

Idea density for predicting Alzheimer’s disease from transcribed speech

Kairit Sirts¹, Olivier Piguet^{2,3} and Mark Johnson⁴

¹Institute of Computer Science, University of Tartu

²School of Psychology and Brain & Mind Centre, The University of Sydney

³Neuroscience Research Australia, The University of New South Wales

⁴Department of Computing, Macquarie University

kairit.sirts@ut.ee, olivier.piguet@sydney.edu.au

mark.johnson@mq.edu.au

Abstract

Idea Density (ID) measures the rate at which ideas or elementary predications are expressed in an utterance or in a text. Lower ID is found to be associated with an increased risk of developing Alzheimer’s disease (AD) (Snowdon et al., 1996; Engelman et al., 2010). ID has been used in two different versions: propositional idea density (PID) counts the expressed ideas and can be applied to any text while semantic idea density (SID) counts pre-defined information content units and is naturally more applicable to normative domains, such as picture description tasks. In this paper, we develop DEPID, a novel dependency-based method for computing PID, and its version DEPID-R that enables to exclude repeating ideas—a feature characteristic to AD speech. We conduct the first comparison of automatically extracted PID and SID in the diagnostic classification task on two different AD datasets covering both closed-topic and free-recall domains. While SID performs better on the normative dataset, adding PID leads to a small but significant improvement (+1.7 F-score). On the free-topic dataset, PID performs better than SID as expected (77.6 vs 72.3 in F-score) but adding the features derived from the word embedding clustering underlying the automatic SID increases the results considerably, leading to an F-score of 84.8.

1 Introduction

Idea density (ID) measures the rate of propositions or ideas expressed per word in a text and it is connected to some very interesting results from neuroscience related to Alzheimer’s disease (AD). In

The old gray [MARE] has a very large [NOSE].

Dependencies	Propositions
det(The, mare)	
amod(old, mare)	(OLD, MARE)
amod(gray, mare)	(GRAY, MARE)
nsubj(mare, has)	(HAS, MARE, NOSE)
det(a, nose)	
advmod(very, large)	(VERY, (LARGE, NOSE))
amod(large, nose)	(LARGE, NOSE)
dobj(nose, has)	(HAS, MARE, NOSE)
punct(., has)	

Table 1: The alignment of the dependency and propositional structures. The example sentence is due to Brown et al. (2008). The predicative proposition (HAS, MARE, NOSE) is represented by two dependency arcs.

particular, two longitudinal studies—the Nun Study (Snowdon et al., 1996) and the Precursors Study (Engelman et al., 2010)—suggest that lower ID, as measured from the essays written in young age, is associated with the higher probability of developing AD in later life.

Two alternative definitions of idea density have been used in relation to AD. **Propositional idea density** (PID) counts the number of *any* ideas expressed in the text, setting no restriction to the topic (Turner and Greene, 1977; Chand et al., 2010). An example sentence with its ideas or propositions is given in Table 1. Based on each proposition a question can be formulated with a yes or no answer. Removing a proposition from a sentence changes the semantic meaning of that sentence. For instance, removing the proposition (GRAY, MARE) from the example makes the overall meaning of the sentence more general. The PID is then computed by normalising the proposition count with the token count and thus the PID of the example given in Table 1 is $6/9 \approx 0.667$.

The existing tool for automatic PID computation,

CPIDR (Brown et al., 2008), is based on counting POS tags. However, we noticed that the propositional structure of a sentence is very similar to its dependency structure, see the first column in Table 1. This motivated us to come up with DEPID, a method for computing PID from dependency structures. In addition, DEPID more easily enables to consider idea repetition which has been shown to be a characteristic feature in Alzheimer’s speech (Bayles et al., 1985; Tomoeda et al., 1996; Bayles et al., 2004), resulting in a modified PID version DEPID-R which excludes the repeated ideas.

Semantic idea density (SID) (Ahmed et al., 2013a,b) relies on a set of pre-defined information content units (ICU). ICU is an object or action that can be seen on the picture or is told in the story and is expected to be mentioned in the narrative. For instance, assuming that the words in capital letters and square brackets in the example sentence shown in Table 1 belong to the set of pre-defined ICUs the SID is computed by normalising the ICU count with the token count: $2/9 \approx 0.222$. Recently, Yancheva and Rudzicz (2016), proposed a method for computing SID based on word embedding clusters. We use their method for computing SID as it does not rely on any pre-defined ICU inventory and thus is applicable also on free-topic datasets.

PID and SID are complementary definitions of idea density with SID being naturally applicable in standardised picture description or story re-telling tasks while PID is more suitable on datasets of spontaneous speech on free topics.

In this paper we study the predictiveness of both PID and SID features in the diagnostic classification task for predicting AD. To that end, we conduct experiments on two very different datasets: DementiaBank, which consists of transcriptions of a normative picture description task, and AMI, which contains autobiographical memory interviews describing life events freely chosen by the subjects.

