Neuroscience and Biobehavioral Reviews 93 (2018) 85–92

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis

J.N. de Boer[a,*,1], A.E. Voppel[b,1], M.J.H. Begemann[b], H.G. Schnack[a,c], F. Wijnen[c], I.E.C. Sommer[b,d]

[a] Department of Psychiatry, University Medical Center Utrecht, Utrecht University & Brain Center Rudolf Magnus, Utrecht, The Netherlands
[b] Department of Neuroscience and Department of Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[c] Utrecht Institute of Linguistics OTS, Faculty of Humanities, Utrecht University, Utrecht, The Netherlands
[d] Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway

ARTICLE INFO

Keywords:
Natural language processing
Vector space
Semantic space
Neurology
Psychiatry

ABSTRACT

Verbal communication disorders are a hallmark of many neurological and psychiatric illnesses. Recent developments in computational analysis provide objective characterizations of these language abnormalities. We conducted a meta-analysis assessing semantic space models as a diagnostic or prognostic tool in psychiatric or neurological disorders. Diagnostic test accuracy analyses revealed reasonable sensitivity and specificity and high overall efficacy in differentiating between patients and controls ($n = 1680$: Hedges' $g = .73$, $p = .001$). Analyses of full sentences (Hedges' $g = .95$ $p < .0001$) revealed a higher efficacy than single words (Hedges' $g = .51$, $p < .0001$). Specifically, models examining psychotic patients (Hedges' $g = .96$, $p = .003$) and those with autism (Hedges' $g = .84$, $p < .0001$) were highly effective. Our results show semantic space models are effective as a diagnostic tool in a variety of psychiatric and neurological disorders. The field is still exploratory in nature; techniques differ and models are only used to distinguish patients from healthy controls so far. Future research should aim to distinguish between disorders and perhaps explore newer semantic space tools like word2vec.

1. Introduction

Language is an essential anamnestic source of information in psychiatry and neurology, since it is an expression of thought and therefore provides a window into the mind (Pinker, 2007). More importantly, disturbed language use holds clue as to what it is that goes astray in the brain. Disorganized language is a core criterion for schizophrenia (American Psychiatric Association, 2013) and can also be seen as a symptom of Alzheimer's disease (Appell et al., 1982). Disturbances in natural language are also seen in several other mental illnesses, including depression, bipolar disorder, autism and personality disorders (Cohen and Elvevåg, 2014). Motor disturbances in speaking are a long standing finding in neurodegenerative disorders such as Parkinson's disease (Logemann et al., 1978), although impairments in high-level language function such as sentence and discourse production also occur in Parkinson's patients (Altmann and Troche, 2011).

While a discussion about whether language should be seen as a module of cognitive function, is beyond the scope of this paper (but see for instance Carruthers, 2002), it is clear that several important cognitive functions are involved in and necessary for successful verbal communication. A person at least needs *planning ability* to develop a plan of what to say, *inhibition* to stick to a chosen message, *perception* to understand an interlocutor and *memory* for the use of the lexicon (Carruthers, 2002; Emmorey, 2001). Since language involves so many complex cognitive functions, it seems plausible that a wide variety of disorders can lead to language disturbances.

However, language, (partly) due to its wide range of applications and nuances, is difficult to analyze and quantify (Hoffman et al., 2013). Recently, through the advent of natural language processing, it has become possible to study the language produced by an individual in an objective and quantifiable way. This creates the opportunity to use language as a marker for diagnosis, prognosis and perhaps even treatment response of a variety of brain disorders in which language disturbances are an early sign of the disorder or relapse of symptoms.

More specifically, analyses of patterns in word meaning seems most likely to serve as a diagnostic and prognostic tool in various brain disorders, because it is a very rich topic and therefore sensitive to subtle changes in produced language. In this review, we quantitatively summarize the recent literature regarding a specific method to objectively quantify disturbances in spoken language, namely semantic space models.

[*] Corresponding author at: Department of Psychiatry, University Medical Center Utrecht, Utrecht, 3508 GA, The Netherlands.
E-mail address: j.n.deboer-18@umcutrecht.nl (J.N. de Boer).
[1] Both authors contributed equally.

https://doi.org/10.1016/j.neubiorev.2018.06.008
Received 14 March 2018; Received in revised form 7 June 2018; Accepted 7 June 2018
Available online 08 June 2018

1.1. Semantic space models

There is a long history of representing the meaning of single words as continuous mathematical objects in a so called ‘semantic space’. Semantic space models attempt to represent word meanings by representing them as points in an abstract multidimensional space. An analogy can be drawn with psycholinguistic research, which has shown that semantically similar words cluster together in the brain cortex based on the context they occur in (Huth et al., 2016). The central assumption underlying all semantic space models is that the meaning of a word is determined by its context. One of the first to draw attention to the context-dependency of meaning was Firth, who famously said: “You shall know a word by the company it keeps” (Firth, 1957:11). Semantic space models additionally assume that words with similar meanings appear in similar contexts. In semantic space models, context is defined as the sets of words that precede and/or follow the target word.

