APHASIOLOGY, 2018
VOL. 32, NO. 1, 41–60
https://doi.org/10.1080/02687038.2017.1355439

Routledge
Taylor & Francis Group

Rhetorical structure and Alzheimer's disease

Mohamed Abdalla[a], Frank Rudzicz[a,b] and Graeme Hirst[a]

[a]Department of Computer Science, University of Toronto, Toronto, Canada; [b]Toronto Rehabilitation Institute-UHN, Toronto, Canada

ABSTRACT

Background: Language is one of the first faculties afflicted by Alzheimer's disease (AD). A growing body of work has focussed on leveraging automated analysis of speech to accurately predict the onset of AD. Previous work, however, did not address the effects of AD on the structure of discourse in spontaneous speech and literature.

Aims: Our goal is to identify the effects of AD on the structure of discourse, both in spontaneous speech and in literature.

Methods & Procedures: We use two existing data sets, DementiaBank and the Carolina Conversations Collection, to explore how AD manifests itself in spontaneous speech. This is done by automatically extracting discourse relations according to Rhetorical Structure Theory. We also study written novels, comparing authors with and without dementia using the same tools.

Outcomes & Results: Several discourse relations, especially those involving elaboration and attribution, are significant indicators of AD in speech. Indicators of the disease in written text, by contrast, involve relations of logical contingency.

Conclusions: Our work highlights how AD can alter discourse structures in both spontaneous speech and written text. Future work should combine discourse analysis with previously studied lexico-syntactic features.

ARTICLE HISTORY
Received 11 August 2016
Accepted 11 July 2017

KEYWORDS
Alzheimer's disease; Rhetorical Structure Theory; computational linguistics; spontaneous speech; longitudinal analysis

Introduction

Memory impairment is the main symptom of typical (late-onset) Alzheimer's disease (AD), without initial posterior cortical atrophy. However, language is one of the first faculties afflicted by AD, with changes presenting a year or more before diagnosis (Ahmed, de Jager, Haigh, & Garrard, 2013). In fact, low idea density and low grammatical complexity in early life can presage declining cognitive test scores decades later (Snowdon et al., 1996). In response to this phenomenon, a growing body of work has used automated linguistic analysis to differentiate individuals with AD, or other dementias, from people without the disease (Almor, Kempler, MacDonald, Andersen, & Tyler, 1999; Szatloczki, Hoffmann, Vincze, Kalman, & Pakaski, 2015; Fraser, Rudzicz, & Rochon, 2013; Fraser, Meltzer, & Rudzicz, 2015). Automated analysis of this type has important implications for clinical assessment, and facilitates the use of relatively large data sets.

CONTACT Frank Rudzicz ✉ frank@cs.toronto.edu Department of Computer Science, University of Toronto, Toronto, Canada.

In the current study, we demonstrate how AD affects the structure of discourse in spontaneous speech and in literature, in contrast with control subjects without the disease (CT). In particular, we apply Rhetorical Structure Theory (RST) (Mann & Thompson, 1988), which is a popular descriptive linguistic framework, to a range of phenomena in the organization of natural discourse, especially in terms of pragmatic relations between segments of text (Taboada & Mann, 2006). Here, we apply this framework to both speech transcripts (of a picture-description task and of free conversation) and novels written by several authors.

Language in AD

Faber-Langendoen et al. (1988) showed that the prevalence of "aphasia" (as assessed by aphasia battery scores) increases with the severity of dementia – in their study, 36% of people with mild AD and 100% of those with severe AD were found to have aphasia. This was manifested mainly in diminished comprehension and written expression. Often, linguistic changes due to dementia correlate significantly with decreased naming ability (Kirshner, Webb, & Kelly, 1984; Reilly, Troche, Grossman, & Budson, 2011; Taler & Phillips, 2008), but also with articulation, word-finding, semantic topic structure (Yancheva & Rudzicz, 2016), and semantic fluency generally (Weiner, Neubecker, Bret, & Hynan, 2008). There is also evidence that AD can be detected through increased incidence of phonological errors (Forbes-McKay, Shanks, & Venneri, 2013), and in the acoustics of emotional speech, through modern signal processing techniques (Bhaduri, Das, & Ghosh, 2016).

Low-level lexical and phonological characteristics have sometimes been used to analyze higher-level *discourse* processing, including connecting high-level semantic themes. Ahmed et al. (2013), for example, approached discourse analysis by measuring lexico-syntactic features in patients with autopsy-confirmed AD, showing that the total number of semantic units in connected speech, including subject-, location-, and object-related nouns correlated with the "Expression" subscore of the Cambridge Cognitive Examination, which includes object and picture naming, and category fluency (Roth, Tym, & Mountjoy, 1986). Similarly, the pragmatic "emptiness" of discourse in AD has often been described in lower-level terms, such as in increased incoherent phrases, and in semantic and graphemic paraphasia (Szatloczki et al., 2015) or in a lack of sensitivity to pronoun appropriateness (Kempler & Goral, 2008). Prosodic cues (Mu¨llerr & Guendouzi, 2002), words per turn, intelligibility (Ripich, Vertes, Whitehouse, Fulton, & Ekelman, 1991), and propositional complexity (Wilson, Rochon, Mihailidis, & Leonard, 2012) have also been used as surrogates for more-abstract discourse features in AD.

However, Glosser and Deser (1991) suggested that AD impairs thematic coherence to a greater extent than microlinguistic syntactic and phonological processes can capture. This can manifest in generally reduced "discourse structuring ability" (Hutchinson & Jensen, 1980), and an inability to follow normal "discourse rules and conventions" (Almor et al., 1999). Similarly, Chapman et al. (2002) showed that AD and mild cognitive impairment result in drastically reduced gist-level aspects of discourse, including inferences across stories and across adjacent sentences, and an inability to identify overarching story themes or even the main idea of a discourse. Seixas Lima et al. (2016) also showed that patients with the semantic variant of primary progressive aphasia produced significantly fewer coherent semantic details than controls in a discourse analysis.

Quantitative analysis of discourse in these high-level terms has been elusive, perhaps due to a lack of procedural methodologies of their computation, which is a gap the present research intends to fill. In this paper, we briefly discuss RST and recent computational "Big Data" approaches to distinguishing people with AD from those without, using linguistic measures. We then demonstrate using RST as a means for that clinical comparison on spontaneous speech from two databases.

Rhetorical Structure Theory

In this paper, *discourse* refers to the coherent composition of a series of textual or spoken linguistic units that can be linked together, both within and across sentence boundaries, into a hierarchical, logical structure. Here, and in related work, *sentences* are defined as a sequence of word tokens delimited by terminal punctuation. As indicated later, the basic units of analysis are *not* sentences, but are typically sub-sentential.

RST (Mann & Thompson, 1988) is a descriptive theoretical framework for discourse analysis. In RST, coherent language is structured by *discourse trees* in which leaf nodes cover nonoverlapping text spans called *elementary discourse units* (EDUs), which are typically sub-sentential clauses, and sometimes phrasal. Adjacent nodes are related through particular discourse relations that structure a text semantically and, importantly, pragmatically; Table 1 defines and exemplifies each of these relations, using the coarse-grained classification given by Carlson and Marcu (2001). Each use of these relations forms new nodes that can be combined with adjacent nodes recursively, so that any complete span of nonoverlapping text or speech forms a discourse *subtree*, as exemplified in Figure 1.