We show that on the DementiaBank data the POS-based PID scores are actually higher for AD patients than they are for normal controls, contrary to the expectations from the AD literature (Engelman et al., 2010; Chand et al., 2012; Kemper et al., 2001). By studying the characteristics of the DementiaBank we are able to adapt DEPID such that its PID values become significantly different between the patient and control groups in the expected direction. Thus, we believe that our proposed DEPID is a better tool for measuring PID as

described by neurolinguists on spontaneous speech transcripts than the POS-based CPIDR.

Secondly, we show that the SID performs better than PID on the constrained-domain DementiaBank corpus but adding the PID feature leads to a small but significant improvement.

Thirdly, we show that on the free-topic AMI dataset the PID performs better than the automatically extracted SID, but adding the features derived from the word embedding clustering underlying the SID, modeling the broad discussion topics, increases the results considerably—an effect which is less visible on the constrained topic DementiaBank.

The contributions of this paper are the following:

1. Development of DEPID, the new dependency-based method for automatically computing PID and its version DEPID-R which enables to detect and exclude idea repetitions;
2. Analysis of the characteristic features of the DementiaBank dataset and the proposal for modifying DEPID to make it applicable to this and other similar closed-topic datasets.
3. Results of extensive diagnostic classification experiments using PID, SID and several related baselines on two very different AD datasets.

2 Idea density and Alzheimer’s disease

ID was first associated with AD in the Nun Study (Snowdon et al., 1996), based on a cohort of elderly nuns participating in a longitudinal study of aging and Alzheimer’s disease. In this work, they studied the autobiographical essays the nuns had written decades ago in their youth. The nuns were divided into three groups based on their ID score computed from the essays, so that each group covered 33.3% percentile of the whole range of ID values. The lowest group was labeled as having low ID and the medium and highest group as having high ID. These groups were established from a sample of 93 nuns. The association between AD and ID was studied on a sample of 25 nuns who had died by the time of the study, for 10 of whom the cause of death had been marked as AD. The study found that most subjects with AD belonged to the low ID group while most of those, who did not develop AD, belonged to the group with high ID, thus suggesting that the low ID in youth might be associated with the development of the AD in later life.

Similar work was conducted on a group of medical students for whom essays from the time of their admission to the medical school several decades earlier were available (Engelman et al., 2010). The results of this study also showed a significantly lower ID on the AD group as compared to the healthy controls, suggesting that ID could be an important discriminative feature for predicting AD.

2.1 Propositional and semantic idea density

Two different versions of ID have been developed over time, both derived from the propositional base structure developed by Kintsch and Keenan (1973) to describe the semantic complexity of texts in reading experiments.

Propositional idea density (PID), which was used both in the Nun Study and the medical students study, is based on counting the semantic propositions as defined by Turner and Greene (1977) and later refined by Chand et al. (2012). Three main types of propositions were described: 1) predications that are based on verb frames; 2) modifications that include all sorts of modifiers, e.g. adjectival, adverbial, quantifying, qualifying etc.; and 3) connections that join simple propositions into complex ones. For each proposition, a question can be formed with a yes or no answer. For instance, based on the example in Table 1, we could form the following questions:

1. Is the mare old?
2. Is the mare gray?
3. Has the mare a nose?
4. Is the nose large?
5. Is the nose very large?

Each of those questions inquires about a different aspect of the whole sentence and is a basis of an idea or proposition.

Semantic idea density (SID) has retained its relation to the propositional base of some text. It relies on a set of information content units (ICUs) that have been pre-defined for a closed-topic task, such as picture description or story re-telling. For instance, different inventories of 7-25 ICUs have been described for the Cookie Theft picture task (Goodglass and Kaplan, 1983), listing objects visible on the picture such as “boy”, “girl”, “cookie” or “kitchen” or actions performed on the scene such as “boy stealing cookies” or “woman drying dishes”. SID is computed by counting the number of ICUs mentioned in the text and then normalising by the total number of word tokens.

2.2 Related work on AD using ID

PID, computed with CPIDR, has been used in few previous works for predicting AD. Jarrold et al. (2010) used PID as one among many features and reported it as significant. They obtained a classification accuracy of 73% on their dataset, which contained short structured clinical interviews, with their best model and feature set that also included the PID feature. PID was also used by Roark et al. (2011) to detect mild cognitive impairment on a story re-telling dataset. However, they found no significant difference between groups in terms of PID and thus, their feature selection procedure most probably filtered it out.

In terms of SID, most previous work has relied on manually defined ICUs (Ahmed et al., 2013b,a). Fraser et al. (2015) extracted binary and frequency-based ICU features. They searched for words related to the ICU objects and looked at the *nsubj*-relations in the dependency parses to detect the ICUs referring to actions. The binary feature was set when any word related to an ICU was mentioned in the text, while frequency-based features counted the total number of times any word referring to an ICU was mentioned.

Recently, Yancheva and Rudzicz (2016) proposed a method for automatically extracting ICUs and computing SID without relying on a manually defined ICU inventory. This work will be reviewed in more detail in section 4. They found that the automatically extracted ICUs and SID performed as well in a diagnostic AD classification task as the human-defined ICUs.