The semantic space model used most often is Latent Semantic Analysis (LSA), a method for representing word meanings by means of statistical computations (Landauer and Dumais, 1997). LSA is based on singular value decomposition, which is a mathematical technique that is similar to factor analysis (Landauer et al., 1998). After processing large text corpora, LSA represents words as points in a high-dimensional semantic space where dimensions are based on the contexts in which words occur. More specifically, LSA takes raw text as its input, which is separated into meaningful units such as sentences or paragraphs. As a first step, LSA represents the text as a matrix in which each unique word is presented as a row and each column is a context in which it occurs. Each of the matrix cells contain the frequency with which the word occurs in the given context. Next, each of these frequencies is weighted by a process that expresses the word’s information value by calculating its entropy. Singular Value Decomposition is then applied to the matrix, collapsing the rows and columns to a multidimensional semantic space. The semantic similarity between two words can be quantified as a distance measure or an angle between two points, such as an Euclidean distance or a cosine angle respectively (Padó and Lapata, 2007). Several studies have shown that meaning representations that were derived by LSA are capable of simulating human cognitive phenomena (Landauer et al., 1998). It is important to note that LSA makes no use of word order and is thus incapable of extracting syntactic relations. By using LSA, it becomes possible to transform language to a quantitative vector on which further analyses can be performed. For a thorough introduction to LSA see Landauer et al. (1998).

1.2. Semantic space measures

As briefly outlined above, semantic similarity is often quantified as a cosine of the angle between two vectors (ranging from -1 to 1). Greater cosines (i.e. smaller angles) indicate greater semantic similarity. It depends both on the research question and the type of data, how the semantic similarity is calculated. For instance in a word association task, semantic similarity can be represented by calculating the cosine of the angle from the cue word to the response. In a verbal fluency task, similarity is often defined as the (average) cosine of the angle between one response to the next response (i.e. how similar all consecutive responses on a verbal fluency task are). This type of similarity can also be used to create derived measures such as ‘clustering’ or ‘switching’. Take for instance the animal verbal fluency task, where people have to name as many animals as possible within a given time frame. When people name the following animals “lion, elephant, zebra” cosines will be high, because responses all fit in the ‘African mammals’ cluster. If a person than *switches* to ‘goldfish’, the cosine between ‘zebra’ and ‘goldfish’ will be low, which can be used to define cluster size and switching.

Semantic similarity can also be calculated over larger chunks of words, such as sentences or utterances. In an interview setting for instance, average cosines can be calculated between the question of the interview and the corresponding answer of the interviewee. In a story retelling setting, similarity between the story produced by the subject and the original story can also be presented as an average cosine value.

Some of the earliest clinical uses of semantic space models came from research into psychosis, where the aim was to quantify the degree of incoherence observed in these patients’ language use. Elvevåg et al. (2007, 2010) first explored the use of LSA in schizophrenia patients, their family members and healthy controls. They found that these groups significantly differ in mean coherence (cosine) scores calculated with LSA as applied to their language output. By comparing the results to traditional rating scales to capture incoherent speech, the authors conclude that LSA is an effective tool to measure coherence of language in patients with schizophrenia. LSA was later used by Bedi et al. (2015) to predict conversion in youths at high risk for psychosis. By combining LSA with measures of syntactic complexity (number of determiners divided by phrase length), the authors were able to predict conversion from at-risk youth to psychosis with 100% accuracy. However, sample sizes were small and only five participants converted to psychosis.

Though automated semantic analysis appears to be a promising new technique, the field lacks an overview of its use as a biomarker for neuropsychiatric disorders. This study therefore aims to provide a qualitative and quantitative overview of the use of semantic space models as a diagnostic or prognostic tool in neurodegenerative and psychiatric populations.

2. Methods

2.1. Literature search

This meta-analysis/systematic review was performed according to the Preferred Reporting for Systematic Reviews and Meta-analysis (PRISMA) Statement (Moher et al., 2009). A systematic search was performed in the databases Pubmed (Medline), Embase, PsychInfo and Cochrane Database of Systematic Reviews (independently by J.B. and A.V). Combinations and synonyms of the following search terms were used: “semantic space” or “semantic vector” and “psychiatric disorder” or “neurological disorder”. See Supplementary Table 1 (Table S1) for a full search string. No year of publication or language filters were used. The search cut-off date was January 19th, 2018. Reference lists of the included studies were searched for cross-references. Authors were contacted in case data was unavailable. When the full text of articles was not available and the authors could not be traced, the abstract was used to extract necessary information to avoid publication bias. Studies that met the following inclusion criteria were included: 1) Studies investigating the clinical use of automated measures of semantic space. 2) Studies including patients with a neurological or psychiatric disorder and a comparison group without a clinical diagnosis. 3) Studies that applied semantic space calculations to any form of language produced by the patient and comparison groups (i.e. written or spoken language output).

Studies using semantic space calculations in addition to other diagnostic measures were included to provide a thorough review of the available literature. Studies were excluded if the language that was analyzed was not produced by a clinically diagnosed group, including for instance electronic patient databases. Methodological quality of the studies was assessed independently by J.B. and A.V., by means of the Quality Assessment of Diagnostic Studies-2 checklist (Whiting et al., 2011), which was adapted to fit the purpose of this study, see Table S2.

