, subsequent values can be predicted with more certainty. Entropy was first applied to language in 1951 by Claude Shannon (Shannon, 1951), who showed that information in a text could be quantified. Taking each unseen character as a variable, the amount of information inherent in that variable is tied to its predictability when previous characters are known. In language this will depend on higher order considerations (such as context or grammatical correctness) rather than just the rate of co-occurrence of individual letters. For example, in the sequence '*The king married the q…*' there is 100% probability that the unseen variable is '*u*', so its identification does not reduce prior uncertainty. For the next character in the same sequence, there is less predictability, as the sentence could conceivably be '*The king married the quick-witted woman'*. An estimate of the Shannon entropy of a passage of text can be made by averaging the values associated with every character.

Entropy can also be applied at a more coarse-grained, sentence level of analysis. Roark et al. (2011) combined entropy with part of speech tagging to calculate the probability of a particular class given the previous one, and found that, compared to a control group, those with MCI had lower average entropy when immediately retelling a story, suggesting that at a grammatical level their speech was more predictable. Further, Hernández-Domínguez, Ratté, Sierra-Martínez & Roche-Bergua (2018) found that entropy of picture descriptions by patients with MCI or AD, and healthy controls, correlated with scores on the MMSE, suggesting more chaotic or disordered speech as cognition declined.

Closely related to entropy is perplexity, a measure of how accurately the probability distribution of words, word-pairs, word-triplets, etc. (i.e. n-grams) in a sample of text predict the words that appear in an unseen portion of the same text. As with entropy, low perplexity indicates high probability of sample prediction, and equates to the number of possible *n*-grams the model would deem likely (Frankenberg et al., 2019).

Wankerl, Nöth, & Evert (2016) trained bi-gram and tri-gram language models using 90% of sentences from 19 of Murdoch's novels, predicting the remaining 10% of sentences. Perplexity decreased across the final three novels of Murdoch's lifetime, from 1987 until her final novel written in 1995, suggesting that the vocabulary of these later works was less diverse. When analysing only the narrative sections of the novels, the pattern of perplexity showed a steady increase across the lifetime, indicating language growing in complexity across Murdoch's career, before declining.

In a longitudinal analysis of spoken language, Frankenberg et al. (2019) found that in people with MCI or AD baseline perplexity correlated with the MMSE score and information processing speed after approximately ten years. Thus lower perplexity may serve as a useful prognostic indicator of future decline, predicting later severity of cognitive decline, though the sample size was again small, with follow-up data available from only five ADs and 15 MCIs.

4. Semantics

Semantics of a language sample are concerned with the meaning and ideas the speaker or writer wishes to convey. Our understanding of semantic memory and language is due in part to an unusual syndrome, semantic dementia (SD), in which specific atrophy of the anterior temporal lobe leads to speech that is fluent but lacking in meaningful concepts (Landin-Romero, Tan, Hodges, & Kumfor, 2016; Bird et al., 2000). AD pathology also gives rise to a semantic impairment, although not specific nor as pronounced as SD (Libon et al., 2013), and this was detectable at the MCI phase for patients with post-mortem confirmed AD (Ahmed, Haigh, de Jager, & Garrard, 2013).

The Computer Language Analysis (CLAN) cross-platform program can be used to analyse semantic content (MacWhinney, 2000). Originally developed for child language data, CLAN has grown to enable the creation and in-depth analysis of a variety of clinical datasets. Mueller et al. (2018) utilised

CLAN to analyse the spoken language of participants enrolled in the longitudinal Wisconsin Registry for Alzheimer's Prevention (WRAP) study, extracting semantic indices: the percentage of nouns, percentage of verbs, and a pronoun index (number of pronouns divided by the total number of nouns and pronouns), along with other features. Compared to controls, a sub-group displaying subtle cognitive deficits that did not meet the threshold for MCI, termed 'early MCI' (eMCI), were found to decline faster over time in these semantic features, and measures of fluency such as filled pauses, when describing a picture, suggesting that speech was fluent but lacked specific content. Interestingly both groups declined in lexical features, but cognitive status was not an indicator of performance. Reflecting findings of Berisha et al. (2015), measures of speech fluency may be a very early predictor of cognitive decline, possibly due to continued error-monitoring (Mueller et al., 2018).

Standard tests of semantic function, such as picture naming, did not correlate with semantic connected speech measures, suggesting that sampling connected speech – the end product of a number of different cognitive processes - results in a different type of data to stand-alone neuropsychological tests (Mueller et al., 2018). While this investigation in to early MCI reveals changes in language in a potentially at-risk cohort, the pathway of those diagnosed with eMCI is not yet known, so findings cannot be generalised to early AD.

4.1. Latent Semantic Analysis

How do we know what a word means, or the ideas that it represents? Humans can come to know the meanings of words even without direct sensory exposure to the concepts for which they stand. 'Innatist' philosophers, beginning with Plato, argued that this implies that some knowledge is hard-wired into the brain at birth, an idea opposed by Locke and the empiricist school. Landauer & Dumais (1997) suggested that many weak intercorrelations between knowledge domains afford learning through inference, an insight summed up by Firth (1957) in a memorable epigram: "You shall know a word by the company it keeps".

Latent Semantic Analysis (LSA) (Landauer, Foltz & Laham, 1998) uses word co-occurrence trends in large corpora to represent words in a semantic space. Beginning with a large contingency table containing all linguistic episodes (typically paragraphs) in a corpus, and the number of times every

word type occurs in each, singular value decomposition is used to reduce dimensions of the matrix to those which depend on particular groups of words tending to occur together across contexts. The output of the process is a high-dimensional vector space of words, where the distance between each word vector, or 'word embedding', is used as a metric of semantic similarity. For example, the words 'doctor' and 'physician' seldom co-occur, but they often occur in similar contexts as they are close in meaning, leading to their embeddings being close in a semantic space.

An LSA web-based platform developed at CU Boulder (http://lsa.colorado.edu/; see Dennis, 2007 for a user guide) has an online interface and allows selection of different semantic spaces built from contrasting text collections, such as encyclopaedia articles or psychology text books; if the corpus used is not representative of discourse being analysed, the results may be unreliable. A number of metrics are available, including sentence to sentence comparison for measuring coherence. The aforementioned Coh-Metrix platform outputs eight measures based on latent semantic variables of texts.

This approach has many applications in NLP, such as automatically grouping news articles according to content, regardless of whether the same words appear in each. In clinical contexts, LSA has been used in a variety of ways to characterise differences and changes in semantic content in clinical populations. For example, it has been used to characterise coherence of thought to detect the severity of thought disorder in schizophrenia (Elvevåg, Foltz, Goldberg & Weinberger, 2007; Holshausen, Harvey, Elvevåg, Foltz, & Bowie, 2014), predict risk of psychosis in patient populations (Bedi et al., 2015; Rezaii et al., 2019; Rosenstein, Foltz & Elvevåg, 2015) and score semantic fluency in patients with autism spectrum disorders (Prud'hommeaux, van Santen, & Gliner, 2017). Combined with neuroimaging, the application of LSA has shed light on the underlying neural systems that support coherent speech in healthy adults (Hoffman, 2019). In the AD field, Dunn, Almeida, Barclay, Waterreus, & Flicker (2002) found that using LSA word embeddings to compute the similarity between patient's attempts at a story recall and the original passage out-performed traditional hand scoring. Their method had the highest correlation with measures of global cognition, and was not subject to the same floor effects, or potential bias.