Carlson and Marcu (2001) describe several lexical and phrasal cues that can be used when manually segmenting language; these cues include pseudo-clefts, temporal expressions (e.g., *after*), adverbials, and correlative subordinators. Some of these cues are used as input to the machine learning approach to discourse segmentation, as described in the "Methodology" section.

To represent a discourse structure in the RST framework, two steps are necessary, in sequence: (1) *discourse segmentation* in which the source text or speech is partitioned into nonoverlapping EDUs, and (2) *discourse parsing* in which EDU segments are combined into subtrees. The following examples show sentences with two clausal EDUs (one superordinate, and the other subordinate, with a discourse marker in bold), demarcated by square brackets (Carlson & Marcu, 2001).

Ex.1: [Such trappings suggest a glorious past] [**but** give no hint of a troubled present.]

(Contrast)

Ex.2: [**Although** Mr. Freeman is retiring,] [he will continue to work as a consultant for American Express on a project basis.] (Contrast)

Ex.3: [Previously, airlines were limiting the programs] [**because** they were becoming too expensive.] (Cause)

In RST, there are two types of discourse relation: *hypotactic* ("mononuclear") and *paratactic* ("multinuclear"). In mononuclear relations, the *nucleus* is a text span that is more salient than the other – the *satellite*; in multinuclear relations, all text spans are

Table 1. The coarse RST relations developed by Mann and Thompson (1988) and defined by Carlson and Marcu (2001).

Relations
Attribution Instances of reported speech, direct or indirect, positive or negative.
Example: [And the girl is saying] ["Be quiet"]
Background The satellite establishes the context or the grounds with respect to which the nucleus is to be interpreted.
Example: [The mother is at the sink …] [She's washing and drying dishes.]
Cause The situation presented in the nucleus is the cause of that in the satellite.
Example: [… she's gonna wipe up some water] [because it's sure running out!]
Comparison Two textual spans are compared along some dimension, which can be abstract, including preference, analogy, and proportion.
Example: [it would sound better][than going the other way]]…
Condition The truth of the proposition associated with the nucleus is a consequence of the fulfillment of the condition in the satellite. The satellite presents a situation that is not realized. This can be hypothetical, a contingency, or otherwise.
Example: [it would be more fun][if you had some variety]
Contrast Two or more nuclei come in contrast with each other along some dimension, including concession and antithesis. Typically, the Contrast relation includes a contrastive discourse cue, such as *but, however*, whereas Comparison does not.
Example: [I 'm laughing about it now][but it was really very serious at the time]
Elaboration The satellite gives additional information or detail about the situation presented in the nucleus. This includes general-specific, part-whole, set-member, object-attribute, and process-step relations.
Example: [a pathway][[that has *um* a tree and shrubbery] [and a part of what might be an extension of the house or a garage]]
Enablement The situation presented in the nucleus is unrealized. The action presented in the satellite increases the chances of the situation in the nucleus being realized.
Example: [telling him to be quiet] [so that mother will not hear]
Evaluation One span assesses the situation in the other span on a qualitative scale. An evaluation can be an appraisal, rating, interpretation, or assessment.
Example: [it's a big town][so it's well chosen]
Explanation The satellite provides a factual, evidential, or purposeful explanation for the nucleus.
Example: [he's about to fall][because the *uh* step-stool is tilting]
Joint A multinuclear relation whose elements can be listed, but which are not in a comparison, contrast or other, stronger type of multinuclear relation.
Example: [the cupboard door is *uh* open][and he was … after the cookies]
Manner-Means A manner satellite explains the way in which something is done. A means satellite specifies a method, mechanism, or conduit for accomplishing some goal.
Example: [I cut off over almost two hours][by going up River Street]
Summary The satellite summarizes the information presented in the nucleus.
Example: [The water is coming out of the sink,][the water from the sink is getting all over the floor.]
Temporal The situation presented in the nucleus (often realized as a superordinate clause) occurs before, after, or at the same time as the situation in the satellite.
Example: [and *uh* looks like he might make it][before he hits the floor]
Topic-Change This is used to link large textual spans when there is a sharp or gradual change in focus going from one segment to the other.
Topic-Comment A general statement or topic of discussion is introduced, after which a specific remark is made on the statement or topic.
Textual Organization A multinuclear relation used to link elements of the structure of the text, for example, to link a title with the body of the text, a section title with the text of a section, etc. It primarily enforces a tree structure on the representation.
Example: [Starting now?][I see a boy on a stool …]
Same-Unit A pseudo-relation used as a device for linking two discontinuous text fragments that are really a single EDU, but which are broken up by an embedded unit.
Example: [And the boy is] ***coughs*** sorry [he's reaching for the cookie jar …]

The examples are derived from the data sets used in this work.

equally salient (Feng, 2014). Nuclei and satellites are rarely determined in isolation but depend on context, even with similar semantics (Carlson & Marcu, 2001). For instance, consider the following examples:

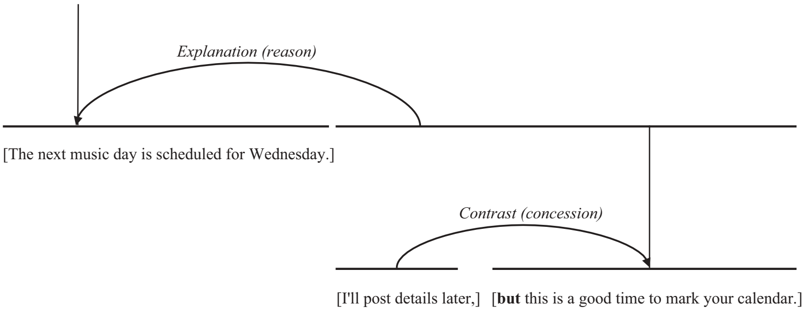

Explanation (reason)
[The next music day is scheduled for Wednesday.]
Contrast (concession)
[I'll post details later,] [**but** this is a good time to mark your calendar.]

Figure 1. RST subtree example (adapted from Mann and Thompson (1988)).

Ex.4: [The earnings were fine and above expectations.] [**Nevertheless**, Salomon's stock fell $1.125 yesterday . . .] (Contrast)

Ex.5: [**Although** the earnings were fine and above expectations,] [Salomon's stock fell $1.125 yesterday]. (Background)

Despite the semantic and lexical relatedness of these examples, both EDUs are nuclei of a multinuclear relation in Ex.4, but the first EDU is the satellite of the second in Ex.5 since, generally, satellites can be removed or substituted without altering the meaning of their nuclei. Furthermore, these relations can connect adjacent sentences or even adjacent paragraphs in a coherent text.