2.2. Outcome measures

The primary outcome measures were sensitivity and specificity of the model, since these values are most informative for diagnostic test accuracy. If these were unavailable, mean scores on the semantic space measure per group, or the statistical significance of the semantic measure difference between groups were used to calculate effect sizes.

Table 1
Main characteristics of studies included in quantitative assessment.

			Diagnostic groups		Mean age (SD)		Linguistic data characteristics			Semantic space characteristics		Included in analyses	
	Condition	*n*	Patient	Comp.	Patient	Comp.	Elicitation method	Output type	Length	Training corpus	Tool	DTA	Efficacy
Bedi et al. (2015)	Psychosis	34	CHR - Con	CHR - Non	22.2 ± 3.4	21.2 ± 3.6	Narrative interview	Sentences	1 hr	TASA	LSA	+	+
Cecchi (2016)	Psychosis	59	CHR - Con	CHR - Non	N/R	N/R	Prose recall	Sentences	N/R	N/R	LSA	+	+
Clark et al. (2016)	Dementia	107	MCI - Con	MCI - Non	73.8 ± 7.9	68.7 ± 8.6	SVF, PVF	Single words	5 x 1 min.	Google n-grams	Other	+	–
Corcoran et al. (2018)	Psychosis	34	CHR - Con	CHR - Non	22.2 ± 3.4	21.2 ± 3.6	Narrative interview	Sentences	1 hr	TASA	LSA	–	–
Elvevåg et al. (2007)	Psychosis	51	Schizophrenia	HC	33.8 ± 7.63	35.4 ± 12.94	Story telling	Sentences	N/R	LSA 300 dimens.	LSA	–	–
Elvevåg et al. (2010)	Psychosis	83	Schizophrenia	HC	N/R	N/R	Narrative interview	Sentences	600-1000 wrds	N/R	LSA	+	–
García et al. (2016)	Parkinson	101	Parkinson's disease	HC	61.5 ± 9.77	60.9 ± 9.47	Narrative interview	Sentences	100 wrds	TASA	LSA	+	+
Hoffman et al. (2014)	Dementia	15	Semantic dementia	HC	60 ± 4.9	N/R	Narrative interview	Sentences	4000-5000 wrds	Lexesp	LSA	–	+
Lee et al. (2017)	ASD	33	ASD	HC	24.2 ± 9.48	19.11 ± 2.20	Describing scenes	Sentences	N/R	Subtitles	LSA	–	+
Losh and Gordon (2014)	ASD	48	ASD	HC	N/R	N/R	Prose recall	Sentences	N/R	N/R	LSA	–	+
Luo et al. (2016)	ASD	159	ASD	HC	34.8 ± 15.9	24.5	Narrative interview	Sentences	84 wrds	Own data	LSI	+	+
Nicodemus et al. (2014)	Psychosis	665	Schizophrenia	HC + healthy family	N/R	N/R	SVF	Single words	1 min.	TASA	LSA	–	+
Pakhomov et al. (2012)	Dementia	86	Alzheimer's disease	MCI	78.2 ± 6.87	76.00 ± 7.33	SVF	Single words	1 min.	WordNet	Gloss vectors	–	+
Pakhomov and Hemmy (2014)	Dementia	239	Alzheimer's disease	Elderly nuns	80.7 ± 3.98		SVF	Single words	1 min.	Wikipedia	LSA	–	+
Pakhomov et al. (2015)[1]- AD	Dementia	27	Alzheimer's disease	HC	69 ± 11	55 ± 13	SVF	Single words	1 min.	Wikipedia	LSA	–	+
Pakhomov et al. (2015)[1]- MCI	Dementia	22	MCI	HC	69 ± 12	55 ± 13	SVF	Single words	1 min.	Wikipedia	LSA	–	+
Prud'hommeaux and Roark (2015)	Dementia	235	ASD	HC	N/R	N/R	Prose recall	Single words	1 min.	DementiaBank,	LSA	–	–
Prud'hommeaux et al. (2017)	ASD	44	ASD	HC	N/R	N/R	SVF	Single words	1 min.	Wikepedia	LSA, word2vec	–	+
Rosenstein et al. (2014)	Psychosis	122	Schizophrenia	HC + healthy family	30.8 ± 9.19	32.5	Prose recall	Sentences	N/R	TASA	LSA	+	–
Rouhizadeh et al. (2015)	ASD	69	Schizophrenia	HC	6.41	5.91	Narrative interview	Sentences	30-60 min.	Own data	Other	–	+
Tagamets et al. (2014)	Psychosis	22	ADHD	HC	40	47	Narrative interview	Sentences	150 wrds	N/R	LSA	–	+
White and Shah (2016)	ADHD	60	CHR - Con	CHR - Non	19.5	20.2	Word association	Single words	25 wrds	N/R	LSA	–	+

Table legend : *n* = number of subjects, HC = healthy controls, CHR = clinical high risk for psychosis, Con = converters, Non = non converters, MCI = Mild Cognitive Impairment, PTSD = Post Traumatic Stress Disorder, ASS = Autism Spectrum Disorders, AD = Alzheimer's Disease, ADHD = Attention Deficit Hyperactivity Disorder, hr = hour, min. = minute, SVF = semantic verbal fluency, PVF = phonological verbal fluency, TASA = Touchstone Applied Science Associates, LSA = Latent Semantic Analysis, LSI = Latent Semantic Indexing, SVD = Single Value Decomposition, N/R = not reported, N/A = not applicable, dimens. = dimensions.
[1]Pakhomov et al. (2015) included different comparison groups or study designs, results are therefore presented separately. If several outcome measures were used for model prediction, the model with highest performance is reported. + = included in quantitative analyses, – = not included in quantitative analyses.