4.2. Idea density

Idea density is a metric that quantifies how conceptually rich a sample of language is - how many ideas is a person expressing, and how concisely? To calculate the idea density of a sample, sentences are first segmented in to propositions, before the ratio of propositions to words is calculated, with higher values indicating a greater number of ideas expressed with fewer words. A lower score could indicate the expression of fewer ideas, or the use of more words to express the same number of ideas (Spencer, Craig, Ferguson, & Colyvas, 2012).

The Computerized Propositional Idea Density Rater tool (CPIDR; Brown, Snodgrass, Kemper, Herman, & Covington, 2008) automatically calculates idea density by tagging each word with its corresponding part of speech, before labelling and counting words as an idea if they are the predicate of a proposition (see figure 3). It offers a speech mode for analysis of transcribed speech, which excludes fillers, repetitions and hesitations.

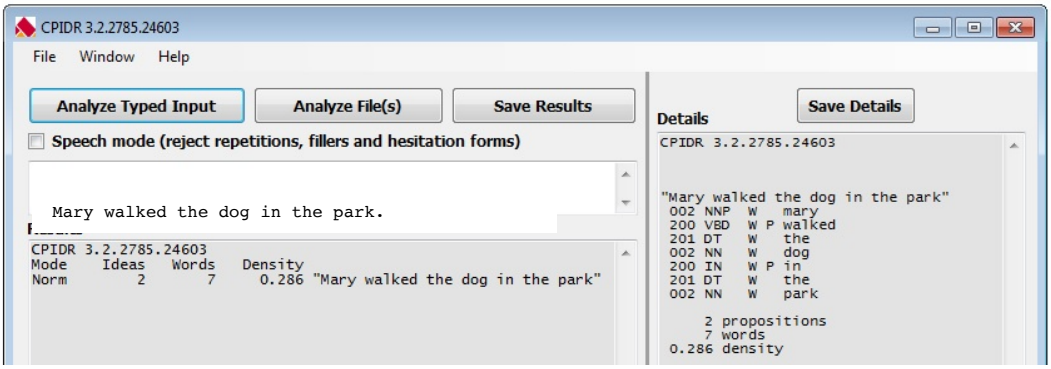

Figure 3. Example output from CPIDR 3.2 for the sentence 'Mary walked the dog in the park'. Output shows that it contains seven words and two propositions, giving an ID of 0.286.

One of the most famous studies investigating language and AD is the Nun Study (Snowdon, 1997). A strong and consistent relationship between idea density of autobiographical essays written on entry to a convent and later cognitive function was found; those who displayed lower idea density at an early age were more likely to have poorer performance on neuropsychological tests years later. Recent computational analysis has replicated these findings in a more representative, yet smaller, AD cohort (Engelman, Agree, Meoni, & Klag, 2010). Chand, Baynes, Bonnici, & Farias (2012) assert four issues which arise from computational analysis of idea density using CPIDR, such as errors at the tagging stage which impact calculation of propositions, but Engelman et al. (2010) found similar results when comparing to manual scoring.

5. Sentiment

The field of NLP has long been concerned with the classification of writing according to the sentiment it expresses, useful for automatically categorising, for example, consumer reviews. In clinical research this approach has been used to detect depression and neurodegenerative disorders using social media posts (Tao, Zhou, Zhang, & Yong, 2016; Wang, Zhang, Ji, Sun, & Wu, 2013). Sentiment can be predicted using machine learning, or using dictionaries of words annotated according to emotional valence, such as Linguistic Inquiry & Word Count (LIWC; Pennebaker, Boyd, Jordan, & Blackburn, 2015). It provides 93 scores relevant to a range of psychological states, personal concerns, and relationship with the past or future, by comparing each word in a sample to its internal dictionaries.

Features relating to time and space were found to be the most important in a study classifying conversations of participants with and without MCI, with 83.33% accuracy (see machine learning section 6), above chance level of 60% given the sample. This dropped to 76.46% on a sub-set of the data matched for education, suggesting the higher education level in the control group played a role in classification (Asgari et al., 2017). Limits of using an annotated dictionary approach should be considered: as words are treated as uni-grams (i.e. without context), negations such as 'I was *not* feeling happy' cannot be taken in to account (Jurafsky & Martin, 2017).

6. Machine learning

Machine learning is a set of computational techniques that aims to learn patterns in data, and apply what has been learnt to generate successful predictions on new data. In clinical medicine, for example, the characteristics of disease are learnt from multiple features, and an individual's disease status predicted given these features (Salvatore & Castiglioni, 2018). Supervised learning is most common, whereby the data used to train and test performance of an algorithm consists of vectors of feature values labelled with their corresponding diagnosis, or 'class'. In the training phase, an algorithm learns weighted values associated with each feature for the class of interest; features which hold predictive value gain large weights, while features of little or no value are smaller or zero respectively. In the testing phase, the algorithm predicts class membership for data that was not seen during training as an indication of performance, typically a portion of the same dataset which is held back. This generalisability to unseen data is key (Raschka, 2015), and in clinical settings will enable predictions in new patients.

Performance metrics include total percentage accuracy, sensitivity and specificity, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC, or AUC), which indicates performance at different thresholds of sensitivity and specificity summarised as a number between zero and one, with 0.5 representing chance level. Accuracy may not be a useful indicator of performance if classes are imbalanced, or the cost of mislabelling as a false negative or false positive is not equal. Health datasets are also often small, and in these cases performance can be estimated using *k*-fold cross-validation, with available data split into *k* different folds of training and test sets, and performance averaged across folds. This can help balance variability in the data set, and detect over-fitting, where a model fits the training data very well, but does not generalise (Hawkins, 2004). Using cross-validation, different model parameters can be tested in order to select the optimal model for good performance on the test set without over-fitting (Schaffer, 1993). For example a regularisation parameter introduces a penalty for weights the model learns, ensuring it does not become too complex (Raschka, 2015).

In a seminal study Fraser et al. (2016) extracted a wide range of 370 linguistic and acoustic features from language samples of 167 participants with AD and 97 controls, part of the 'DementiaBank' Cookie Theft picture description dataset (Becker, Boller, Lopez, Saxton, & McGonigle, 1994). Using cross-validation, a maximum accuracy of 81.92% was achieved using a sub-set of 35 features, selected according to their correlation with the class. A further factor analysis of the top 50 features found four factors: semantic impairment, acoustic abnormality, syntactic impairment and information impairment, in order of variance in the data explained. There was no single profile of impairment, suggesting heterogeneity in linguistic decline possibly due to spread of pathology. Interestingly, values of the semantic and syntactic factors correlated in the control, but not patient, groups, suggesting a decoupling of language abilities in AD (Fraser et al., 2016). Features of the semantic impairment factor were similar to those which Mueller et al. (2018) found differentiated their 'early MCI' group from controls.

Orimaye et al. (2018) used a deep learning, neural network, approach to classify smaller groups from the DementiaBank set. Deep learning represents a subset of complex algorithms which contain an

extra layer, or layers, capable of learning interactions between features and directly from an input, without the need necessarily to first extract features (Najafabadi et al., 2015). An AUC of 0.83 for classifying AD, and 0.80 for MCI (both compared to a control group) was achieved, with models with more layers achieving better performance, demonstrating the effectiveness of deep learning in this domain. However, it is not possible to extract information regarding feature importance from such models, rendering them more of a 'black box' approach (Jarrold et al., 2014), and transparency decreases as the number of layers increases.