Manually segmenting and parsing text in RST is open to some interpretability, but extensive protocols exist for that purpose (Carlson & Marcu, 2001). Several projects have demonstrated the reliability of RST; for instance, den Ouden (2004) showed higher agreement between human judges in RST analysis than in alternative methods on the task of hierarchical discourse analysis, building on earlier work that showed high consistency between analysts using RST in complex text (den Ouden, van Wijk, Terken, & Noordman, 1998). Extensive testing of agreement in RST is reported by Marcu, Romera, and Amorrortu (1999) on three levels: assignment of text spans, assignment of nuclei, and assignment of relations, establishing the reliability of each. However, since segmentation and parsing each involve potentially ambiguous data, automated systems must take stochastic approaches in practice; those applied in the present work are described in the following subsections.

The developers of RST point out several vulnerabilities (Taboada & Mann, 2006). For instance, the approach is intimately tied to the clause structure of the given language; here, we focus only on English, in which the theory originated. Moreover, in spoken language, units are often considered to be intonational, rather than independent clauses. Fortunately, there is evidence that the cognitive representation of RST-like relations transcends both text and speech. For instance, den Ouden (2004) found that discourse aspects of read texts correspond to characteristic prosodic cues; for example, pause and pitch range correlate with the level of relation embedding, with nuclearity, and with particular relations (e.g., causal relations are associated with shorter pauses than noncausal relations, and nuclei are uttered more slowly than satellites) (Taboada & Mann, 2006). Noordman, Dassen, Swerts, and Terken (1999) found similar results.

In this work, we apply RST both to written text and spontaneous speech. We examine the rhetorical relations described by Mann and Thompson (1988), as shown in Table 1. In general, because of data sparsity (especially in conversational data), we consider coarse relations by summing together their respective, encompassed fine relations. Additional details are provided in the “Experiments: spontaneous speech” section.

Discourse treebanks

Computational approaches generally require data from which to form statistical models. The RST Discourse Treebank (RST-DT) is a corpus of 385 documents from the *Wall Street Journal*, annotated in the RST framework (Carlson, Marcu, & Okurowski, 2001). This corpus is perhaps the most widely used benchmark for research in RST-style discourse analysis, and encapsulates the definitions of EDUs and discourse relations of Mann and Thompson (1988)’s seminal work. For consistency, RST-DT is generally annotated with *clauses* as the basis of EDUs, with some exceptions (e.g., clauses that are subjects or objects of main verbs are not EDUs).

As a point of comparison, the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is a superset of the material in RST-DT, but uses Discourse Lexicalized Tree Adjoining Grammar (D-LTAG) (Webber, 2004) as the dominant framework. D-LTAG is based on predicate-argument structures, with lexically based discourse relations (e.g., the connective *because* is a predicate that takes two text spans as arguments). Unlike RST, D-LTAG neither guarantees a complete coverage of a corpus nor imposes a hierarchical structure on that corpus. The latter point makes RST a more appropriate formalism for the present work, as it inherently models the global understanding (and understandability) of a body of language (Feng & Hirst, 2014).

Related computational work

The automatic analysis of discourse has been applied to several tasks in language processing, including text summarization (Louis, Joshi, & Nenkova, 2010), natural language generation (Prasad, Joshi, Dinesh, Lee, & Miltsakaki, 2005), and question–answering (Chai & Jin, 2004). Typically, discourse provides enough additional information to improve the results according to *extrinsic* evaluation criteria in these kinds of tasks (e.g., the validity of answers to user-provided questions), but little work has evaluated computational discourse analysis *intrinsically* (i.e., in terms of the validity of the discourse models themselves).

Discourse annotations can be either positive (present) or negative (absent), hence Type I (false positive, FP) errors and Type II (false negative, FN) are possible, in addition to true positives (TP) and true negatives (TN) (Scholman, Evers-Vermeul, & Sanders, 2016). There are two primary criteria for measuring the quality of automatically produced discourse annotations, namely *Precision* and *Recall*, defined as

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN}. \tag{2}$$

Precision is the positive predictive value and measures the proportion of all positive annotations produced by the system that are in fact true positives. *Recall* is the true positive rate, or the proportion of all possible positive annotations that are actually made. Since both criteria are important, their harmonic mean, F_1, is sometimes used:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{3}$$

For discourse tree-building, there are two commonly applied strategies:

(1) *Greedy strategies* consider only one possible solution at each parsing step. By optimizing the decision at each intersection, these strategies identify the most appropriate pairs of adjacent EDUs to be merged into larger discourse units. However, as with any greedy approach, these strategies do not reconsider any prior decisions, and thus the final tree may be nonoptimal. The HILDA discourse parser is a commonly used greedy tree-building strategy (Duverle & Prendinger, 2009; Hernault, Bollegala, & Ishizuka, 2010). At each step of this bottom-up approach,[1] a *structural* classifier is used to determine whether two adjacent EDUs or subtrees should be combined, and a *labelling* classifier chooses which relation should be assigned to that new subtree.

(2) *Non-greedy strategies* consider multiple (possibly all) solutions at each step, and choose the optimal one according to some stochastic function. For example, Joty, Carenini, and Ng (2012) apply conditional random fields (CRFs, described in the "Methodology" section) for segmentation, then build the discourse tree from the bottom-up, following a parsing algorithm similar to the Cocke–Younger–Kasami method, which is a probabilistic parsing algorithm for context-free grammars (Lee, 2002).

Recently, the CRF parsing approach of Joty, Carenini, Ng, and Mehdad (2013) outperformed HILDA in terms of precision and recall on annotations in RST-DT. The apparent success of the non-greedy approach may be attributed to the fact that it incorporates contextual information by using CRFs as local classifiers. This accuracy comes at a computational cost, with a worst-case complexity of $\mathcal{O}(n^3)$ (the time required increases polynomially with the size of the input), in contrast with HILDA's much faster $\mathcal{O}(n)$ (the time required increases only linearly with the size of the input), given n EDUs.

Despite evidence of structural changes to discourse in AD, no prior computational work has attempted to capture these changes in a statistical model. Instead, work has focussed on more accessible features of language. For example, Fraser et al. (2015) obtained state-of-the-art accuracy in identifying AD from short narratives during a picture-description task using 370 acoustic, lexical, syntactic, and semantic features. Some of these features have been used in low-level discourse processing, including part-of-speech tags (e.g., plural noun, gerund verb), the lengths of utterances in words, and paraphasias.

Orimaye and Golden (2014) similarly avoid full discourse analysis by limiting their exploration to syntactic dependency parses. In contrast, the current work applies the RST framework to the speech and language of individuals with AD to quantitatively assess differences in their discourse structure. In particular, we demonstrate that the cognitive decline associated with AD affects the frequencies of production of specific discourse relations.

Methodology

We use a fast, linear-time bottom-up discourse parser to extract RST relations, as described later. This involves segmenting the original text, building trees from those segments, and subsequently selecting the most relevant features.

Discourse segmentation

The first step is to identify atomic, nonoverlapping regions of text that can be later combined into discourse trees. This is a stochastic process that can affect the correctness of tree-building, so its accuracy is essential. There have traditionally been two approaches to segmenting a text: the first is to consider each word token in the sentence independently, and to use a binary classifier such as a support vector machine (Cortes & Vapnik, 1995) or logistic regressor to decide whether a new EDU begins at that token (Fisher & Roark, 2007). The second approach is to use sequential labelling by considering the sentence as a whole and assigning a label to each token, using local context. Both Hernault et al. (2010) and Feng and Hirst (2014) showed that this latter approach is more effective, using a statistical method called CRFs.