3 Computation of PID

Automating the computation of PID is difficult because it is essentially a semantic measure. The instructions given by Turner and Greene (1977) for counting the propositions assume the comprehension of the semantic meaning of the text, while the raw text lacks the necessary semantic annotations. However, it has been noticed that the propositions roughly correspond to certain POS tags. In particular, Snowdon et al. (1996) mention that elementary propositions are expressed using verbs, adjectives, adverbs and prepositions. This observation is the basis of the CPIDR program (Brown et al., 2008), a tool for automatically computing PID scores from text. CPIDR first processes the text with a POS-tagger, then counts all verbs, adjectives, adverbs, prepositions and coordinating

Dep rel	Proposition type
advcl	Causal connection
advmod	Qualifying modification
amod	Qualifying modification
appos	Referential predication
cc	Conjunctive connective
csubj	Predication with a clausal subject
csubjpass	Predication with a passive clausal subject
det ^a	Quantifying modification
neg	Negative modification
npadvmod	Qualifying modification
nsubj ^b	Predication subject
nsubjpass	Predication with passive subject
nummod	Quantifying modification
poss	Possessive modification
predet	Qualifying modification
preconj	Conjunctive or disjunctive connection
prep	Proposition denoting purpose, location, intention, etc.
quantmod	Quantifying modification
tmod	Qualifying modification
vmod	Qualifying modification

Table 2: Dependency relations encoding propositions.

^aexcept *a*, *an* and *the*

^bexcept *it* and *this*

conjunctions as propositions, and then applies a set of 37 rules to adjust the final proposition count.

3.1 DEPID—dependency-based PID

We propose that the dependency structure is better suited for PID computation than the POS tag counting approach adopted by the existing CPIDR program (Brown et al., 2008) because the dependency structure resembles more closely the semantic propositional structure, see Table 1. We treat each dependency type as a separate feature and manually set the feature weights to either one or zero depending on whether this dependency relation encodes a proposition or not. We make these decisions based on the dependency type descriptions in the Stanford dependency manual (de Marneffe and Manning, 2008). The dependency types with non-zero weights are listed in Table 2. The PID is then computed by summing the counts of those dependency relations and normalising by the number of word tokens. We call our dependency-based PID computation method DEPID.

We computed the Spearman correlations between CPIDR, DEPID and manual proposition counts on the 69 example sentences given in chapter 2 in (Turner and Greene, 1977)¹ and the 177

¹Similar to Brown et al. (2008), we exclude the example

	Spearman r
CPIDR vs Manual	0.795
DEPID vs Manual	0.839
DEPID vs CPIDR	0.864

Table 3: Spearman correlations between CPIDR, DEPID and manual proposition counts on the examples given in Turner and Greene (1977) and Chand et al. (2010).

example sentences given in (Chand et al., 2010), making up the total of 276 sentences. These correlations are given in Table 3. We observe that by just counting the dependency relations given in Table 2, we obtain proposition counts that correlate better with the manual counts than the POS-based CPIDR counts.

3.2 DEPID-R

It is known that the Alzheimer’s language is generally fluent and grammatical but in order to maintain the fluency the deficiencies in semantic or episodic memory are compensated with empty speech (Nicholas et al., 1985), such as repetitions, both on the word level but also on the idea, sentence or narrative level. DEPID easily enables to track repeated ideas in the narrative. We consider a proposition as repetition of a previous idea when the *deprel*(DEPENDENT LEMMA, HEAD LEMMA) tuples of the two propositions match. For instance, a sentence “*I had a happy life.*” contains three propositions: *nsubj*(I, HAVE), *obj*(LIFE, HAVE) and *amod*(HAPPY, LIFE). Another sentence “*I’ve had a very happy life.*” later in the same narrative only adds a single proposition to the total count—*advmod*(VERY, HAPPY)—as this is the only new piece of information that was added.

We modify DEPID to exclude the repetitive ideas of a narrative by only counting the proposition *types* expressed with the lexicalised *deprel*(DEPENDENT LEMMA, HEAD LEMMA) dependency arcs. We call this modified version of dependency-based PID computation method DEPID-R. The relation between DEPID-R and DEPID is that DEPID counts the *tokens* of the same propositions.

4 Computation of SID

Recently, Yancheva and Rudzicz (2016) proposed a method for automatically computing SID without

17, but for examples 18, 54, 55, 56, we include all paraphrases.

	DB		AMI	
	AD	Ctrl	AD	Ctrl
Subjects	169	98	20	20
Samples	257	241	36	20
Mean samples	1.52	2.46	1.80	1.00
Mean words	104	114	1674	1509
Std words	58	59	778	688

Table 4: Statistics of the DementiaBank (DB) and AMI datasets. *Mean samples* is the average number of samples per subject. *Mean* and *std words* are the mean number of words per sample and the respective standard deviation.

the use of manually defined ICUs. Their method relies on clustering word embeddings of the nouns and verbs found in the transcriptions, assuming that the embeddings of the words related to the same semantic unit are clustered together.