Yancheva, Fraser, & Rudzicz (2015) were able to predict MMSE scores from spoken picture descriptions with a mean absolute error of 3.83, which reduced when only participants with multiple, longitudinally obtained, samples were included, evidencing the need for longitudinal sample collection and analysis. Syntactic and semantic features were found to be most predictive of MMSE score, in keeping with other studies of semantic features in connected speech (Ahmed et al., 2013; Rentoumi, Raoufian, Ahmed, de Jager, & Garrard, 2014).

Fraser, Lundholm Fors, Eckerström, Öhman, & Kokkinakis (2019) reported predictions at the individual level, using a classifier to predict an individual's probability of having MCI, and varying the threshold required to obtain the label of MCI to investigate sensitivity and specificity of the model. Overall performance was best when combining predictions from different classifiers built using connected speech features and other tasks, achieving an accuracy of 84% and AUC of 0.90. Moving away from group level predictions, towards those at the individual level, will have greater impact for measuring clinical risk, prognosis and treatment response.

6.1. Neural word embeddings

Similar to LSA (section 4.1), neural word embeddings represent the meaning of words in a high-dimensional vector space, but are built using deep learning. In an early approach called Word2Vec, a model is trained to predict either a target word in the centre of a window given the context (a continuous bag-of-words model), or the context of a window given the target word (skip-gram model). The learned weights of this model are used to build a high-dimensional semantic space, in which each word is represented by a unique vector (Mikolov, Chen, Corrado, & Dean, 2013). Word2Vec has

been successfully used to measure prose recall in schizophrenia and predict classes of patients and controls (Chandler, Foltz, Cheng, et al., 2019). In the GloVe ('global vectors') approach, global co-occurrence statistics across the whole corpus are utilised along with a smaller window looking at context (Pennington, Socher, & Manning, 2014). The Python Gensim library (Rehurek & Sojka, 2010) enables a user to build a custom semantic space from their own corpora, using Word2Vec, GloVe and LSA approaches, or utilise pre-trained word embeddings. All, however, entail important pre-processing steps (Iter, Yoon, & Jurafsky, 2018).

Using the average word embedding of a sample, representing average 'meaning', Mirheidari et al. (2018) achieved only 69.8% accuracy classifying controls vs AD in the DementiaBank dataset, suggesting a loss of accuracy when lexico-syntactic properties are ignored. This is supported by Yancheva & Rudzicz (2016), who found that adding automatically generated semantic topics from clusters of word embeddings to features utilised in Fraser et al. (2016) improved performance, using the same DementiaBank dataset to classify controls and AD patients. Weissenbacher et al. (2016) also automatically generated semantic content information, using word embeddings to find words with similar meaning to those used by controls, ensuring a comprehensive score. Added to other features, they achieved 86.1% accuracy classifying patients with AD and MCI compared to controls.

Investigating MCI only, Fraser, Lundholm Fors, & Kokkinakis (2019) found that using a multilingual approach, including data from both English and Swedish speakers when creating information topics using word embeddings, improved model performance. Additional data from patients of another language was more effective than additional data from healthy controls of the same language, with overall accuracy reaching 72% using information content. Thus, while semantic information captured using neural embeddings alone may not lead to optimal detection of AD, these methods are being utilised in innovative ways to automate steps in analysis and augment data sets.

In a different task, Mirheidari et al. (2018) achieved 100% accuracy on a small dataset classifying groups of patients diagnosed with any neurodegenerative disorder, or 'functional memory disorder' (i.e. lacking an organic cause), using conversations with an 'intelligent virtual agent' (IVA) asking similar questions to a Neurologist. Whilst an important classification, due to the need to make ongoing

referral decisions in primary care (Mirheidari et al. (2018), it is not clear how well this model would generalise to a larger dataset, or data collected under less well-controlled conditions.

There are limitations to the approaches outlined to creating word embeddings, such as an inability to model polysemy, where the same word has multiple meanings. Newer approaches seek to overcome this issue: ELMo (Embeddings from Language Models), which learns word embeddings from a whole sentence (Peters et al., 2018), BERT (Bidirectional Encoder Representations from Transformers; Devlin, Chang, Lee, & Toutanova, 2018) and EARP (Embeddings Augmented by Random Permutations; Cohen & Widdows, 2018) consider word order through deeper neural networks, capturing context dependent differences in vectors. These newer techniques are performing at state of the art levels in a range of language tasks, and although are yet to be applied to the field of dementia (to our knowledge), may lead to increased accuracy and precision when modelling AD discourse.

7. Automatic Speech Recognition

To fully automate the process of diagnostic classification and scoring, language samples need to be quickly and accurately transcribed, a goal that can be achieved through ASR, which uses a range of computational methods, including machine learning to automatically generate words from audio recordings, and can circumvent the need for human transcribers, which is costly and unscalable (Zhou, Fraser, & Rudzicz, 2016). As current ASR systems are not 100% accurate, work has investigated their utility in clinical fields.

Early studies suggested that its use may negatively impact subsequent machine learning classification tasks, with performance dropping as errors in ASR transcription, measured using the Word Error Rate (WER), increase (Lehr, Prud, Shafran, & Roark, 2012; Zhou et al., 2016), More recently, a deep learning ASR approach led to an increase in classification accuracy of up to 22% in some, but not all tasks using the ‘IVA’ dataset described above (Mirheidari et al., 2018). Thus both approach and dataset quality may be key, with improvements in these domains leaving researchers with more time and resources to collect data.

Outside of the research setting use of ASR in the dementia field may be problematic as its accuracy is particularly effected by age of voice and frailty. WERs gradually increase with age (Vipperla, Renals, & Frankel, 2008) and were found to be 10-12% higher for older voices than adults (Pellegrini et al., 2012; Vipperla, Renals, & Frankel, 2010). These errors may 'propagate' downstream (Errattahi, Hannani, & Ouahmane, 2018). Adapting ASR systems for older voices can help to reduce errors: Zhou et al. (2016) found that using a small, domain specific dataset led to fewer errors than using large, out-of-domain data, and Kwon, Kim & Choeh (2016) improved accuracy by preprocessing data in-line with elderly speech patterns. Given that early detection of AD will rely on ASR capabilities in adult voices, as opposed to older, current systems may be appropriate.

8. Discussion

We have described the most important advances in NLP and machine learning, and shown how these have stimulated interest in computational studies of the impact of AD on discourse, with research moving towards answering clinically important questions quickly, objectively and with reproducible results. The tools and approaches available are expanding, with developments in neural approaches opening new pathways for investigation of discourse. They show potential to be deployed as clinical applications, but they also hold the promise of helping to better understand the underlying mechanisms of AD as they are manifested through the assessment of different components of language. However, research outlined throughout remains retroactive. For these tools to be used effectively and implemented into practice, there still remain several practical issues that need to be overcome before the field can progress from research to clinical application, and we outline these below.

8.1. Availability of large and diverse datasets

Datasets suitable for NLP analysis are scarce and most often consist of samples of spoken or written English (Fraser, Lundholm Fors, & Kokkinakis, 2019). Many studies have been performed on a small sample of authors, or using the DementiaBank dataset, and thus results may not generalise to other populations, with variation in education, age, and language. This is particularly problematic for machine learning studies which train and test a model on one dataset, as algorithms can be very 'brittle', with performance dropping when applied to new data, such as in a clinical setting. To obtain an accurate measure of performance, algorithms could be tested on a separately collected dataset.