CRFs are dynamical models that connect sequences of two variables: token observations τ_i and label sequences l_i, where $i = 0, \ldots, t$ for sequences of length $t + 1$ tokens. In our case, observations are individual word tokens and each label is either *B* if the token begins an EDU or *C* otherwise. In this implementation, the conditional probabilities associated with a label l_i depend on the previous (l_{i-1}) and subsequent (l_{i+1}) labels, and the previous (τ_{i-1}) and current (τ_i) tokens. The first token in a text, τ_0 is always considered to begin an EDU; hence l_0 is always **B**. Figure 2 shows the labelling for a sequence used in our experiments by the CRF graphical model used in this work. The parameters of a CRF allow it to be "unrolled" to sequences of arbitrary length. Those parameters are optimized according to training data in order to minimize error; in our case, this was provided by the standard RST-DT data set described earlier (Feng, 2014).

Here, we use a two-pass segmentation algorithm which first uses adjacent tokens to infer labels using contextual information, producing an initial label sequence, and then applies global features to refine that initial sequence. Basic features (used in both passes) include part-of-speech tags for each token, neighbouring punctuation, and the

Text: [well the girl is telling the boy] [**to** get the cookies down] [**but** don't tell your mother .]
Labels: **B** C C C C C C **B** C C C C **B** C C C C C

Figure 2. A sample label sequence from a conditional random field (CRF), indicating EDU boundaries, given a data sequence (DementiaBank 007–1) from our experiments.

depth of the largest syntactic constituent starting from or ending with the current token, as determined by the reranking parser of Charniak and Johnson (2005). Global features (used in the second pass) include lemmas to the left and right, the distance to the nearest marked EDU boundary, and the number of syntactic constituents formed by the sequence between the current token and the nearest marked EDU boundary. By using the RST-DT data set to train the parameters of such a CRF, Feng and Hirst (2014) obtained state-of-the-art precision of 96.1%, recall of 95.9%, and a combined F_1 score of 96.0% over both label types on held-out data. We use this system in the experiments later.

Bottom-up discourse parsing

After EDUs are segmented, they are combined using a bottom-up tree-building parser to form a discourse tree over the text or transcript with EDUs as leaf nodes. This tree is then modified using higher-level information and subsequently combined with other sentence-level discourse trees to form the final structure.

As with segmentation, there are two classes of discourse parsers: greedy and non-greedy, as discussed in the section on related computational work. Here, we employ the approach of Feng and Hirst (2014), which combines the greedy bottom-up tree-building process of HILDA with two non-greedy linear-chain CRFs in cascade to serve as local classifiers to select relations. Adding contextual information to HILDA unites the strengths of both approaches into a more optimal parser, both in terms of efficiency and discourse parsing accuracy.

Given a pair of text spans S_L and S_R, the rich linguistic features used to drive parsing include: (1) *n*-gram prefixes and suffixes, (2) lexical heads, (3) syntactic tag prefixes and suffixes, (4) word pairs across S_L and S_R, (5) dependency parse features (after Lin, Kan, and Ng (2009)), (6) semantic similarity of verbs in VerbNet[2] and WordNet, and (7) cue phrases from Knott (1994).

This discourse parser achieved an accuracy of 95.6% and an F_1 score of 89.5% in structural reconstruction in RST-DT (Feng & Hirst, 2014); we use this parser in the following experiments.

Feature analysis

In the following analysis, we use a one-way ANOVA to identify the most informative RST relations for distinguishing people with AD and those without, CT. Specifically, ANOVA is run on each relation independently, and the grouping variable is the presence or absence of AD, which is provided in each data set. Equation 4 shows the F statistic used to evaluate RST features. Here, $\bar{Y}_i$ is the sample mean of the ith group, n_i is the number of observations in that group, $\bar{Y}$ is the overall mean in the data, K is the number of groups, Y_{ij} is the jth observation of the ith group, and N is the overall sample size. In the present work, we correct for multiple comparisons where appropriate.

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{(N-K)\cdot\sum_i n_i(\bar{Y}_i-\bar{Y})^2}{(K-1)\sum_{ij}\left(Y_{ij}-\bar{Y}_i\right)^2} \tag{4}$$

Experiments: spontaneous speech

Here, we examine how AD manifests itself in connected, spontaneous speech, with regards to discourse unit creation and organization, across both task-directed and conversational speech.

Speech data

We study two existing data sets of English speech in AD, i.e., DementiaBank (MacWhinney, Fromm, Forbes, & Holland, 2011) and the Carolina Conversations Collection (CCC) (Pope & Davis, 2011). DementiaBank (part of the TalkBank project) contains English audio and transcriptions of verbal interviews, collected between 1983 and 1988 at the University of Pittsburgh. DementiaBank also contains demographic information about the speakers, including age, sex, and years of education. Information on this cohort, including an extensive neuropsychological and physical assessment, was made available by Becker, Boller, Lopez, Saxton, and McGonigle (1994).

In DementiaBank, verbal interviews were recorded during the "`Cookie Theft" picture description component of the Boston Diagnostic Aphasia Examination (Kaplan, Goodglass, & Weintraub, 2001). Participants were asked by the interviewer to "`tell [them] everything [they] see going on in this picture". The speech was manually transcribed at the word level in accordance to the TalkBank CHAT protocol (MacWhinney, 2000). For the current work, we split the participants into two cohorts: the AD group includes participants with a diagnosis of either probable or possible AD (total $N = 196$), and the control (CT) cohort is composed of older adults without AD ($N = 98$). The AD participants in DementiaBank produce an average of 104.3 (*SD*: 59.0) words per narrative, while the control participants produce an average of 114.4 (*SD*: 59.5) words per narrative, although the distribution in both cases is somewhat right-skewed.