Direction of the effect size depends greatly on the language production task since more similar language is not per definition a positive feature. In spontaneous speech settings, interview settings or story retelling settings (i.e. *full sentences*), the effect size direction was considered positive if controls were more similar or coherent (i.e. higher cosine angles) in their language compared to the clinically diagnosed group. Only tasks that measured mental flexibility, higher similarity in patients was considered a positive effect size. In verbal fluency tasks, effect size was considered positive if controls showed lower similarity and thus more clustering or switching between words. These effect directions are based on the expectation that the switching between clusters reflects mental flexibility, which is considered to be a positive function. In word association tasks the effect size was positive if controls produced more similar words than patients.

2.3. Statistical analyses

2.3.1. Diagnostic test accuracy

Meta-Disc software version 1.4 (Zamora et al., 2006) was applied to obtain pooled sensitivity and specificity, using the number of true positives, true negatives, false positives and false negatives. To evaluate whether studies could be combined to calculate a pooled diagnostic accuracy, the Q-value and I^2-statistic were calculated for each analysis. The Q-statistic tests the heterogeneity of the studies and displays a distribution with k-1 degrees of freedom (k = number of studies). A Q-value higher than the degrees of freedom indicates significant between-studies variability. I^2 reflects the inconsistency between studies, ranging from 0 to 100%. According to Higgins et al. (2003), values of 25%, 50% and 75% can be interpreted as low, moderate and high inconsistencies, respectively. A random effects model was used because of the variation in population, language elicitation method, method of calculating semantic space and the limited number of studies for some diagnostic groups. A summary receiver operating characteristic (SROC) curve was calculated, from which we determined the area under the curve (AUC) which is robust against study heterogeneity and is considered a useful summary of the SROC (Walter, 2002). An AUC of 1 represents a perfect diagnostic test, while an area of .5 is at chance level. The desired AUC greatly depends on the gold standard in diagnosing the clinical condition under investigation; therefore there are no standard guidelines for interpretation of AUC in terms of high, medium or low accuracy (Hajian-Tilaki, 2013).

2.3.2. Effect size

Comprehensive Meta-Analysis (CMA) software version 2.0 was used to perform effect size analyses, using a random-effects model (Borenstein et al., 2005). For every individual study, Hedges' g was calculated for each outcome measure based on mean scores on the semantic space measure per group, using a random effects model. When these values were not reported, exact F-, t- or p-values were used. All effect sizes were calculated twice independently to check for errors. The Q-value and I^2-statistic were calculated for each analysis to evaluate to what extend studies could be taken together to share a common population effect size. Funnel plots were inspected for symmetry in order to check for publication bias and outliers. Potential asymmetry was assessed with Egger's test, using a significance level of $\alpha = .05$ (2-tailed). All effect sizes with a p-value of .05 or smaller were considered statistically significant. Effect sizes were interpreted according to the guidelines by Cohen, with an effect size of 0.20 indicating a small effect, 0.50 a medium and over 0.80 a large effect (Cohen, 1988).

3. Results

The search yielded a total of twenty-one relevant articles investigating the use of semantic space models in a clinical population, see Fig. S1 for a flow-diagram of the search (Bedi et al., 2015; Cecchi, 2016; Clark et al., 2016; Corcoran et al., 2018; Elvevåg et al., 2010, 2007; García et al., 2016; Hoffman et al., 2014; Lee et al., 2017; Losh and Gordon, 2014; Luo et al., 2016; Nicodemus et al., 2014; Pakhomov et al., 2015, 2012; Pakhomov and Hemmy, 2014; Prud'hommeaux et al., 2017; Prud'hommeaux and Roark, 2015; Rosenstein et al., 2014; Rouhizadeh et al., 2015; Tagamets et al., 2014; White and Shah, 2016). From these, eighteen studies were suitable to be included in the meta-analysis, investigating a total of 1995 participants (see Table 1) (Bedi et al., 2015; Cecchi, 2016; Clark et al., 2016; Elvevåg et al., 2010; García et al., 2016; Hoffman et al., 2014; Lee et al., 2017; Losh and Gordon, 2014; Luo et al., 2016; Nicodemus et al., 2014; Pakhomov et al., 2015, 2012; Pakhomov and Hemmy, 2014; Prud'hommeaux et al., 2017; Rosenstein et al., 2014; Rouhizadeh et al., 2015; Tagamets et al., 2014; White and Shah, 2016). These eighteen studies included the following brain disorders: psychosis spectrum disorders, autism spectrum disorders (ASD), Attention Deficit Hyperactivity Disorder (ADHD), dementia and Parkinson's disease.