While *more* data for training and testing algorithms within this domain may increase performance, *diverse* datasets are required for results that will generalise to the clinic; novel methods such as augmenting datasets using other languages are providing promising results (Fraser, Lundholm Fors, & Kokkinakis, 2019).

In terms of stimuli, there has been much focus on the cookie theft picture, which may miss important features of different or longer discourse samples, and Fraser, Lundholm Fors & Eckerström, et al., (2019) found better results combining different tasks. Longer samples require time consuming and costly transcription, though ASR promises automation of the analytic pipeline. There is also still relatively little known about how language changes across the lifetime in healthy ageing for the wider population. Better normative data for specific linguistic features outlined will enable more accurate interpretation of clinical results, with longitudinal studies, such as WRAP which includes language samples (Johnson et al., 2018) meaning we are closer to achieving these aims.

8.2. Ethics of data collection & sharing

Collection of the large and diverse datasets required entails ethical guidelines for data collection and storage to ensure participant safety and protection of personal information. While these constraints do not apply to openly shared data (e.g. blog posts), diagnoses are less reliable and production conditions unknown. To achieve the large, diverse datasets necessary to advance the field, sharing of data amongst researchers is crucial. Fraser, Linz, Lindsay, & König (2019) outline these complex issues in depth, along with examples of good data sharing and recommendations, such as obtaining consent from participants for re-use of data, and considering the type of discourse samples collected, such as avoiding personal histories to maintain privacy.

8.3. Diagnostic specificity

One intrinsic problem with detecting AD through discourse is that disease can only be confirmed at post-mortem. Clinical diagnosis is often difficult; for those diagnosed with AD in life, sensitivity to post-mortem thresholds for disease have been found to range from between 70.9% and 87.3%, and specificity between 44.3% and 70.8% (Beach, Monsell, Phillips, & Kukull, 2012). A few studies have utilised data from post-mortem confirmed AD, such as Ahmed et al. (2013) and studies of Iris Murdoch (e.g. Garrard et al., 2005), but most studies listed in Table 2 lack this diagnostic certainty.

The increasing availability of brain tissue through brain banking could lead to increased diagnostic specificity not only in computational linguistics, but all forms of clinical research.

8.4. Clinical acceptability

Once in the clinic, for a dementia 'flag-raising' tool to be useful its use must be acceptable by both clinicians and patients. Trust in a tool or approach being used is key, and we break this down to two important factors: interpretability and accountability. Interpretability involves the level at which a tool's output, such as the decision a patient is at risk of disease, can be explained – how was this decision reached? How does it map on to clinical understanding of disease mechanisms and symptomatology? There is usually a trade-off between interpretability and performance; deep learning methods can achieve higher performance than traditional machine learning approaches, but their interpretability is low (Chandler, Foltz, & Elvevåg, 2019). Research which attempts to open this black box is gathering pace, but must be a consideration for clinical utility. In terms of accountability, who is accountable if something goes wrong, such as an error in the system that leads to a false negative, or a data breach? Clear accountability will help foster trust.

Patients may have additional concerns regarding acceptability, such as around transparency in how their data will be collected and used; whilst there are already laws and frameworks in place for clinical data, the use of technology, for example with remote monitoring, brings new challenges. The ease of use of a tool, or its intrusiveness, should also be considered. Some of these questions are starting to be addressed; Mirheidari et al. (2019) found that patients found it less intimidating to speak to a 'virtual' neurologist on a screen than to a human examiner. Others, such as around accountability, are only now being discussed, and have some way to go before decisions are reached. Contributions from all 'stake-holders', including researchers, clinicians, patients, health services and commercial enterprises, will likely be required. The current and future challenges outlined are all linked - good generalisability will foster trust, as an algorithm will not be biased or more unreliable for certain groups, be that gender, ethnicity, or age or socioeconomic status (Chandler, Foltz, & Elvevåg, 2019).

There is a wider question of whether early detection of disease is beneficial for patients given that there is currently no disease modifying treatment, however, as well as the need to identify those at

risk to enable drug development and testing, it is widely considered an acceptable goal, and research has found that patients largely welcome early detection (Prince, Bryce & Ferri, 2011; Department of Health 2012; Alzheimer's Research UK, 2019).

9. Conclusion

Undoubtedly NLP and machine learning techniques, and their application to AD, is gathering pace. The wealth of open source tools should enable greater homogeneity and reproducibility of methods, with researchers able to access tools and share code to study language features objectively, without the necessity for advanced programming skills. The techniques can provide insights into the underlying neurobiology of language as well as practical tools. This enables a multidisciplinary field, which benefits from clinical knowledge, although as we have outlined, researchers must understand processes and pitfalls of different approaches. Newer language models, investigated in larger cohorts, may bring new insights in to early language change in AD and MCI, leading to increased detection of at-risk individuals and optimal monitoring and assessment post-diagnosis. However, it may be some time before findings have clinical implications, given the challenges yet to overcome.

Tool	Requirements	Metrics	Reference
Coh-metrix 3.0	Online interface (Firefox or Chrome browser)	108 indices of cohesion, syntactic complexity, lexical diversity, word information and readability metrics	Graesser et al., (2004)
Computer Language Analysis (CLAN)	MAC, Windows or UNIX. Transcripts must follow specific guidelines	Enables basic and complex analysis, see manual for details	Macwhinney (2000)
Computerized Linguistic Analysis System (CLAS)	Java based	Measures of syntactic complexity	Pakhomov et al., (2011)
Computerized Propositional Idea Density Rater (CPIDR 3.2)	Mac, Windows, UNIX or LINUX	Propositional idea density	Brown et al., (2008)
Gensim	Proficiency in Python coding language	Python library which enables training of word embeddings using a variety of approaches including Word2Vec, or loading of pre-trained embeddings	Rehurek & Sojka (2010)
L2 Syntactic Complexity Analyzer (L2SCA 3.3.3)	Implemented in Python, runs on LINUX, MAC or UNIX systems with Java installed. Online interface with batch mode available	14 indices of syntactic complexity	Lu (2010)
Latent Semantic Analysis @ CU Boulder	Online interface	5 applications of LSA, with difference semantic spaces available	Dennis (2007)
Lexical Complexity Analyzer (LCA)	Implemented in Python on LINUX, MAC or UNIX. Online interface with single (allows comparisons of two texts) or batch mode	25 measures of lexical density, variation and sophistication	Lu (2012)
Linguistic Inquiry & Word Count (LIWC2015)	Purchase of license. MAC or Windows	Over 90 indices relating to POS, psychological constructs and language markers	Pennebaker et al., (2015)

NLTK 3.4	Proficiency in Python coding language	Simple processing steps like tokenizing or tagging, to more complex code	Bird et al., (2009)
Tool for the Automatic Analysis of Lexical Diversity (TAALED, 1.3.1)	MAC, Windows or Python	A wide variety of lexical diversity metrics including TTR and MATTR	Kyle (n.d)
Tool for the Automatic Analysis of Lexical Sophistication (TAALES 2.2)	MAC, Windows or LINUX	Over 400 classic and new indices of lexical sophistication	Kyle & Crossley (2015)
Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC 1.3.8)	MAC, Windows or Python	372 indices of syntactic complexity, including those calculated by L2SCA	Kyle (2016)

Table 1. NLP tools available for computation of different linguistic features, referenced throughout this review.