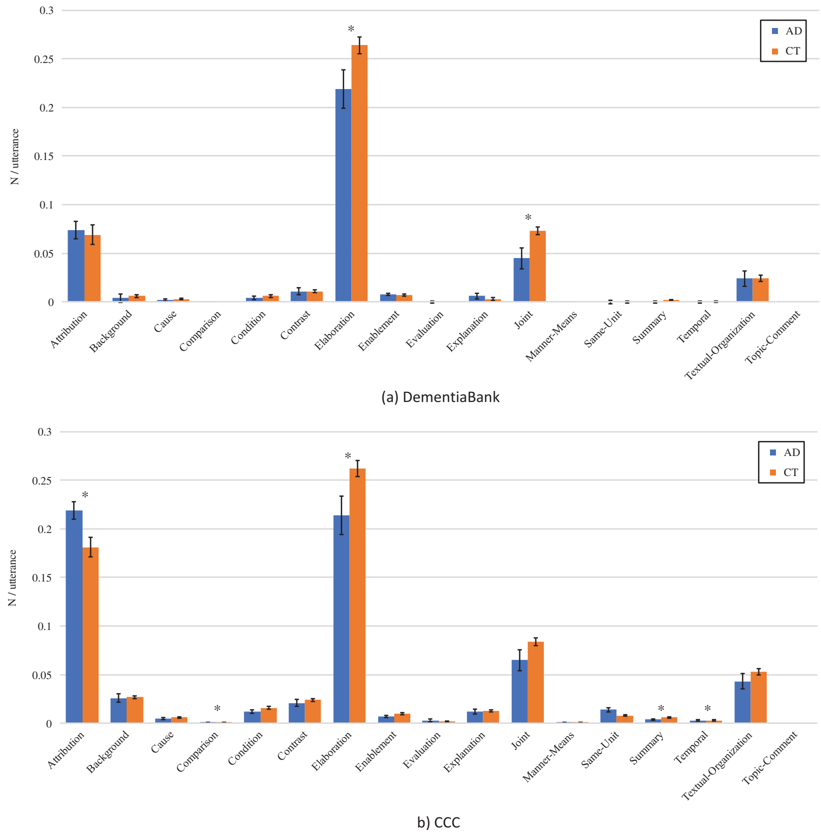
The CCC is a digital collection of natural and unstructured conversations. It is hosted at the Medical University of South Carolina, is supported by the National Libraries of Medicine, and has roots in a collection developed at the University of North Carolina at Charlotte. We examine over 400 conversations between seniors, totalling over 70.45 h of participant-only speech, each averaging 12.18 min. The interviewers were trained to have effective natural conversations with older people with dementia, based on the techniques developed by Davis and Smith (2009) and Pope and Ripich (2006), to promote interaction and participation. The participants (all 60+ years of age) are multi-ethnic, and primarily from North Carolina. In CCC, there are 55 people in the AD group (producing an average of 629.67 (*SD*: 610.32) words per conversation), and 10 in the CT group (with 564.26 words per conversation (*SD*: 205.9)). Although audio recordings exist in both DementiaBank and CCC, we focus on the textual transcripts in the current work.

In addition to the details provided earlier, paralinguistic annotations (e.g., paraphasias) were removed, and punctuation retained. In both data sets, we removed any dialogue produced by the interviewer or third parties (according to provided annotations), to focus on the interviewee.

As may be expected (Fraser et al., 2015), individuals with AD speak more slowly than those without the disease. On average, in DementiaBank, individuals with AD produce 120.6 words per minute (*SD*: 52.2) and those without produce 146.4 (*SD*: 21.6), which is significantly different at $p < 0.01$, given a one-way ANOVA. In CCC, this average is 90.0 (*SD*: 31.2) for those with AD and 103.8 (*SD* 26.3) for those without ($p < 0.01$). For this reason, our basis of analysis is the number of RST relations per utterance, rather than per unit of time, where an utterance is an uninterrupted turn in the dialogue.

Results: RST in spontaneous speech

We use ANOVA, as described in the "Methodology" section, to determine which RST relations are significantly different between groups. The histograms of Figure 3(a,b)



(a) DementiaBank

b) CCC

Figure 3. Histograms of average RST relation counts per utterance from Table 1, for AD and CT groups, in the DementiaBank (a) and CCC (b) data sets, with standard error bars ($\sigma/\sqrt{n_i}$, where n_i is the number of observations in relation *i*). Significantly different relations, after correction for multiple comparisons, are denoted with an asterisk.

Table 2. The means (μ) and variances (σ^2) of the RST relations, normalized per utterance of speech, with significant class discrimination according to the ANOVA method, for DementiaBank and CCC data.

	Feature	*p*-Value	AD	CT
			μ (σ^2)	μ (σ^2)
DementiaBank	Elaboration	0.016	0.22 (0.18)	0.26 (0.17)
	Joint	0.002	0.05 (0.08)	0.07 (0.09)
CCC	Attribution	0.014	0.22 (0.11)	0.18 (0.17)
	Comparison	0.049	0.01 (0.02)	0.00 (0.05)
	Elaboration	0.025	0.21 (0.24)	0.26 (0.14)
	Summary	0.013	0.01 (0.02)	0.01 (0.02)
	Temporal	0.028	0.01 (0.01)	0.01 (0.01)

show the counts for each RST relation type shown in Table 1, normalized per utterance. Unsurprisingly, CT speakers generally produce more RST relations per utterance, which is true across all relations except most notably "Attribution" in both data sets. Only "Elaboration" is significantly different between groups in both data sets. To evaluate these 16-dimensional distributions, we compute the Lawley–Hotelling trace and the associated statistics for all data using a multivariate analysis of covariance (MANCOVA) using population (AD or CT) and data set as the two binary grouping variables, and controlling for sex, age, and education as covariates in the model. There are significant linear effects of data set ($F_{2\times2,797} = 478.2, p < 1.0 \times 10^{-4}$) and population ($F_{2\times2,797} = 95.4, p < 1.0 \times 10^{-4}$) on the vector of RST relations, and a significant interaction between data set and population ($F_{2\times2,797} = 51.0, p < 1.0 \times 10^{-4}$).

Table 2 provides statistics for those RST relations with significant differences between groups. As may be expected, people with AD are less likely to provide additional detail to nuclei EDUs, by this analysis. Attribution is the second most frequent RST relation in both data sets, across both groups, but is only significantly discriminative in CCC.

As an aside, we expand the evaluation in the DementiaBank data to all pathologies indicated, namely (1) probable AD ($N = 173$); (2) possible AD + other dementia ($N = 50$); (3) vascular dementia ($N = 15$); (4) other dementia including Parkinson's ($N = 3$); (5) complaints of problems but none diagnosed ($N = 3$); (6) mild cognitive impairment, language, and memory type (MCI-lang, $N = 19$); (7) mild cognitive impairment + general anxiety, depression, or cerebrovascular disease (MCI-psych, $N = 7$); and (8) CT ($N = 121$). Sets (1) and (2) form partitions within the data originally analyzed. After Bonferroni correction, we find significant differences between "probable AD" and CT on Joint ($F_1 = 1.05$, $p < 3.398 \times 10^{-5}$) and Elaboration ($F_1 = 12.29$, $p < 0.0012$), between "probable AD" and MCI-psych on Temporal ($F_1 = 6.54$, $p < 6.3110 \times 10^{-5}$), and between MCI-psych and CT on Elaboration ($F_1 = 13.0$, $p < 0.0017$).

Discussion and future work

This paper demonstrates that AD has a significant effect on specific discourse relations in speech, using a novel application of a computational method on relatively large data sets. This expands on other studies that have previously associated dementia with shrinking vocabulary (Le, Lancashire, Hirst, & Jokel, 2011; Fraser et al., 2015), abrupt topic changes (Sunderman, 2012), and discontinuity in semantic cohesion (Ripich et al., 1991; Seixas Lima et al., 2016). Importantly, given its reliability (Den Ouden, 2004), the established connection of RST analysis with human communication and cognition (Taboada & Mann, 2006) motivates this novel application of this theory to language use in cognitive decline.

We observed significant differences in the discourse structure between people with AD and healthy controls in transcripts of spontaneous speech in DementiaBank and CCC. These differences occur despite relatively short conversations. We also observed, in DementiaBank (where the diagnoses were available), that fine-grained differences exist even between specific subtypes of dementia (including two variants of MCI). Whether the differences in the use of RST relations between these pathological subgroups depend on the aetiologies of those pathologies is yet to be determined. Moreover, the fact that not all subgroups are differentiable in this analysis remains an open challenge, and will require additional data.