They first perform K-means clustering on the word embeddings. Then, for each cluster they compute the mean distance μ_{cl} and its standard deviation σ_{cl} . The mean distance is the average Euclidean distance of all vectors assigned to a cluster from the centroid of that cluster. Finally, for each word they compute the scaled distance as a z-score of the Euclidean distance d_E between the word embedding and its closest cluster centroid:

$$d_{scaled} = \frac{d_E - \mu_{cl}}{\sigma_{cl}}$$

The words with $d_{scaled} < 3$ are counted as automatic ICUs. SID is then computed by dividing the number of ICUs with the total number of word tokens in the transcription.

In addition to SID, [Yancheva and Rudzicz \(2016\)](#) experiment with distance-based features also derived from the same clustering. The distance feature for each cluster is computed as the average of the scaled distances of the words (nouns or verbs) in the transcript assigned to that cluster. These cluster features are not directly related to the concept of SID but they could be viewed as an automatic approximation of features derived from the human annotated ICUs.

5 Experiments

5.1 Data

We conduct experiments on two very different AD datasets. The first dataset is derived from the DementiaBank ([Becker et al., 1994](#)), which is part of a

publicly available Talkbank corpus.² It contains descriptions of the Cookie Theft picture ([Goodglass and Kaplan, 1983](#)) produced by subjects diagnosed with dementia as well as of healthy control cases. The data is manually transcribed and annotated in the CHAT format ([MacWhinney, 2000](#)), containing a range of annotations denoting various speech events. This is the same dataset used by [Yancheva and Rudzicz \(2016\)](#) and similar to them, we use the interviews of all control subjects and subjects whose diagnose is either AD or probable AD.

The second dataset, collected at NeuRA³, contains autobiographical memory interviews (AMI) of both AD patients and healthy control subjects. Each interview consists of four stories, each story describing events from a particular period of the subject’s life: teenage years, early adulthood, middle adulthood and last year. Each story has three logical parts: free recall, general probe and specific probe. In the free recall part the subject is asked to talk freely about events he remembers from the given life period. In the general recall part the interviewer helps to narrow down to a particular specific event. In the specific probe part the interviewer asks a number of predefined questions about this specific event. We use all four stories of an interview as a single sample but extract only the free recall part of each story as this is the most spontaneous part of the interview.

We preprocess both data sets similarly, following the procedure described in ([Fraser et al., 2015](#)) as closely as possible. We first extract only the patient’s dialogue turns. Then we remove any tokens that are not words (e.g. laughs). In DementiaBank corpus, such tokens can be detected by various CHAT annotations. We also remove filled pauses such as *um*, *uh*, *er*, *ah*. The statistics of both datasets are given in Table 4.

5.2 Analysis of the idea density

First, we perform a statistical analysis of the different ID measures in Table 5 on both datasets using the independent samples Wilcoxon rank-sum test to test the difference between group means.

The DEPID computed PID values are systematically lower than the CPIDR values on both datasets, suggesting that either CPIDR overestimates or the DEPID underestimates the number of propositions. In order to check that we manually annotated the

²<https://talkbank.org/DementiaBank/>

³Neuroscience Research Australia

Data	Method	AD mean (sd)	Ctrl mean (sd)
DB	CPIDR*	0.518 (0.069)	0.491 (0.057)
DB	DEPID*	0.371 (0.052)	0.356 (0.046)
DB	DEPID-R	0.339 (0.049)	0.334 (0.042)
DB	DEPID-R-ADD*	0.168 (0.064)	0.194 (0.059)
DB	SID*	0.380 (0.051)	0.427 (0.045)
AMI	CPIDR	0.524 (0.023)	0.532 (0.017)
AMI	DEPID	0.468 (0.022)	0.473 (0.017)
AMI	DEPID-R*	0.334 (0.027)	0.366 (0.027)
AMI	DEPID-R-ADD+*	0.291 (0.032)	0.337 (0.032)
AMI	SID*	0.346 (0.034)	0.385 (0.024)

Table 5: The statistics of the ID values for AD and control groups. DEPID-R ignores the repeated ideas. DEPID-R-ADD for DementiaBank additionally excludes conjunctions, sentences with *I* and *you* subjects and sentences with vague meaning. DEPID-R-ADD+ for AMI only ignores sentences with vague meaning. SID is computed based on the clustering of the whole dataset. Star (*) after the method name indicates that the difference in group means is statistically significant ($p < 0.001$).

propositions of 20 interviews from DementiaBank according to the guidelines given by Chand et al. (2012). We found that both CPIDR and DEPID overestimate the PID values although CPIDR does it to much greater extent. CPIDR both overestimates the number of propositions and underestimates the number of tokens in certain cases leading to higher PID scores. For example, CPIDR does not count contracted forms, such as “’s” in “it’s” or “n’t” in “don’t” as distinct tokens. Because there are many such forms in DementiaBank transcriptions, this behaviour considerably lowers CPIDR token counts. Also, CPIDR counts each auxiliary verb in present participle constructions as a separate proposition although these auxiliaries only mark syntax, thus leading to an artificially high proposition count. For instance, the clauses “she is reaching” and “he is taking” both contain two propositions according to CPIDR, whereas they both really contain only one semantic idea.