Example studies	Dataset & task	NLP Approach	Sample size	Best performance
Question 1: Does this patient have dementia?				
Weissenbacher et al., (2016)	Arizona Alzheimer's Disease Center (ADC) Written picture description	Word embedding features & other computationally extracted linguistic variables	154 HC, 47 MCI or AD	HC vs MCI+AD: 86.1% accuracy
Fraser et al., (2016)	DementiaBank Spoken Cookie Theft	Computational extraction of 370 speech & linguistic features, with factor analysis	97 HC, 167 AD	HC vs AD: 81.92% accuracy
Masrani et al., (2017)	6 online blogging sites	Computational extraction of linguistic features	3 HC, 2 AD, 1 Dementia with Lewy Bodies	HC vs dementia: 84.8 AUC
Toledo et al., (2017)	Universidade de São Paulo Cinderella Story narration	Computational extraction of linguistic features (Coh-Metrix Dementia tool)	20 HC, 20 amnestic MCI (aMCI), 20 mild AD	Significant differences found between mild AD and other groups. Unable to detect statistical differences for features between HC and aMCI (question 2)
Mirheidari et al., (2018)	DementiaBank (Spoken Cookie Theft) Hallam (Neurologist & patient conversations) IVA ('Virtual Neurologist' & patient conversations)	ASR or manual transcription. Word embedding features for classification	473 audio/text files 45 conversations 18 conversations (participant breakdown not specified)	DementiaBank HC vs AD: 69.8% accuracy Hallam HC vs FMD: 70.8% FMD vs DPD: 93.7% HC vs DPD: 75.9% IVA HC vs FMD: 100% FMD vs MCI: 75% (question 2) HC vs MCI: 81.25% (question 2)
Question 2: Is this patient at risk of developing Alzheimer's disease?				

Garrard et al., (2005)	3 novels written by Iris Murdoch	Computational & manual extraction of linguistic variables	1 AD	Significantly higher average word frequency detected in final novel, prior to AD diagnosis
Garrard (2009)	Unscripted speeches by Harold Wilson	Computational extraction of frequency features	1 AD, other speakers in the House of Commons pooled as controls (*n* unknown)	Word frequency in AD speech more similar to controls in later years, suggesting a change over time
Engelman et al., (2010)	Precursor's Study Medical School Admission Essays	Idea density (CPIDR tool)	36 HC, 18 AD	Higher idea density significantly lowered odds ratio for AD (OR=0.16)
Roark et al., (2011)	Layton Aging & Alzheimer's Disease Center Spoken story recall	Computational extraction of speech & linguistic variables	37 HC, 37 MCI	86.1 AUC
Lehr et al., (2012)	Oregon Health and Science University's Layton Aging and Alzheimer's Disease Center Spoken story recall	ASR to transcribe recordings. Computational alignment of recall to original story for feature extraction	37 HC, 35 MCI	80.9 AUC AUC decreased as WER increased
Wankerl et al., (2016)	19 novels written by Iris Murdoch	Perplexity language model	1 AD	Perplexity decreased across the final 3 novels of Murdoch's career, starting 10 years prior to her AD diagnosis
Asgari et al., (2017)	Participants enrolled in a clinical trial Spoken semi-structured interview	Sentiment features (LIWC tool)	27 HC, 14 MCI	HC vs MCI: 83.33% accuracy (chance=60%) 76.46% on education-matched subset
Orimaye et al., (2018)	DementiaBank Spoken Cookie Theft	Deep learning language model	99 HC, 99 AD, 19 MCI	HC vs AD: 83.0 AUC HC vs MCI: 80.0 AUC (subgroup of 19 matched HC)

Frankenberg et al., (2019)	ILSE Spoken semi-structured interview	Perplexity language model	31 HC, 15 MCI, 5 AD	Perplexity significantly correlated with cognition measures 10-12 years later for the patient group, but not HC
Fraser, Lundholm Fors, Eckerström, Öhman & Kokkinakis (2019)	Gothenburg MCI Study Spoken picture description, eye tracking, neuropsychological testing	Computational extraction of linguistic variables	29 HC, 26 MCI	HC vs MCI: 84% accuracy, 0.90 AUC (combining multiple tasks & classifiers)
Fraser, Lundholm Fors & Kokkinakis (2019)	Gothenburg MCI Study, Karolinska corpus & DementiaBank Spoken picture description	Word embedding features in a multi-lingual approach (Swedish & English)	229 HC, 50 MCI	HC vs MCI: 72% accuracy for Swedish speakers (0.77 sensitivity 0.69 specificity) 63% accuracy for English speakers (0.53 sensitivity, 0.74 specificity)
Question 3: Is there linguistic change over time?				
Le et al., (2011)	51 novels written by 3 renowned authors	Computational extraction of linguistic variables	1 HC, 1 AD, 1 suspected AD	Trough in vocabulary & syntactic complexity in AD in late 40's-early 50s
Pakhomov et al., (2011)	4 novels written by Iris Murdoch	Syntactic features (CLAS tool)	1 AD	Significant decline in some measures over the lifespan, but not all, with accelerated decline in the middle of her career
Ahmed et al., (2013)	OPTIMA Spoken Cookie Theft	Computational extraction of linguistic variables	9 HC, 9 AD	Significant linear trends in five discourse composite scores from MCI to moderate AD Post-mortem confirmed AD (question 4)
Van Velzen et al., (2014)	78 novels written by 6 renowned authors	Computational extraction of linguistic variables	3 HC, 2 AD, 1 suspected AD	Significant decrease over time for noun:pronoun ratio for authors diagnosed or suspected of AD, but not HC

Berisha et al., (2015)	Transcripts of President's Reagan & Bush speech Q&As	Computational extraction of linguistic variables (NLTK)	1 HC, 1 AD	A significant change was found in linguistic variables for Regan only, from transcripts 13 years prior to AD diagnosis
Mueller et al., (2018)	Wisconsin Registry for Alzheimer's Prevention (WRAP) Spoken Cookie Theft	Computational extraction of linguistic variables	200 HC, 64 early MCI (eMCI)	eMCI declined faster over time in measures of speech fluency & semantic content compared to HC, but not grammatical complexity & lexical diversity
Question 4: In the presence of linguistic change, what is the underlying pathology?				
Rentoumi et al., (2014)	OPTIMA Spoken Cookie Theft	Computational extraction of linguistic variables	18 pure AD (ADp), 18 mixed AD pathology (ADm)	ADp vs ADm: 75% accuracy
Question 5: In the presence of cognitive change, what is the degree of linguistic decline?				
Yancheva et al., (2015)	DementiaBank Spoken Cookie Theft	Computational extraction of 477 linguistic features	90 HC, 165 AD	Linguistic features predicted MMSE scores with a MSE of 3.83
Yancheva & Rudzicz (2016)	DementiaBank Spoken Cookie Theft	Automatically generated information content units, lexico-syntactic & acoustic features	98 HC, 168 AD	HC vs AD: Standard features: 76% accuracy Information content unit features: 74% accuracy Feature sets combined: 80% accuracy
Hernández-Domínguez et al., (2018)	DementiaBank Spoken Cookie Theft	Computational extraction of speech & linguistic variables	74 HC, 19 MCI, 169 AD	A number of measures significantly correlated with cognitive severity & MMSE score Study also answers question 1 for HC vs AD: 94% accuracy & 93.0 AUC, & HC vs MCI+AD: 87% accuracy, 87.0 AUC

Table 2. Studies utilising NLP methods to investigate connected language in AD, organised according to five questions of clinical interest.