Despite the differences in task across the data sets, Elaboration, Attribution, and Joint are the three most frequent RST relations across all groups, although Attribution is significantly different only between groups in CCC, and Joint is significant only in DementiaBank. This difference may be important – if healthy older adults are more likely than those with AD to attribute speech or positions to third parties in free conversation (as in CCC) than in more directed picture description (as in DementiaBank), the former may be more suitable for elicitation of this aspect of the theory of mind. Indeed, the ability to attribute mental states, thoughts, feelings, and positions to others is the hallmark of the theory of mind, and is significantly afflicted by AD (Heitz et al., 2016). Elaboration, however, significantly differentiates groups in both data sets, which may merely be due to the relative facility of healthy older adults with deeper, or more complex, semantic relations. We also note that the only RST relation to be more frequent among people with AD, across both data sets, is the Attribution relation, although this is significant only in CCC. Unlike Elaboration, satellites in Summary provide no new information over the nucleus, which is consistent with the relatively repetitive nature of speech in AD. While the techniques presented in this work are promising, ongoing work needs to overcome several limitations. Specifically, more data will need to be collected in different tasks and from people with different ethnolinguistic backgrounds to establish the generalizability of these results. We are also interested in examining how accurately different fine-grained types of dementia can be distinguished with this approach.

The developers of RST admit that it was designed for written monologue (Mann & Thompson, 1988), and not necessarily for spoken dialogue. Ongoing work must further establish the validity of RST on speech, especially speech produced with AD. It will also be important to gauge any potential differences in accuracy between RST analysis of people with and without AD, and to further explain the differences we have observed between populations. Given lexico-syntactic and semantic differences in AD (Fraser et al., 2015), variance in tree structure should also be taken into account. Other taxonomies more specific to dialogues, such as trouble-indicating behaviours (Orange & Purves, 1996), should continue to be added to procedures amenable to automatic processing. Rudzicz, Wang, Begum, and Mihailidis (2015), for example, examine how individuals with AD exhibit pragmatic confusion in dyadic speech-based interaction.

Ongoing work: longitudinal written narratives

Longitudinal changes to language, in the presence of dementia, can sometimes be undetectable at smaller timescales. For instance, Kemper, Marquis, Thompson, and Marquis (2001) found that grammatical complexity and propositional content decline with age over decades, even for healthy older adults, and that AD manifests as an acceleration of these declines. Le et al. (2011) performed various linguistic analyses on 51 novels by three prolific English authors and showed that degradation in lexical measures, such as type/token ratios and word-type introduction rates, over decades could indicate cognitive decline. As an aside, we apply an RST analysis to these same data, in order to see if these surface-level lexical changes have deeper counterparts.

Novels by four English authors were digitized (Le et al., 2011), specifically: 20 novels by Iris Murdoch (diagnosed with AD; average of 10,582 sentences per book; average sentence length of 12 word tokens), 16 novels by Ross Macdonald (diagnosed with AD; average of 7924 sentences per book; average sentence length of 9 word tokens), 15 novels by P.D. James (not diagnosed with AD; average 8372 sentences per book; average sentence length of 13 word tokens), and 16 novels by Agatha Christie (not officially diagnosed, but compelling evidence exists of cognitive decline similar to AD (Lancashire & Hirst, n.d.); average of 6432 sentences per book; average sentence length of 10 word tokens).[3]

We compare the counts of the RST relations between the works of Murdoch and Macdonald on one hand and James on the other, over all of their novels. There is more variation in these data than in CCC or DementiaBank, given the open nature of written narratives. As a means of standardization, we normalize the counts of RST relations by the total number of sentences in each book. Here, 9 of the 16 RST relations are significant after Bonferroni correction, as depicted in Table 3. This is consistent with the results from our experiments on speech data in the previous section – the significant features were generally produced less frequently by authors with AD. Table 4 shows regression and correlation statistics for the two most significant relations, Enablement and Condition. The use of the Condition relation is strongly correlated with age across all authors, and decreases only for Macdonald, but the frequency of Condition is again more indicative of cognitive health than the change in the frequency of Condition over time. The directions of change for the frequencies of RST relations appear less discriminative than their absolute

Table 3. The *p*-values, means, and variances of the RST relations whose frequencies, normalized by novel length in sentences, are significantly different across groups (Murdoch and Macdonald on one hand, and James on the other), according to a one-way ANOVA.

Feature	*p*-Value	AD	CT
		μ (σ^2)	μ (σ^2)
Comparison	7.47×10^{-5}	0.003 (9.00×10^{-7})	0.004 (8.00×10^{-7})
Elaboration	9.84×10^{-4}	0.264 (0.009)	0.354 (0.001)
Attribution	6.49×10^{-5}	0.247 (0.003)	0.311 (0.000)
Enablement	6.23×10^{-9}	0.017 (2.69×10^{-5})	0.026 (3.10×10^{-6})
Condition	1.00×10^{-14}	0.022 (1.44×10^{-5})	0.035 (9.26×10^{-6})
Background	1.87×10^{-8}	0.043 (8.23×10^{-5})	0.060 (2.48×10^{-5})
Explanation	1.40×10^{-3}	0.005 (5.25×10^{-6})	0.007 (1.63×10^{-6})
Contrast	1.76×10^{-5}	0.037 (1.87×10^{-4})	0.057 (1.51×10^{-4})
Manner-Means	1.60×10^{-3}	0.004 (2.91×10^{-6})	0.005 (7.35×10^{-7})

Table 4. Regression and correlation statistics for each author, between age and for the two most significant RST relations from Table 3 (Enablement and Condition). The authors diagnosed with AD used these relations less frequently than the author without AD.

	Enablement		Condition	
Author	Regression	Correlation	Regression	Correlation
Agatha Christie	$R^2 = 0.06$	$r = 0.25, p = 0.36$	$R^2 = 0.43$	$r = 0.66, p = 0.006$
Iris Murdoch	$R^2 = 0.04$	$r = -0.19, p = 0.42$	$R^2 = 0.26$	$r = 0.51, p = 0.02$
P.D. James	$R^2 = 0.01$	$r = 0.10, p = 0.72$	$R^2 = 0.27$	$r = 0.52, p = 0.05$
Ross Macdonald	$R^2 = 0.24$	$r = -0.49, p = 0.05$	$R^2 = 0.30$	$r = -0.55, p = 0.03$

frequencies, across all authors. This observation is in contrast to work on the same data (Le et al., 2011), which showed that various lexico-syntactic features vary significantly over time. This also provides a longitudinal analysis that is not possible with the previously discussed spontaneous speech data, and which involves different discourse relations. The fact that Condition and Enablement are both related to logical contingencies, realized and unrealized respectively, may deserve further study.