Both CPIDR and DEPID PID values differ significantly between AD and control groups on DementiaBank but the mean values are opposite to what was expected—AD patients have significantly higher PID than controls. When the repeated ideas are not counted (DEPID-R), the difference between groups becomes non-significant. However, we were curious about why the association between the lower PID values and the AD diagnosis cannot be observed on DementiaBank. Thus, we investigated this issue and found that the DementiaBank

interviews have certain additional characteristics that contribute to the automatic proposition count being too high.

Conjunctive propositions First, we noticed that most *and-conjunctions* are used as lexical fillers in DementiaBank, whereas both CPIDR and DEPID count all conjunctions as propositions. In order to address this problem we excluded the *cc* dependency type from the set of propositions.

Sentences with pronominal subjects Secondly, we noticed that the sentences with subject either *I* or *you* most probably do not say anything about the picture but rather belong to the meta conversation. Two examples of such sentences are for instance “*what else can I tell you about the picture?*” or “*I’d say that’s about all.*”. To solve this problem we did not count propositions from sentences, where the subject was either *I* or *you*.

Vague sentences Finally, we observed that the AD patients seem to utter more *vague sentences* that do not contain any concrete ideas, such as for instance “*the upper one is there*” or “*they’re doing more things on the outside.*”. Both CPIDR and DEPID extract propositions from syntactic structures and thus they count pseudo-ideas from those sentences as well. To detect such vague sentences we evaluated the specificity of all sentences using SpeciTeller (Li and Nenkova, 2015). SpeciTeller predicts a specificity score between 0 and 1 for each sentence using features extracted from the sentence surface-level, specific dictionaries and distributional word embeddings. We did not count propositions from sentences whose specificity score was lower than 0.01.

After incorporating all those three measures to DEPID we finally obtain PID values on DementiaBank that are significantly different for patients and controls in the expected direction—the AD patients have significantly lower PID values than control subjects. Note that those measures only affect the proposition count and not the number of tokens. Also note that although these measures were motivated by the observations made on one particular (DementiaBank) dataset, they can be expected to be applicable to other similar closed-topic datasets, containing picture descriptions or story re-tellings.⁴

⁴Unfortunately, aside from DementiaBank there are no other publicly available AD datasets and thus we could not test whether our expectations hold true.

On AMI data, the difference between group means is non-significant for both CPIDR and DEP-ID values. However, when the repeated ideas are excluded (DEPID-R), the mean PID for AD patients is significantly lower than for controls, as expected. It should be noted that the first two problems observed on DementiaBank—conjunctions and pronominal subjects—are not actual on the free-recall AMI data. In autobiographical memory interviews many sentences are expected to have *I* as subject. Also, the *and*-conjunctions are more likely to convey real ideas there rather than carry the role of lexical fillers. However, AD patients can utter more sentences with very vague meaning in AMI data as well and thus, in the last row of the Table 5 we show the DEP-ID PID values with vague sentences excluded for AMI dataset as well. We see that the PID values decrease for both patients and controls and the difference between groups remains statistically significant.

SID values differ significantly between the AD and control groups on both datasets with AD patients having significantly lower SID values as expected. The clustering underlying the automatically computed SID is trained on the whole dataset for both DementiaBank and AMI data.

5.3 Classification setup

We test both PID and SID in the diagnostic binary classification task on both DementiaBank and AMI datasets. When computing PID, the repeated ideas are excluded (DEPID-R). In addition, for DementiaBank, we also use the additional measures described in Section 5.2 (DEPID-R-ADD) as, according to Table 5, just DEP-ID-R cannot be expected to be predictive on that type of dataset. We compute the SID as described in Section 4. In following (Yancheva and Rudzicz, 2016), we cluster the 50-dimensional Glove embeddings⁵ of all nouns and verbs found in the transcripts with k-means. Similar to them, we set the number of clusters to 10 on both datasets.

For single feature models (SID or PID) we use a simple logistic regression classifier. For models with multiple features we use the elastic net logistic regression with an elastic net hyperparameter $\alpha = 0.5$. We train and test with 10-fold cross-validation on subjects and repeat each experiment 100 times. We report the mean and stan-

⁵<http://nlp.stanford.edu/projects/glove/>

Data	Features	Precision	Recall	F-score
DB	CPIDR	59.8 (0.7)	59.1 (0.5)	58.8 (0.5)
DB	PID	61.1 (0.7)	60.3 (0.6)	60.0 (0.5)
DB	SID	71.4 (0.6)	70.7 (0.5)	70.5 (0.5)
DB	SID+PID	73.7 (0.9)	72.1 (0.6)	72.2 (0.6)
AMI	CPIDR	45.1 (3.2)	63.4 (1.8)	51.9 (2.3)
AMI	PID	79.2 (1.9)	80.0 (0.5)	77.6 (0.9)
AMI	SID	73.7 (3.0)	75.3 (1.5)	72.3 (2.1)
AMI	SID+PID	82.9 (3.8)	78.0 (1.8)	77.7 (1.8)

Table 6: Classification results of various ID measures. The PID is DEP-ID-R-ADD for DementiaBank and DEP-ID-R for AMI.