3.1. Diagnostic test accuracy

A total of seven studies provided sufficient data to be included in a meta-analysis of diagnostic test accuracy (Bedi et al., 2015; Cecchi, 2016; Clark et al., 2016; Elvevåg et al., 2010; García et al., 2016; Luo et al., 2016; Rosenstein et al., 2014). Heterogeneity and I^2 for diagnostic test accuracy measures were high. All studies were pooled, which yielded an overall sensitivity of 61.7% and a specificity of 85.6% and an AUC of the SROC of 0.830 see Figs. 1 and 2. The analyses were repeated in the subset of studies investigating patients with a psychosis (k = 4), which improved the overall sensitivity to 71% and the specificity to

		Sample size			
Study name	Patient group	Control	Patient	Sensitivity (95% CI)	Specificity (95% CI)
Bedi et al. 2015	psychosis	29	5	1.00 (0.48 - 1.00)	1.00 (0.88 - 1.00)
Cecchi et al. 2016	psychosis	40	19	0.84 (0.60 - 0.97)	0.85 (0.72 - 0.94)
Clark et al. 2016	dementia	83	24	0.71 (0.49 - 0.87)	0.88 (0.79 - 0.94)
Elvevag et al. 2010	psychosis	53	30	0.87 (0.75 - 0.95)	0.60 (0.41 - 0.77)
García et al. 2016	Parkinson	50	51	0.40 (0.28 - 0.53)	0.71 (0.53 - 0.85)
Luo et al. 2016	ASD	132	27	0.54 (0.33 - 0.73)	0.74 (0.65 - 0.81)
Rosenstein et al. 2014	psychosis	94	28	0.55 (0.43 - 0.66)	0.94 (0.90 - 0.96)
	Overall	**481**	**184**	**0.62 (0.56 - 0.67)**	**0.86 (0.83 - 0.88)**

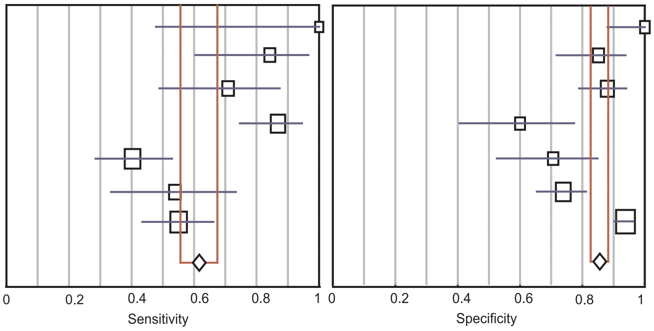


Fig. 1. Meta-analysis of the diagnostic test accuracy of semantic space models. Squares indicate individual study sensitivity and specificity respectively, which are scaled to study sample size. Blue bars indicate 95% confidence intervals per study. Diamonds indicate pooled sensitivity and specificity, with red bars indicating their 95% confidence interval. CI = confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

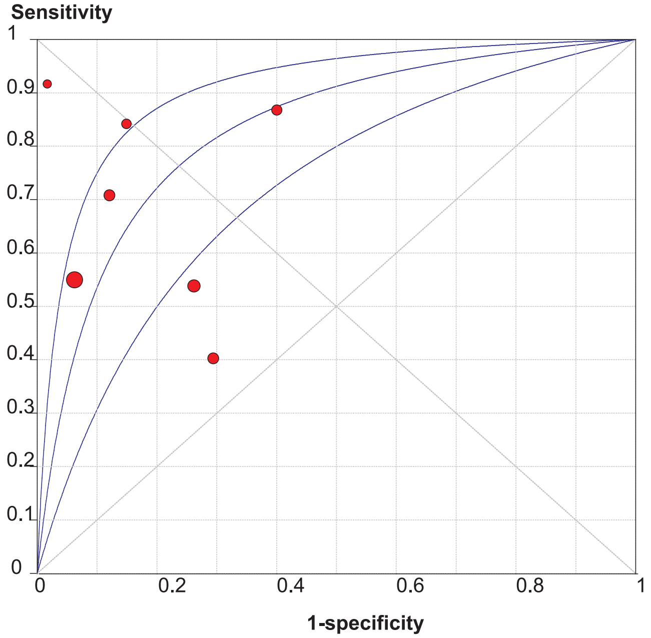


Fig. 2. Summary receiver operating characteristic (SROC) curve of the diagnostic test accuracy of semantic space models. Red dots indicate individual studies and are scaled to study sample size. Blue lines represent the SROC, with its 95% confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

91%, see Fig. S2 for the corresponding graphs.

3.2. Effect size

An overall cross-diagnostic meta-analysis of effect sizes showed a significant medium to large effect of semantic space models in separating patients from healthy controls (n = 1680; Hedges' g = .73, p = .001), see Table 2 and Fig. 3. The Q-value revealed slight but significant between-studies variability (Table 2). Eggers' Test was significant for the overall effect size, indicating the possibility of publication bias. Visual inspection of the funnel plots suggests a slight under publication of negative results.

All studies used spoken language of both diagnostic groups to calculate semantic space, though the method of language elicitation differed between studies. In general, studies could be divided into those that elicited single words (mainly word association task or verbal fluency tasks), and studies that elicited sentences (e.g. interview or story retelling). Therefore, sub-analyses were performed for studies that analyzed single words versus studies that analyzed sentences, see Fig. 4. Between group analyses revealed that the overall pooled effect size for sentences (Hedges' g = .95, p < .0001) was higher than the effect size for single words (Hedges' g = .51, p < .0001), p = 021.