Conclusion

The Elaboration and Attribution relations are significant indicators of AD across the free conversations of our speech data, and the written narratives of our ongoing work, and are also relevant in other domains. For example, Wolf and Gibson (2005) collected 135 newswire texts and annotated them with *coherence* relations. Although their discourse relations were based mostly on the work of Hobbs (1985), there was a large conceptual overlap between the discourse units (including Condition, Contrast, Attribution, and Elaboration). The analysis of their database showed that Elaboration and Attribution compose the majority of discourse relations with 44.6% and 14.5%, respectively. The frequent occurrence of these relations, in both speech and text, regardless of the task at hand, suggests further study, especially within a clinical context.

Despite its rising prevalence, AD remains under-diagnosed (Okie, 2011). Controversy surrounds routine screening for early diagnosis of cognitive disorders, including AD, due in part to the stresses involved in explicit assessment. To the extent that automated methods for assessment can be run, with consent but without explicit intervention by either healthcare providers or patients, during everyday activities such as conversation, these methods may offer unique benefits to clinical assessment. Our analyses suggest that RST can be applied to the clinical study of AD, and that several significant differences in discourse emerge from that analysis, with considerable consistency across tasks and types of data. Future work should extend similar techniques to different task types, and combine statistical analysis in RST with lexico-syntactic features in automated assessment.

Notes

1. Bottom-up approaches build more abstract structures from smaller ones, starting with the atomic units themselves.
2. http://verbs.colorado.edu/ mpalmer/projects/verbnet.

3. Murdoch and Macdonald constitute the AD group, while James is the CT group in further statistical group analysis. Literary work can involve considerable review and editing, with assistants and editors changing the author's original writing. However, there is no evidence that this is the case for the authors we consider. Le et al. (2011), drawing on Lancashire (2010), reviews the writing and editing processes of Christie, Murdoch, and James, and conclude that the later novels of each author do not deviate from the author's earlier practices; in particular, Murdoch allowed no one else to edit her writing at all.

Acknowledgements

Frank Rudzicz's contribution is funded by an NSERC Discovery grant (RGPIN 435874) and by a Young Investigator award by the Alzheimer Society of Canada. Graeme Hirst's contribution is funded by an NSERC Discovery grant. The original acquisition of the DementiaBank data was supported by NIH grants AG005133 and AG003705 to the University of Pittsburgh, and maintenance of the data archive is supported by NIH-NIDCD grant R01-DC008524 to Carnegie Mellon University. The authors thank Vanessa Wei Feng for providing the software tool used in the analysis of RST relations. We acknowledge and thank Professor Ian Lancashire for providing the Ross Macdonald data.

Disclosure statement

Frank Rudzicz is a co-founder of a software company, called WinterLigtht Labs Inc, that commercializes automatic speech-based assessment of cognitive disorder. No intellectual property of that company has been used in this research.

Funding

Frank Rudzicz's contribution is funded by an Natural Sciences and Engineering Research Council of Canada grant [RGPIN 435874] and by a Young Investigator award by the Alzheimer Society of Canada. Graeme Hirst's contribution is funded by an NSERC Discovery grant. The original acquisition of the DementiaBank data was supported by NIH grants [AG005133 and AG003705] to the University of Pittsburgh, and maintenance of the data archive is supported by NIH-NIDCD grant [R01-DC008524] to Carnegie Mellon University.

References

Ahmed, S., de Jager, C. A., Haigh, A.-M., & Garrard, P. (2013). Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology*, *27*, 79–85. doi:10.1037/a0031288

Almor, A., Kempler, D., MacDonald, M. C., Andersen, E. S., & Tyler, L. K. (1999). Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease. *Brain and Language*, *67*, 202–227. doi:10.1006/brln.1999.2055

Becker, J. T., Boller, F., Lopez, O. I., 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*, 585–594. doi:10.1001/archneur.1994.00540180063015

Bhaduri, S., Das, R., & Ghosh, D. (2016). Non-invasive detection of Alzheimer's disease – multifractality of emotional speech fractal and multifractal analysis and speech. *Journal of Neurology and Neuroscience*, *7*, 1–7. doi:10.21767/2171-6625.100084

Carlson, L., & Marcu, D. (2001). *Discourse tagging reference manual (Tech. Rep.)*. Los Angeles, CA: University of Southern California Information Sciences Institute.

Carlson, L., Marcu, D., & Okurowski, M. E. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of Second SIGDial Workshop on Discourse and Dialogue* (pp. 1–10). Aalborg: Association for Computational Linguistics.

Chai, J. Y., & Jin, R. (2004). Discourse structure for context question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004* (pp. 23–30). Association for Computational Linguistics.

Chapman, S. B., Zientz, J., Weiner, M., Rosenberg, R., Frawley, W., & Burns, M. H. (2002). Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging. *Alzheimer Disease & Associated Disorders*, *16*, 177–186. doi:10.1097/00002093-200207000-00008

Charniak, E., & Johnson, M. (2005). Coarse-to-fine *n*-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)* (Vol. 1, pp. 173–180). Association for Computational Linguistics.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, *20*, 273–297. doi:10.1007/BF00994018

Davis, B. H., & Smith, M. K. (2009). Infusing cultural competence training into the curriculum: Describing the development of culturally sensitive training on dementia communication. *Kaohsiung Journal of Medical Sciences*, *25*, 503–509. doi:10.1016/S1607-551X(09)70557-1

den Ouden, H. (2004). *Prosodic realizations of text structure* (PhD). Tilburg: University of Tilburg

den Ouden, H., Van Wijk, C., Terken, J., & Noordman, L. (1998). *Reliability of discourse structure annotation* (Vol. 33, Tech. Rep.). IPO Center for Research on User-System Interaction, Technical University of Eindhoven.

Duverle, D. A., & Prendinger, H. (2009). A novel discourse parser based on support vector machine classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Vol. 2, pp. 665–673). Association for Computational Linguistics.

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*, 365–370. doi:10.1002/(ISSN)1531-8249

Feng, V. W. (2014). *RST-Style Discourse Parsing and Its Applications in Discourse Analysis* (PhD Thesis). Toronto: University of Toronto.

Feng, V. W., & Hirst, G. (2014). A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL14)* (p. 511–521). Assocation for Computational Linguistics.

Fisher, S., & Roark, B. (2007). The utility of parse-derived features for automatic discourse segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)* (pp. 488–495). Assocation for Computational Linguistics.

Forbes-McKay, K., Shanks, M. F., & Venneri, A. (2013). Profiling spontaneous speech decline in Alzheimer's disease: A longitudinal study. *Acta Neuropsychiatrica*, *25*, 320–327. doi:10.1017/neu.2013.16

Fraser, K., Rudzicz, F., & Rochon, E. (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *Proceedings of Interspeech 2013* (pp. 2177–2181). Lyon, France: Assocation for Computational Linguistics.

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2015). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, *49*, 407–422. doi:10.3233/JAD-150520

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

Heitz, C., Noblet, V., Phillipps, C., Cretin, B., Vogt, N., Philippi, N., … Blanc, F. (2016). Cognitive and affective theory of mind in dementia with Lewy bodies and Alzheimer's disease. *Alzheimer's Research & Therapy*, *8*, 10. doi:10.1186/s13195-016-0179-9

Hernault, H., Bollegala, D., & Ishizuka, M. (2010). A sequential model for discourse segmentation. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 315–326). Springer.