dard deviation of the 100 macro-averaged cross-validated runs. For each experiment we report class-weighted precision, recall and F-score.⁶

5.4 Classification results

The classification results using various ID measures are shown in Table 6. On both datasets, PID and SID are better from the CPIDR baseline although the difference is considerably larger on the free-recall AMI dataset. On DementiaBank, SID performs better than PID and combining SID and PID also gives a small consistent cumulative effect, improving the F-score by 1.7%. On AMI data, the SID performs surprisingly well, considering that the automatic ICUs were extracted from only 10 clusters and the number of clusters was not tuned to that dataset at all. However, PID performs ca 5% better than SID in terms of all measures. Combining PID and SID gives some improvements in precision at the cost the decrease in recall and gives no cumulative gains in F-score. These results are fully in line with our expectations that the syntax-based DEP-ID performs better on the free-topic dataset, while the SID is better on closed-domain dataset.

For better comparison with Yancheva and Rudzicz (2016) we also experimented with the distance-based cluster features, which are derived from the clusters underlying the automatic SID (see section 4). We also show additional semantic baselines using LIWC features (Tausczik and Pennebaker, 2010) and bag-of-word (BOW) features extracting the counts of nouns and verbs normalised by the number of tokens. These results are shown in Table 7. On DementiaBank dataset, cluster features alone do not perform too well and using cluster features together with PID and SID gives only minor improvements. On the other hand, both the

⁶Classification accuracy is omitted because it is equivalent to the class-weighted recall.

Data	Features	Precision	Recall	F-score
DB	Clusters	62.3 (1.6)	62.2 (1.7)	62.2 (1.7)
DB	C+PID	67.4 (1.7)	64.9 (1.5)	65.1 (1.5)
DB	C+SID	73.4 (1.4)	71.5 (1.3)	71.6 (1.3)
DB	C+SID+PID	74.4 (1.5)	72.5 (1.2)	72.7 (1.2)
DB	LIWC	80.0 (0.9)	78.4 (0.7)	78.5 (0.7)
DB	BOW	80.6 (1.1)	79.1 (1.0)	79.3 (1.0)
AMI	Clusters	76.9 (7.7)	71.2 (5.2)	70.5 (5.8)
AMI	C+PID	81.2 (5.0)	75.7 (3.8)	75.3 (3.8)
AMI	C+SID	83.5 (5.0)	77.9 (4.1)	77.7 (4.4)
AMI	C+SID+PID	84.6 (4.4)	78.1 (3.8)	78.4 (4.0)
AMI	LIWC	74.2 (4.7)	67.8 (3.5)	66.8 (3.3)
AMI	BOW	65.1 (7.2)	65.3 (4.1)	61.6 (4.7)

Table 7: Classification results on DementiaBank (DB) and AMI using cluster features (C) combined with PID and SID, and LIWC and BOW baselines. The PID is DEPID-R-ADD for DementiaBank and DEPID-R for AMI.

Data	Features	Precision	Recall	F-score
DB	Clusters	68.0 (1.2)	65.5 (0.9)	65.7 (0.8)
DB	C+PID	69.6 (1.1)	67.1 (0.7)	67.4 (0.7)
DB	C+SID	75.3 (1.0)	73.3 (0.7)	73.5 (0.7)
DB	C+SID+PID	76.6 (1.1)	74.8 (0.8)	75.0 (0.7)
AMI	Clusters	86.0 (3.6)	80.4 (2.2)	80.5 (2.1)
AMI	C+PID	88.4 (3.9)	83.0 (2.7)	83.2 (2.8)
AMI	C+SID	88.6 (3.0)	84.8 (1.7)	84.8 (1.7)
AMI	C+SID+PID	87.3 (3.8)	82.4 (2.6)	82.7 (2.7)

Table 8: Classification results on DementiaBank (DB) and AMI using cluster features (C) combined with PID and SID. The clusters are pre-trained on the whole dataset. The PID is DEPID-R-ADD for DementiaBank and DEPID-R for AMI.

LIWC and BOW baselines perform very well on DementiaBank with BOW features giving the total highest precision of 80.6%, recall of 79.1% and F-score of 79.3%. In fact, these results are very close to the state-of-the-art on this dataset: a recall of 81.9% (Fraser et al., 2015) and an F-score of 80.0% (Yancheva and Rudzicz, 2016). Note however that the BOW features are conceptually much simpler than the acoustic and lexicosyntactic features extracted by Yancheva and Rudzicz (2016) and Fraser et al. (2015).

On the free-recall AMI data, the cluster features perform surprisingly well while the results of the LIWC and BOW baselines are lower. Adding cluster features to ID behaves inconsistently—in case of SID the F-score improves while adding cluster features to PID lowers the F-score. It is also worth noticing that results on AMI data including cluster features vary quite a bit, in some cases having standard deviation even as high as 7.7%.