Further subgroup analyses showed significant high effect sizes for separating patients with ASD (Hedges' g = .84, p < .0001) and patients with psychosis (Hedges' g = .96, p= .003) from healthy controls, while the positive combined effect size for dementia papers showed an overall significant but medium size (Hedges' g = 49, p= .012), see Table 2 and Fig. 2. None of the diagnostic subgroups showed a significant Q-value and I^2 was low, indicating low between-studies variability. Eggers' Test was not significant in any of the diagnostic subgroups, indicating no publication bias within diagnoses.

3.3. Qualitative assessment

Most studies (n=17) used LSA to calculate semantic space measures, see Table 1. Latent Semantic Indexing (LSI), Single Value Decomposition (SVD) and word2vec were also used. Training corpora varied between studies; Wikipedia texts and Touchstone Applied Science Associates (TASA) corpus were mostly used. Overall, there were roughly three different types of semantic space variables: clustering, coherence and similarity. Clustering analyses group words into different semantic clusters and measures that a person switches to a different cluster as well as the size of these clusters. Clustering was therefore mostly used in studies that analyzed verbal fluency data, while coherence and similarity were used more often in the full-sentence data studies.

4. Discussion

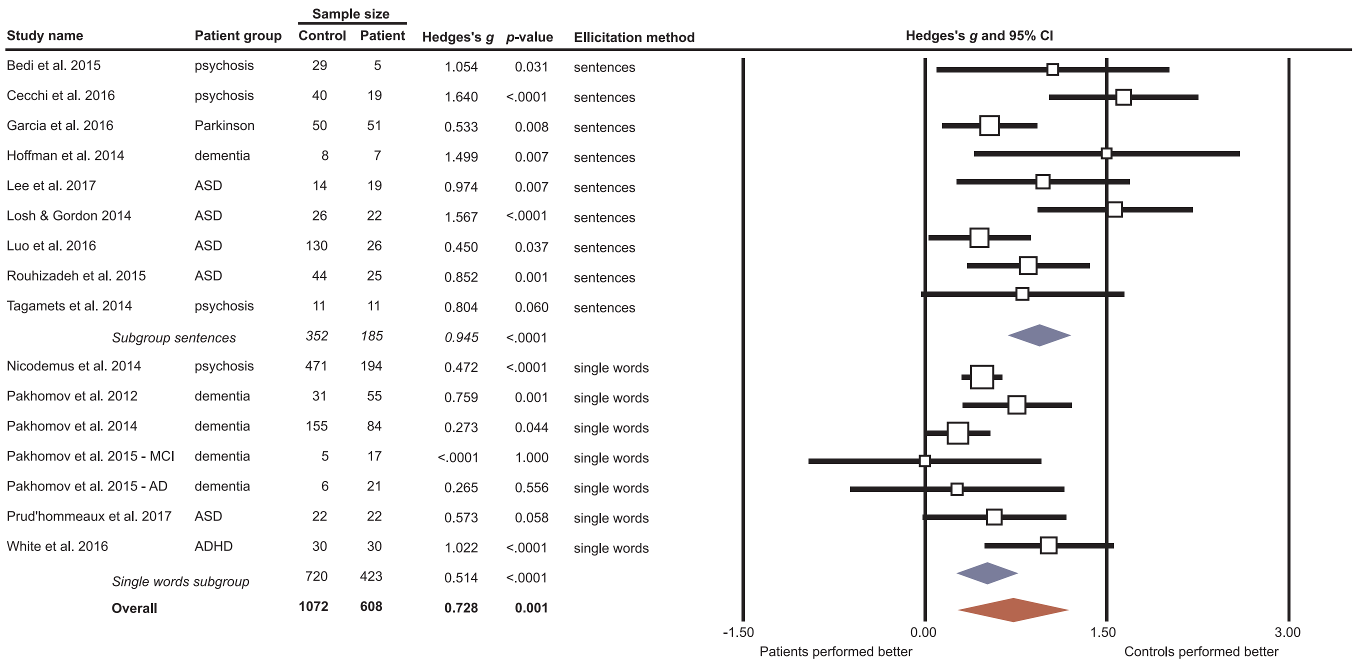
This study provides an overview of the clinical use of semantic space models in psychiatry and neurology populations. Our results indicate that these models have reasonable diagnostic accuracy, especially for diagnosing psychosis patients. We found a large overall effect size. Subanalyses revealed that models perform better on full sentences than single words as their language input. Best results were found in differentiating ASD patients and psychotic patients from healthy controls, though dementia patients also had a medium effect size. The wide range of disorders where we find positive significant effect sizes indicates the applicability of semantic space models in a wide range of disorders and pathologies.

It is important to note that semantic space models are not the only relevant linguistic marker for brain disorders. For linguists, the production of language occurs in the following stages (e.g. Dell et al., 1997): semantic (the generation of meaning), syntactic (the generation of the grammatical form of the utterance), morphological (the formation of words) and phonological (how these words should sound). Lastly, the speech is produced through articulation. In this review we only looked at a quite specific measure of semantics, while we expect broader aspects of language production to be affected as well. Indeed,

Table 2
Statistical results regarding all effect size outcome measures.