References

Ahmed, S., Haigh, A. M. F., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, *136*, 3727–3737. https://doi.org/10.1093/brain/awt269

Alzheimer's Research UK. (2019). *Detecting and diagnosing Alzheimer's disease: Enhancing our understanding of public attitudes to improving early detection and diagnosis.* https://www.alzheimersresearchuk.org/about-us/our-influence/policy-work/reports/detecting-diagnosing-alzheimers-disease-2020/

Arevalo-Rodriguez, I., Smailagic, N., Roqué i Figuls, M., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., … Cullum, S. (2015). Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database Syst Rev*, (3), 1–20. https://doi.org/10.1002/14651858.CD010783.pub2.www.cochranelibrary.com

Asgari, M., Kaye, J., & Dodge, H. (2017). Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, *3*(2), 219–228. https://doi.org/10.1016/j.trci.2017.01.006

Balota, D. A., & Chumbley, J. I. (1984). Are Lexical Decisions a Good Measure of Lexical Access? The Role of Word Frequency in the Neglected Decision stage. *Journal of Experimental Psychology*, *10*(3), 340–357.

Bateman, R., Schindler, S. E., Bollinger, J. G., Ovod, V., Mawuenyega, K. G., Li, Y., … Fagan, A. M. (2019). Blood Amyloid-beta Predicts Amyloid PET Conversion. *Alzheimer's & Dementia*, *15*(7), P526. https://doi.org/10.1016/j.jalz.2019.06.4440

Beach, T. G., Monsell, S. E., Phillips, L. E., & Kukull, W. (2012). Accuracy of the Clinical Diagnosis of Alzheimer Disease at National Institute on Aging Alzheimer's Disease Centers, 2005– 2010. *J Neuropathol Exp Neurol.*, *71*(4), 266–273. https://doi.org/10.1097/NEN.0b013e31824b211b

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology, 51*(6), 585-594.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., ... & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, *1*, 15030.

Berisha, V., Wang, S., LaCross, A., & Liss, J. (2015). Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of Presidents Ronald Reagan and George Herbert Walker Bush. *Journal of Alzheimer's Disease: JAD*, *45*(3), 959–963.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc. http://nltk.org/book

Bird, H., Lambon Ralph, M. A., Patterson, K., & Hodges, J. R. (2000). The Rise and Fall of Frequency and Imageability: Noun and Verb Production in Semantic Dementia. *Brain and Language*, *73*(1), 17–49. https://doi.org/10.1006/brln.2000.2293

Bondi, M. W., Edmonds, E. C., Jak, A. J., Clark, L. R., Delano-Wood, L., Mcdonald, C. R., … Salmon, D. P. (2014). Neuropsychological Criteria for Mild Cognitive Impairment Improves Diagnostic Precision, Biomarker Associations, and Progression Rates for the Alzheimer's Disease Neuroimaging Initiative 1. *J Alzheimers Dis January*, *1*(421), 275–289. https://doi.org/10.3233/JAD-140276

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, *40*(2), 540–545. https://doi.org/10.3758/BRM.40.2.540

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics*, *30*(7), 489–518. https://doi.org/10.3109/02699206.2016.1145740

Brysbaert, M., & New, B. (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. https://doi.org/10.3758/BRM.41.4.977

Chand, V., Baynes, K., Bonnici, L. M., & Farias, S. T. (2012). A rubric for extracting idea density from oral language samples. *Current Protocols in Neuroscience*, *58*(1), 1–15. https://doi.org/10.1002/0471142301.ns1005s58

Chandler, C., Foltz, P. W., Cheng, J., Bernstein, J. C., Rosenfeld, E. P., Cohen, A. S., … Elvevåg, B. (2019). Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology* (pp. 137–147). https://doi.org/10.18653/v1/w19-3016

Chandler, C., Foltz, P. W., & Elvevåg, B. (2019). Using Machine Learning in Psychiatry : The Need to Establish a Framework That Nurtures Trustworthiness, *46*(1), 1–4. https://doi.org/10.1093/schbul/sbz105

Cohen, T., & Widdows, D. (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* (pp. 465-475).

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Crystal, D. (1987). How many words? *English Today, 3*(4), 11-14.

Dennis, S. (2007). How to use the LSA website. In T.K. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (eds.), *Handbook of latent semantic analysis* (pp. 57–70). Mahwah, NJ: Erlbaum.

Department of Health (2012). *Prime Minister's Challenge on Dementia - Delivering major improvements in dementia care and research by 2015.* https://www.gov.uk/government/publications/prime-ministers-challenge-on-dementia

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv Preprint ArXiv:1810.04805.*

Dunn, J. C., Almeida, O. P., Barclay, L., Waterreus, A., & Flicker, L. (2002). Latent Semantic Analysis: A New Method to Measure Prose Recall. *Journal of Clinical and Experimental Neuropsychology, 24*(1), 26–35. https://doi.org/10.1076/jcen.24.1.26.965

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, *93*(1-3), 304-316.

Engelman, M., Agree, E. M., Meoni, L. A., & Klag, M. J. (2010). Propositional density and cognitive function in later life: Findings from the precursors study. *Journals of Gerontology Series B:*

Psychological Sciences and Social Sciences, *65*(6), 706–711. https://doi.org/10.1093/geronb/gbq064

Ernecoff, N. C., Wessell, K. L., Gabriel, S., Carey, T. S., & Hanson, L. C. (2018). A Novel Screening Method to Identify Late-Stage Dementia Patients for Palliative Care Research and Practice. *Journal of Pain and Symptom Management, 55*(4), 1152-1158.

Errattahi, R., Hannani, A. El, & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, *128,* pp. 32–37. https://doi.org/10.1016/j.procs.2018.03.005

Faber-Langendoen, K., Morris, J. C., Knesevich, J. W., LaBarge, E., Miller, J. P., & Berg, L. (1988). Aphasia in senile dementia of the Alzheimer type. *Annals of Neurology*, *23*(4), 365–70. https://doi.org/10.1002/ana.410230409

Firth, J. R. "A synopsis of linguistic theory, 1930-1955". In J. R. Firth et al. (eds.) *Studies in Linguistic Analysis.* Oxford: Blackwell, 1957.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes, 25*(2-3), 285-307.

Forbes-McKay, K., Shanks, M. F., & Venneri, A. (2013). Profiling spontaneous speech decline in Alzheimer's disease: a longitudinal study. *Acta Neuropsychiatrica*, *25*(6), 320–7. https://doi.org/10.1017/neu.2013.16

Frankenberg, C., Weiner, J., Schultz, T., Knebel, M., Degen, C., Wahl, H. W., & Schroeder, J. (2019). Perplexity - A new predictor of cognitive changes in spoken language? - Results of the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE). *Linguistics Vanguard*, *5*(s2). https://doi.org/10.1515/lingvan-2018-0026

Fraser, K. C., Linz, N., Lindsay, H., & König, A. (2019). The importance of sharing patient-generated clinical speech and language data, In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology* (pp. 55-61). https://doi.org/10.18653/v1/w19-3007

Fraser, K. C., Lundholm Fors, K., Eckerström, M., Öhman, F., & Kokkinakis, D. (2019). Predicting MCI Status From Multimodal Language Data Using Cascaded Classifiers. *Frontiers in Aging*

Neuroscience, *11*, 205. https://doi.org/10.3389/fnagi.2019.00205

Fraser, K. C., Lundholm Fors, K., & Kokkinakis, D. (2019). Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech and Language*, *53*, 121–139. https://doi.org/10.1016/j.csl.2018.07.005

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, *49*, 407–422. https://doi.org/10.3233/JAD-150520

Garrard, P., Maloney, L. M., Hodges, J. R., & Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, *128*(2), 250–260. https://doi.org/10.1093/brain/awh341

Garrard, P. (2009). Cognitive archaeology: Uses, methods, and results. *Journal of Neurolinguistics, 22*(3), 250-265.