Finally, we experimented with a scenario where

the word embedding clusters are pre-trained on the whole dataset, in which case the clustering and thus also the SID feature reflect the structure of both training and test folds. This scenario assumes re-training the clustering and the classification model for each new test item/set. Although the classification model is then informed by the test set, we do not see it as test set leakage as the clustering is unsupervised. These results, given in Table 8, show that all results on both datasets improve, whereas the improvements are considerably larger on AMI dataset, which is expected because the model trained on the free-topic AMI data is likely to gain more on knowing the topics discussed in the test item/set. This scenario gives the highest F-score of 84.8% on this dataset when adding cluster features to SID.

Note, that the cluster features F-score trained on the full dataset is slightly lower than the 68% reported by Yancheva and Rudzicz (2016). This difference is probably due to the differences in hyperparameters and experimental setup: we use an elastic-net regularised logistic regression classifier while they used a random forest, we perform 10-fold cross-validation while they divided the DementiaBank into 60-20-20 train-dev-test partitions. However, the classification performance of cluster features together with SID are in the same range as their reported 74%.

6 Discussion

This is the first work we are aware of that compares the same methods for predicting AD on two different datasets. Moreover, most previous work has been conducted either on constrained-topic datasets, containing picture descriptions (Orimaye et al., 2014; Fraser et al., 2015; Yancheva and Rudzicz, 2016; Rentoumi et al., 2014), or semi-constrained structured interviews about some particular topic (Thomas et al., 2005; Jarrold et al., 2010, 2014), while our AMI dataset contains free recall samples and thus is probably more spontaneous than the previously used datasets.

We expected PID to perform well on the free-recall AMI dataset, which proved to be the case. However, we were surprised that the small number of automatically extracted clusters perform so well on that dataset too. This raises the natural question what topics those clusters represent. To shed light on this question, we studied the clustering trained on the whole AMI dataset. There were three clusters for which values differed sig-

nificantly⁷ between AD and control subjects: C0 ($p < 0.001$), C6 ($p < 0.001$) and C9 ($p = 0.0044$). C0, which could be denoted as a cluster describing experiences, contained a diverse mix of words, which close to the cluster center denoted specific aspects of something or connoted emotions such as “rudeness”, “flirting” and “usher”, while the farther words contained a range of aspects relevant to people’s lives such as “billiards”, “bronchitis” and “depression”. C6 contained close to the cluster center simple work-related words, e.g. “working”, “employed” and “student”, while farther from the center there were more words referring to family members and even further away became the words referring to specific professions such as “psychologists”, “barrister” and “chemist”. The values of C6 feature for AD patients were significantly lower than for controls. Finally, the cluster C9 contained simple business-related words close to the cluster center, such as “manage”, “product” and “account”, while the words got more specific farther away from the centroid, e.g. “licensed”, “reorganisation” and “textile”.

Also, we checked how many words were considered as ICUs (words with $d_{scaled} < 3.0$ to their closest cluster center) on AMI data and found that most words were counted. This suggests that the automatically computed SID is in fact very close to the simple proportion of nouns and verbs in the transcripts. In order to check this, we extracted the normalised counts of nouns and verbs from all transcripts in both datasets and used it to train single feature logistic regression classifiers. We obtained the precision 67.6, recall 66.8 and F-score 66.6 on DementiaBank and precision 77.1, recall 76.0 and F-score 74.3 on AMI dataset. Also, we found that on DementiaBank the simple bag-of-words baseline obtained the results very close to the current state-of-the-art that uses much more complex feature sets, including both acoustic and lexicosyntactic features (Fraser et al., 2015). These two observations suggest that there is still room for studying simple feature sets for predicting AD.

7 Conclusion

We experimented with two different definitions of idea density—propositional idea density and semantic idea density—in the classification task for predicting Alzheimer’s disease. In the AD and psycholinguistic literature, PID has been automatically

calculated using CPIDR (Engelman et al., 2010; Ferguson et al., 2014; Bryant et al., 2013; Moe et al., 2016). We show that CPIDR has a number of flaws when applied to AD speech, and we propose a new PID computation method DEPID which is more highly correlated with manual estimates of PID. We recommend that AD researchers use our automatic measure, DEPID-R, which also excludes repeating ideas from the total idea count, in place of CPIDR.

This is the first comparison between PID and SID and also the first computational study that evaluates the predictive models for Alzheimer’s disease on two very different datasets. While on the closed-topic picture description dataset SID performs better, including PID also adds a small improvement to the classification results. On the open-domain dataset we found that the PID was more predictive than SID as expected. However, the small number of automatically extracted cluster features underlying the SID, modeling the broad discussion topics, led to even better results.

In future we plan to study the usefulness and applicability of both PID and SID also in other clinical tasks, such as in clinical diagnostic tasks for depression or schizophrenia. Another possible avenue for future work would include combining dependency-base PID and embedding-based SID into a unified idea density measure that would take into account both the propositional structure as well as the semantic content of words.

Acknowledgements

This research was supported by a Google award through the Natural Language Understanding Focused Program, and under the Australian Research Council’s Discovery Projects funding scheme (project number DP160102156), and in part by funding to ForeFront, a collaborative research group dedicated to the study of frontotemporal dementia and motor neuron disease, from the National Health and Medical Research Council (NHMRC) (APP1037746), and the Australian Research Council (ARC) Centre of Excellence in Cognition and its Disorders Memory Program (CE11000102). OP is supported by an NHMRC Senior Research Fellowship (APP1103258).