Outcome measure	Studies k	Subjects n	Hedges' g (95% CI)	p-value	I^2	Q-df	Q-value	Egger's test p-value
Diagnostic groups								
ASD	5	350	**.84 (.47 – 1.22)**	< .0001	0.97	4	4.04	.201
Dementia	5	389	**.49 (.11 - .88)**	.012	10.34	4	4.46	.518
Psychosis	4	780	**.96 (.32 – 1.60)**	.003	0.00	3	1.95	.231
Elicitation type								
Sentences	9	537	**.95 (.68– 1.21)**	< .0001	0.00	8	7.08	.007
Single words	7	1143	**.51 (.26 - .77)**	< .0001	6.754	6	6.44	.731
Overall	16	1680	**.73 (.31 – 1.15)**	.001	11.06	15	16.87	.026

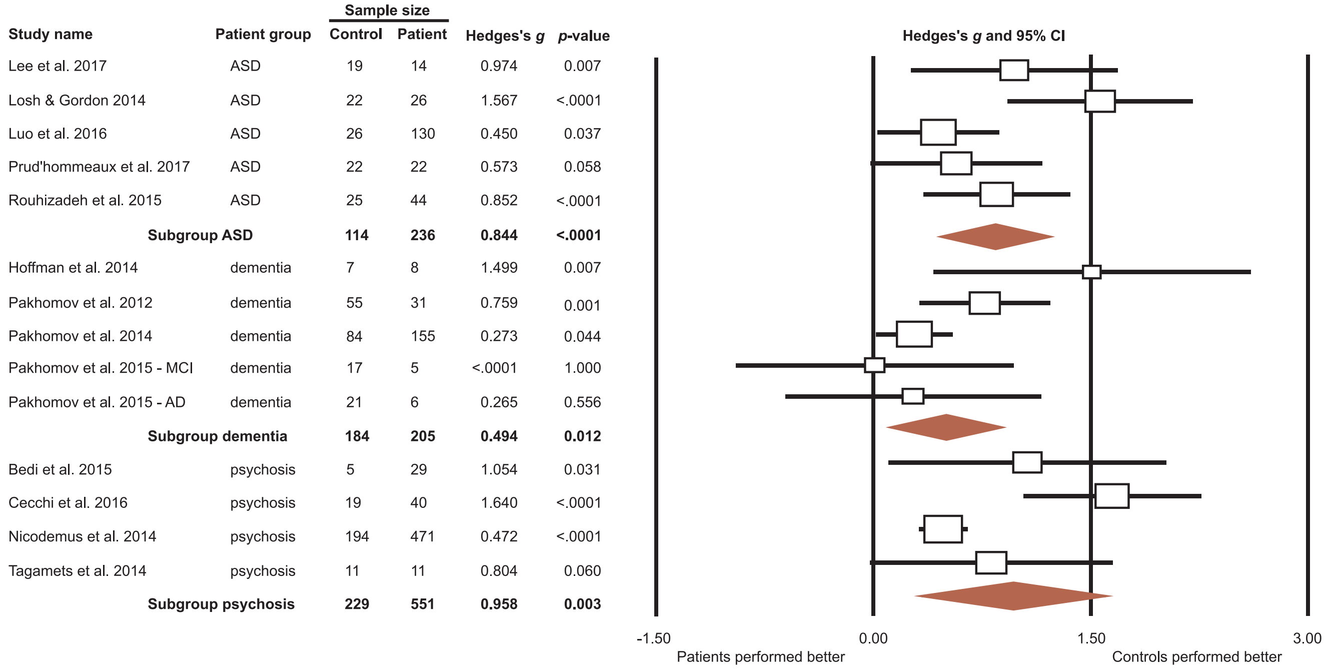
Table Legend : k = number of studies, n = number of subjects, CI = confidence interval, ASD = Autism Spectrum Disorders, df = degrees of freedom. Significant effect sizes in bold. Note: The article by Pakhomov et al. (2015) was divided into two different populations in the forest plot due to two different patient-control comparisons, and is therefore counted as two studies here.



Study name	Patient group	Sample size Control	Sample size Patient	Hedges's *g*	*p*-value	Ellicitation method
Bedi et al. 2015	psychosis	29	5	1.054	0.031	sentences
Cecchi et al. 2016	psychosis	40	19	1.640	<.0001	sentences
Garcia et al. 2016	Parkinson	50	51	0.533	0.008	sentences
Hoffman et al. 2014	dementia	8	7	1.499	0.007	sentences
Lee et al. 2017	ASD	14	19	0.974	0.007	sentences
Losh & Gordon 2014	ASD	26	22	1.567	<.0001	sentences
Luo et al. 2016	ASD	130	26	0.450	0.037	sentences
Rouhizadeh et al. 2015	ASD	44	25	0.852	0.001	sentences
Tagamets et al. 2014	psychosis	11	11	0.804	0.060	sentences
Subgroup sentences		*352*	*185*	*0.945*	<.0001	
Nicodemus et al. 2014	psychosis	471	194	0.472	<.0001	single words
Pakhomov et al. 2012	dementia	31	55	0.759	0.001	single words
Pakhomov et al. 2014	dementia	155	84	0.273	0.044	single words
Pakhomov et al. 2015 - MCI	dementia	5	17	<.0001	1.000	single words
Pakhomov et al. 2015 - AD	dementia	6	21	0.265	0.556	single words
Prud'hommeaux et al. 2017	ASD	22	22	0.573	0.058	single words
White et al. 2016	ADHD	30	30	1.022	<.0001	single words
Single words subgroup		720	423	0.514	<.0001	
Overall		**1072**	**608**	**0.728**	**0.001**	