Glosser, G., & Deser, T. (1991). Patterns of discourse production among neurological patients with fluent language disorders. *Brain and Language*, *40*(1), 67–88. https://doi.org/10.1016/0093-934X(91)90117-J

Glosser, G., & Deser, T. (1992). A comparison of changes in macrolinguistic and microlinguistic aspects of discourse production in normal aging. *Journal of Gerontology*, *47*(4), 266-272.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers, 36*(2), 193–202.

Gupta, T., Hespos, S. J., Horton, W. S., & Mittal, V. A. (2018). Automated analysis of written narratives reveals abnormalities in referential cohesion in youth at ultra high risk for psychosis. *Schizophrenia Research, 192,* 82–88. https://doi.org/10.1016/j.schres.2017.04.025

Halpern, R., Seare, J., Tong, J., Hartry, A., Olaoye, A., & Aigbogun, M. S. (2019). Using electronic health records to estimate the prevalence of agitation in Alzheimer disease/dementia. *International journal of geriatric psychiatry*, *34*(3), 420-431.

Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. https://doi.org/10.1021/ci0342472

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

10, 260–268. https://doi.org/10.1016/j.dadm.2018.02.004

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266.

Hoffman, P. (2019). Reductions in prefrontal activation predict off-topic utterances during speech production. *Nature communications*, *10* (1), 1-11.

Holshausen, K., Harvey, P. D., Elvevåg, B., Foltz, P. W., & Bowie, C. R. (2014). Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. *Cortex*, *55*(1), 88–96. https://doi.org/10.1016/j.cortex.2013.02.006

Horton, W. S., Spieler, D. H., & Shriberg, E. (2010). A Corpus Analysis of Patterns of Age-Related Change in Conversational Speech. *Psychol Aging*, *25*(3), 708–713. https://doi.org/10.1038/jid.2014.371

Iter, D., Yoon, J., & Jurafsky, D. (2018). Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 136–146). https://doi.org/10.18653/v1/w18-0615

Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., … Trojanowski, J. Q. (2013). Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, *12*(2), 207-216. https://doi.org/10.1016/S1474-4422(12)70291-0

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., & Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, (pp. 27–37). https://doi.org/10.3115/v1/W14-3204

Johnson, S. C., Koscik, R. L., Jonaitis, E. M., Clark, L. R., Mueller, K. D., Berman, S. E., … Sager, M. A. (2018). The Wisconsin Registry for Alzheimer's Prevention: A review of findings and current directions. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *10*, 130–142. https://doi.org/10.1016/j.dadm.2017.11.007

Jurafsky, D., & Martin, H. Speech and Language Processing (Draft). 2017. *URL: https://web.stanford.edu/~jurafsky/slp3*.

Kilgarriff, A. (2006). BNC database and word frequency lists. Retrieved May 13, 2013, from http://www.kilgarriff.co.uk/bnc-readme.html

Krithara, A., Aisopos, F., Rentoumi, V., Nentidis, A., Bougatiotis, K., Vidal, M. E., ... & Torrente, M. (2019). iASiS: Towards Heterogeneous Big Data Analysis for Personalized Medicine. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 106-111). IEEE.

Kwon, S., Kim, S. J., & Choeh, J. Y. (2016). Preprocessing for elderly speech recognition of smart devices. *Computer Speech & Language, 36,* 110-121.

Kyle, K. (n.d.). TAALED. Retrieved January 28, 2020, from https://www.linguisticanalysistools.org/taaled.html

Kyle, K. (2016). Measuring Syntactic Development in L2 Writing : Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Retrieved from https://scholarworks.gsu.edu/alesl_diss/35

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Doctoral Dissertation). Retrieved from http://scholarworks.gsu.edu/alesl_diss/35.

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly, 49*(4), 757–786. https://doi.org/10.1002/tesq.194

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes, 25(2-3),* 259-284.

Landin-Romero, R., Tan, R., Hodges, J. R., & Kumfor, F. (2016). An update on semantic dementia: genetics, imaging, and pathology. *Alzheimer's Research & Therapy*, *8*(1), 52. https://doi.org/10.1186/s13195-016-0219-5

Laske, C., Sohrabi, H. R., Frost, S. M., López-De-Ipiña, K., Garrard, P., Buscema, M., … O'bryant, S. E. (2015). Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimer's and Dementia, 11(*5*), 561-578*. https://doi.org/10.1016/j.jalz.2014.06.004

Le, X., Lancashire, I., Hirst, G., & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and Linguistic Computing*, *26*(4), 435–461. https://doi.org/10.1093/llc/fqr013

Lehr, M., Prud, E., Shafran, I., & Roark, B. (2012). Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Libon, D. J., Rascovsky, K., Powers, J., Irwin, D. J., Boller, A., Weinberg, D., … Grossman, M. (2013). Comparative semantic profiles in semantic dementia and Alzheimer's disease. *Brain*, *136*(8), 2497–2509. https://doi.org/10.1093/brain/awt165

López-de-Ipiña, K., Alonso, J.-B., Travieso, C., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., … Lizardui, U. (2013). On the Selection of Non-Invasive Methods Based on Speech Analysis Oriented to Automatic Alzheimer Disease Diagnosis. *Sensors*, *13*(5), 6730–6745. https://doi.org/10.3390/s130506730

Lovestone, S. (2014). Blood biomarkers for Alzheimer's disease. *Genome Medicine*, *6*(8), 8–11. https://doi.org/10.1186/s13073-014-0065-7

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96(2),* 190-208.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Manning, C. D. (2011, February). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics* (pp. 171-189). Springer, Berlin, Heidelberg.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: the penn treebank. *Computational Linguistics*, *19*(2), 313-330.

Marini, A., Boewe, A., Caltagirone, C., & Carlomagno, S. (2005). Age-related differences in the production of textual descriptions. *Journal of psycholinguistic research*, *34*(5), 439-463.

Masrani, V., Murray, G., Field, T., & Carenini, G. (2017). Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In *BioNLP 2017* (pp. 232-237).

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes*, *47*(4), 292–330. https://doi.org/10.1080/01638530902959943

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, *arXiv preprint arXiv:1301.3781*. https://doi.org/10.1162/153244303322533223

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., & Christensen, H. (2019). Dementia detection using automatic analysis of conversations. *Computer Speech and Language*, *53*, 65–79. https://doi.org/10.1016/j.csl.2018.07.006

Mirheidari, B., Blackburn, D., Walker, T., Venneri, A., Reuber, M., & Christensen, H. (2018). Detecting signs of dementia using word vector representations. In *Interspeech* (pp. 1893-1897). https://doi.org/10.21437/Interspeech.2018-1764

Mueller, K. D., Koscik, R. L., Clark, L. R., Hermann, B. P., Johnson, S. C., & Turkstra, L. S. (2018). The Latent Structure and Test-Retest Stability of Connected Language Measures in the Wisconsin Registry for Alzheimer's Prevention (WRAP). *Archives of Clinical Neuropsychology, 33*(8), 993–1005. https://doi.org/10.1093/arclin/acx116

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. https://doi.org/10.1186/s40537-014-0007-7

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., … Chertkow, H. (2005). the Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1029/WR017i002p00410

Orimaye, S. O., Wong, J. S. M., & Golden, K. J. (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* (pp. 78-87).