Hobbs, J. R. (1985). *On the coherence and structure of discourse* (Tech. Rep.). Technical report CSLI-85-37. Center for the Study of Language and Information, Stanford, CA: Stanford University.

Hutchinson, J., & Jensen, M. (1980). A pragmatic evaluation of discourse communication in normal and senile elderly in a nursing home. In L. Obler & M. Albert (Eds.), *Language and communication in the elderly* (pp. 59–74). Lexington, MA: D. C. Heath and Company.

Joty, S., Carenini, G., Ng, R., & Mehdad, Y. (2013). Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)* (pp. 486–496). Sofia Bulgaria.

Joty, S., Carenini, G., & Ng, R. T. (2012). A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the Joint meeting of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning (EMNLP-CoNLL 2012)* (pp. 904–915). Association for Computational Linguistics.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test* (2nd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Kemper, S., Marquis, J., Thompson, M., & Marquis, J. (2001). Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging*, *16*, 600–614. doi:10.1037/0882-7974.16.4.600

Kempler, D., & Goral, M. (2008). Language and dementia: Neuropsychological aspects. *Annual Review of Applied Linguistics*, *28*, 72–90. doi:10.1017/S0267190508080045

Kirshner, H. S., Webb, W. G., & Kelly, M. P. (1984). The naming disorder of dementia. *Neuropsychologia*, *22*, 23–30. doi:10.1016/0028-3932(84)90004-6

Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, *18*, 35–62. doi:10.1080/01638539409544883

Lancashire, I. (2010). *Forgetful muses: Reading the author in the text*. Toronto: University of Toronto Press.

Lancashire, I., & Hirst, G. (n.d.). Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: A case study. In *19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice* (pp. 1–5), 8–10 March 2009, Toronto.

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*, 435–461. doi:10.1093/llc/fqr013

Lee, L. (2002). Fast context-free grammar parsing requires fast Boolean matrix multiplication. *Journal of the Association for Computing Machinery*, *49*, 1–15. doi:10.1145/505241.505242

Lin, Z., Kan, M.-Y., & Ng, H. T. (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)* (pp. 343–351). Association for Computational Linguistics.

Louis, A., Joshi, A., & Nenkova, A. (2010). Discourse indicators for content selection in summarization. In *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, (pp. 147–156). Association for Computational Linguistics.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, *25*, 1286–1307. doi:10.1080/02687038.2011.589893

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Interdisciplinary Journal for the Study of Discourse*, *8*, 243–281.

Marcu, D., Romera, M., & Amorrortu, E. (1999). Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging* (pp. 48–57). Association for Computational Linguistics.

Müller, N., & Guendouzi, J. A. (2002). Transcribing discourse: Interactions with Alzheimer's disease. *Clinical Linguistics & Phonetics*, *16*, 345–359. doi:10.1080/02699200210135875

Noordman, L., Dassen, I., Swerts, M., & Terken, J. (1999). Prosodic markers of text structure. In K. van Hoek, A. Kibrik, & L. Noordman (Eds.), *Discourse studies in cognitive linguistics* (pp. 133–148). Amsterdam: John Benjamins Publishing.

Okie, S. (2011). Confronting Alzheimer's disease. *The New England Journal of Medicine*, *365*, 1069–1072. doi:10.1056/NEJMp1107288

Orange, J. B., & Purves, B. (1996). Conversational discourse and cognitive impairment: Implications for Alzheimer's disease. *Journal of Speech-Language, Pathology and Audiology*, *20*, 139–150.

Orimaye, S. O., & 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). Baltimore, Maryland.

Pope, C., & Davis, B. H. (2011). Finding a balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory*, *7*, 143–161. doi:10.1515/cllt.2011.007

Pope, C., & Ripich, D. N. (2006). Speak to me, listen to me: Ethnic and gender variations in talk and potential consequences in interactions for people with Alzheimer's disease. In B. Davis (Ed.), *Alzheimer talk, text and context* (pp. 37–59). Basingstoke: Palgrave Macmillan.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 1–4). European Language Resources Association (ELRA).

Prasad, R., Joshi, A., Dinesh, N., Lee, A., & Miltsakaki, E. (2005). The penn discourse treebank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for NLG*. Birmingham

Reilly, J., Troche, J., Grossman, M. (2011). Language processing in dementia. In A. E. Budson & Kowall, N. W. (Eds.), *The handbook of Alzheimer's disease and other dementias*. Hoboken, NJ: Wiley-Blackwell.

Ripich, D. N., Vertes, D., Whitehouse, P., Fulton, S., & Ekelman, B. (1991). Turn-taking and speech act patterns in the discourse of senile dementia of the Alzheimer's type patients. *Brain and Language*, *40*, 330–343. doi:10.1016/0093-934X(91)90133-L

Roth, M., Tym, E., & Mountjoy, C. Q. (1986). CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. *British Journal of Psychiatry*, *149*, 698–709. doi:10.1192/bjp.149.6.698

Rudzicz, F., Wang, R., Begum, M., & Mihailidis, A. (2015). Speech interaction with personal assistive robots supporting aging at home for individuals with Alzheimer's disease. *ACM Transactions on Accessible Computing*, *7*, 1–22. doi:10.1145/2785580

Scholman, M., Evers-Vermeul, J., & Sanders, T. J. (2016). A step-wise approach to discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse*, *1*, 1–28.

Seixas Lima, B., Graham, N. L., Leonard, C., Black, S. E., Wai, D. F. T., Freedman, M., … Rochon, E. (2016). The effect of semantic memory impairment on the speech of svPPA patients: A discourse-level analysis. In *Proceedings of the International Clinical Phonetics and Linguistics Association Conference (ICPLA)*. Halifax, Canada.

Snowdon, D. A., Kempler, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markes-Bery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in later life. *Journal of American Medical Association*, *275*, 528–532. doi:10.1001/jama.1996.03530310034029

Sunderman, M. (2012). *The Effects of Alzheimer's Disease on Expressive Language Over Time* (Unpublished doctoral dissertation). Columbus, OH: Ohio State University.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in Aging Neuroscience*, *7*, 1–7. doi:10.3389/fnagi.2015.00195

Taboada, M., & Mann, W. C. (2006). Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, *8*, 423–459. doi:10.1177/1461445606061881

Taler, V., & Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Psychology*, *30*, 501–556.

Webber, B. (2004). D-LTAG: Extending lexicalized TAG to discourse. *Cognitive Science*, *28*, 751–779. doi:10.1207/s15516709cog2805_6

Weiner, M. F., Neubecker, K. E., Bret, M. E., & Hynan, L. S. (2008). Language in Alzheimer's disease. *The Journal of Clinical Psychiatry*, *69*, 1223–1227. doi:10.4088/JCP.v69n0804

Wilson, R., Rochon, E., Mihailidis, A., & Leonard, C. (2012). Examining success of communication strategies with Alzheimer's disease during an activity of daily living. *Journal of Speech, Language, and Hearing Research*, *55*, 328–341. doi:10.1044/1092-4388(2011/10-0206)

Wolf, F., & Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, *31*, 249–287. doi:10.1162/0891201054223977

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 (ACL 2016)*. Association for Computational Linguistics.