References

Samrah Ahmed, Celeste A. de Jager, Anne-Marie Haigh, and Peter Garrard. 2013a. Semantic process-

⁷We used the Wilcoxon signed rank test.

- ing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology* 27(1):79–85.
- Samrah Ahmed, Anne-Marie F. Haigh, Celeste A. de Jager, and Peter Garrard. 2013b. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain* 136(12):3727–3737.
- Kathryn A. Bayles, Cheryl K. Tomoeda, Alfred W. Kaszniak, Lawrence Z. Stern, and Karen K. Eigans. 1985. Verbal perseveration of dementia patients. *Brain and Language* 25(1):102–116.
- Kathryn A. Bayles, Cheryl K. Tomoeda, Patrick E. McKnight, Nancy Helm-Estabrooks, and Josh N. Hawley. 2004. Verbal perseveration in individuals with Alzheimer's disease. *Seminars in Speech and Language* 25(4):335–347.
- James T. Becker, Francois Boller, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis. *Archives of Neurology* 51(6):585–594.
- Cati Brown, Tony Snodgrass, Susan J. Kemper, Ruth Herman, and Michael A. Covington. 2008. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods* 40(2):540–545.
- Lucy Bryant, Elizabeth Spencer, Alison Ferguson, Hugh Craig, Kim Colyvas, and Linda Worrall. 2013. Propositional Idea Density in aphasic discourse. *Aphasiology* 27(8):992–1009.
- Vineeta Chand, Kathleen Baynes, Lisa M. Bonnici, and Sarah Tomaszewski Farias. 2010. Analysis of Idea Density (AID): A Manual. Technical report, University of California at Davis.
- Vineeta Chand, Kathleen Baynes, Lisa M. Bonnici, and Sarah Tomaszewski Farias. 2012. A rubric for extracting idea density from oral language samples. *Current Protocols in Neuroscience* 1.
- Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford Dependencies manual. Technical report, Stanford University.
- Michal Engelman, Emily M. Agree, Lucy A. Meoni, and Michael J. Klag. 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.
- Alison Ferguson, Elizabeth Spencer, Hugh Craig, and Kim Colyvas. 2014. Propositional idea density in women's written language over the lifespan: computerized analysis. *Cortex* 55:107–121.
- Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2015. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's disease* 49(2):407–422.
- Harold Goodglass and Edith Kaplan. 1983. *The Assessment of Aphasia and Related Disorders*. Lea & Febiger.
- William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 27–37.
- William L. Jarrold, Bart Peintner, Eric Yeh, Ruth Krasnow, Harold S. Javitz, and Gary E. Swan. 2010. Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic alzheimer's disease. In *Proceedings of the 2010 International Conference on Brain Informatics*. pages 299–307.
- Susan Kemper, Janet Marquis, and Marilyn Thompson. 2001. Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging* 16(4):600–614.
- Walter Kintsch and Janice Keenan. 1973. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology* 5(3):257–274.
- Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. pages 2281–2287.
- Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk, 3rd edition..* Lawrence Erlbaum Associates.
- Aubrey M. Moe, Nicholas J. K. Breitborde, Mohammed K. Shakeel, Colin J. Gallagher, and Nancy M. Docherty. 2016. Idea density in the life-stories of people with schizophrenia: Associations with narrative qualities and psychiatric symptoms. *Schizophrenia Research* 172(1):201–205.
- Marjorie Nicholas, Loraine K. Obler, Martin L. Albert, and Nancy Helm-Estabrooks. 1985. Empty Speech in Alzheimer's Disease and Fluent Aphasia. *Journal of Speech and Hearing Research* 28(3):405–410.
- Sylvester O. Orimaye, Jojo Sze-Meng Wong, and Karen J. Golden. 2014. Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. pages 78–87.

- Vassiliki Rentoumi, Ladan Raoufian, Samrah Ahmed, Celeste A. de Jager, and Peter Garrard. 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):3–17.
- Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *IEEE transactions on audio, speech, and language processing* 19(7):2081–2090.
- David A. Snowdon, Susan J. Kemper, James A. Mortimer, Lydia H. Greiner, David R. Wekstein, and William R. Markesbery. 1996. Linguistic ability in early life and cognitive function and Alzheimer’s disease in late life. Findings from the Nun Study. *JAMA* 275(7):528–532.
- Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29(1):24–54.
- Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *IEEE International Conference Mechatronics and Automation, 2005*. pages 1569–1574.
- Cheryl K. Tomoeda, Kathryn A. Bayles, Michael W. Trosset, Tamiko Azuma, and Anna McGeagh. 1996. Cross-sectional analysis of Alzheimer disease effects on oral discourse in a picture description task. *Alzheimer Disease and Associated Disorders* 10(4):204–215.
- Althea Turner and Edith Greene. 1977. The construction and use of a propositional text base. Technical report, University of Colorado.
- Maria Yancheva and Frank Rudzicz. 2016. Vector-space topic models for detecting Alzheimer’s disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. pages 2337–2346.