Orimaye, S. O., Wong, J. S. M., & Wong, C. P. (2018). Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *PLoS ONE*, *13*(11), 1–15. https://doi.org/10.1371/journal.pone.0205636

Oulhaj, A., Wilcock, G. K., Smith, A. D., & De Jager, C. A. (2009). Predicting the time of conversion to MCI in the elderly: Role of verbal expression and learning. *Neurology*, *73*(18), 1436–1442. https://doi.org/10.1212/WNL.0b013e3181c0665f

Pakhomov, S., Chacon, D., Wicklund, M., & Gundel, J. (2011). Computerized assessment of syntactic

complexity in Alzheimer's disease: A case study of Iris Murdoch's writing. *Behavior Research Methods*, *43*(1), 136–144. https://doi.org/10.3758/s13428-010-0037-9

Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M. S., & Braga, D. (2012). Impact of age in ASR for the elderly: Preliminary experiments in European Portuguese. In *Advances in Speech and Language Technologies for Iberian Languages* (pp. 139-147). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35292-8_15

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. https://doi.org/10.15781/T29G6Z

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *ArXiv Preprint ArXiv:1802.05365.* https://doi.org/10.18653/v1/n18-1202

Prince, M., Bryce, R. & Ferri, C. (2011). *World Alzheimer Report 2011: The benefits of early diagnosis and intervention.* Alzheimer's Disease International. https://www.alz.co.uk/research/world-report-2011.

Prud'hommeaux, E., van Santen, J., & Gliner, D. (2017). Vector space models for evaluating semantic fluency in autism. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 32–37). https://doi.org/10.18653/v1/P17-2006

Raschka, S. (2015). *Python machine learning.* Packt Publishing Ltd.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.*

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophrenia*, *5*(1), 1-12

Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C. A., & Garrard, P. (2014). Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's*

Disease, 42(s3), S3-S17.

Ripich, D. N., Carpenter, B. D., & Ziol, E. W. (2000). Conversational cohesion patterns in men and women with Alzheimer's disease: a longitudinal study. *International Journal of Language & Communication Disorders*, *35*(1), 49–64.

Ritchie, K. (2004). Mild cognitive impairment: An epidemiological perspective. *Dialogues in Clinical Neuroscience*, *6*(4), 401–408.

Ritchie, K., Carrière, I., Su, L., O'Brien, J. T., Lovestone, S., Wells, K., & Ritchie, C. W. (2017). The midlife cognitive profiles of adults at high risk of late-onset Alzheimer's disease: The PREVENT study. *Alzheimer's and Dementia*, *13*(10), 1089–1097. https://doi.org/10.1016/j.jalz.2017.02.008

Roark, B., Mitchell, M., Hosom, J. P., Hollingshead, K., & Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(7), 2081–2090. https://doi.org/10.1109/TASL.2011.2112351

Rosenstein, M., Foltz, P. W., DeLisi, L. E., & Elvevåg, B. (2015). Language as a biomarker in those at high-risk for psychosis. *Schizophrenia Research, 165*(2-3), 249–250. https://doi.org/10.1016/j.schres.2015.04.023

Salvatore, C., & Castiglioni, I. (2018). A wrapped multi-label classifier for the automatic diagnosis and prognosis of Alzheimer's disease. *Journal of Neuroscience Methods*, *302*, 58–65. https://doi.org/10.1016/j.jneumeth.2017.12.016

Schaffer, C. (1993). Technical Note: Selecting a Classification Method by Cross-Validation. *Machine Learning*, *13*(1), 135–143. https://doi.org/10.1023/A:1022639714137

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal*, *30*(1), 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Snowdon, D. A. (1997). Aging and Alzheimer's Disease : Lessons From the Nun Study. *The Gerontologist, 37*(2), 150-156.

Spencer, E., Craig, H., Ferguson, A., & Colyvas, K. (2012). Language and ageing - Exploring propositional density in written language - Stability over time. *Clinical Linguistics and Phonetics*, *26*(9), 743–754. https://doi.org/10.3109/02699206.2012.673046

Tao, X., Zhou, X., Zhang, J., & Yong, J. (2016). Sentiment Analysis for Depression Detection on Social Networks. In *International Conference on Advanced Data Mining and Applications* (pp. 807-810). Springer, Cham.

Toledo, C. M., Aluísio, S. M., dos Santos, L. B., Brucki, S. M. D., Trés, E. S., de Oliveira, M. O., & Mansur, L. L. (2018). Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer's disease, mild cognitive impairment, and no cognitive impairment. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *10*, 31–40. https://doi.org/10.1016/j.dadm.2017.08.005

Tsantali, E., Economidis, D., & Tsolaki, M. (2013). Could language deficits really differentiate Mild Cognitive Impairment (MCI) from mild Alzheimer's disease? *Archives of Gerontology and Geriatrics*, *57*(3), 263–270. https://doi.org/10.1016/j.archger.2013.03.011

Van Velzen, M., & Garrard, P. (2008). From hindsight to insight–retrospective analysis of language written by a renowned Alzheimer's patient. *Interdisciplinary Science Reviews*, *33*(4), 278-286.

Vipperla, R., Renals, S., & Frankel, J. (2008). Longitudinal study of ASR performance on ageing voices. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2550–2553.

Vipperla, R., Renals, S., & Frankel, J. (2010). Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing, 2010*(1), 525783. https://doi.org/10.1155/2010/525783

Wang, X., Zhang, C., Ji, Y., Sun, L., & Wu, L. (2013). A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 201-213). Springer, Berlin, Heidelberg.

Wankerl, S., Nöth, E., & Evert, S. (2016). An Analysis of Perplexity to Reveal the Effects of Alzheimer's Disease on Language. In *Speech Communication; 12. ITG Symposium* (pp. 1-5). VDE.

Weissenbacher, D., Travis, J. A., Wojtulewicz, L., Amylou, D., Locke, D., Caselli, R., & Gonzalez, G. (2016). Towards Automatic Detection of Abnormal Cognitive Decline and Dementia Through Linguistic Analysis of Writing Samples. In *Proceedings of NAACL-HLT* (pp. 1198-1207).

Yancheva, M., Fraser, K., & Rudzicz, F. (2015). Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies* (pp. 134-139). https://doi.org/10.1016/j.bandl.2008.08.002

Yancheva, M., & Rudzicz, F. (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) (pp. 2337-2346).

Zhou, L., Fraser, K. C., & Rudzicz, F. (2016). Speech recognition in Alzheimer's disease and in its assessment. In *Interspeech* (pp. 1948-1952). https://doi.org/10.21437/Interspeech.2016-1228

Zhou, X., Wang, Y., Sohn, S., Therneau, T. M., Liu, H., & Knopman, D. S. (2019). Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing. *International journal of medical informatics, 130,* 103943.