Fig. 3. Meta-analysis of the group-comparison efficacy of semantic space models. Studies are grouped by language output type. Squares indicate Hedges' *g* for individual studies and are scaled to sample size. Black bars indicate 95% confidence intervals. Blue diamonds represent subgroup pooled effect sizes. The red diamond represents the overall pooled effect size. CI = confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

several meta-analyses have been published on other linguistic measures of verbal communication in a variety of brain disorders. For instance, measures of speed and volume of language production have been reviewed in schizophrenia (Cohen et al., 2014), acoustic markers such as formant analysis or spectrum analysis were performed in patients with mania (Zhang et al., 2018) and Parkinson's disease (Godino-Llorente et al., 2017) and measures of semantic and syntactic complexity (e.g. length of utterance, number of coordinating sentences) have been assessed in Alzheimer's disease and mild cognitive impairment (Dodge et al., 2015; Orimaye et al., 2017). Another example is the use of linguistic analysis in the differential diagnosis between psychogenic non-epileptic attack disorder and epilepsy, which shows that the way a patient describes an epileptic attack can be used to differentiate between these disorders (Ekberg and Reuber, 2015; Jenkins et al., 2016). Furthermore, there is ample evidence that language comprehension is also disturbed in some brain disorders, see for instance Condray et al.,



Study name	Patient group	Sample size Control	Sample size Patient	Hedges's *g*	*p*-value
Lee et al. 2017	ASD	19	14	0.974	0.007
Losh & Gordon 2014	ASD	22	26	1.567	<.0001
Luo et al. 2016	ASD	26	130	0.450	0.037
Prud'hommeaux et al. 2017	ASD	22	22	0.573	0.058
Rouhizadeh et al. 2015	ASD	25	44	0.852	<.0001
Subgroup ASD		**114**	**236**	**0.844**	**<.0001**
Hoffman et al. 2014	dementia	7	8	1.499	0.007
Pakhomov et al. 2012	dementia	55	31	0.759	0.001
Pakhomov et al. 2014	dementia	84	155	0.273	0.044
Pakhomov et al. 2015 - MCI	dementia	17	5	<.0001	1.000
Pakhomov et al. 2015 - AD	dementia	21	6	0.265	0.556
Subgroup dementia		**184**	**205**	**0.494**	**0.012**
Bedi et al. 2015	psychosis	5	29	1.054	0.031
Cecchi et al. 2016	psychosis	19	40	1.640	<.0001
Nicodemus et al. 2014	psychosis	194	471	0.472	<.0001
Tagamets et al. 2014	psychosis	11	11	0.804	0.060
Subgroup psychosis		**229**	**551**	**0.958**	**0.003**

Fig. 4. Meta-analysis of the efficacy of semantic space models per diagnostic group. Squares indicate Hedges' *g* for individual studies and are scaled to sample size. Black bars indicate 95% confidence intervals. Red diamonds represents the pooled effect sizes per patient group. CI = confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

1996; Gavilán and García-Albea, 2011 for their findings in patients with schizophrenia. Research focusing on using linguistic measures of verbal communication as a diagnostic or prognostic marker of brain disorders should thus not focus on semantic space calculations alone; instead a wide spectrum of analyses should be incorporated. This could also aid in further describing the linguistic characteristics of the different disorders.

4.1. Strengths and limitations

We have comprehensively reviewed applications of semantic space in psychiatric and neurodegenerative clinical contexts. The result of our meta-analysis shows an overall significant difference in semantic space measures between various clinically relevant groups and controls, whether healthy matched controls, cohort studies, at-risk populations or family members.

A limitation of this study is that the studies are very heterogeneous since they assess different disorders, use different methods to collect language and employ various calculations of semantic space. Furthermore, most studies did not focus of diagnostic test accuracy, but rather on finding differences between groups. Though it is interesting to see semantic space calculations differ quite strongly between patients and healthy controls (effect sizes were high), measures of diagnostic test accuracy are necessary to analyze the quality of these methods in a clinical setting. Only a small proportion of the studies provided sufficient data to calculate diagnostic test accuracy, which made it difficult to do further subgroup analyses. The overall pooled sensitivity and specificity is therefore still hard to interpret from a clinical perspective, since it provides an accuracy of diagnosing multiple disorders. For that reason, the analyses of diagnostic test accuracy were repeated for the psychosis group alone. Given the small number of studies, these analyses could not be performed for other subgroups. Future research investigating these techniques should therefore aim to additionally measure the sensitivity and specificity of their methods.

4.2. Recommendations

Future studies should also focus on newer methods to calculate semantic space, since they provide a promising alternative to LSA. A novel and easily accessible tool for calculating vector spaces, embedding and similarity is word2vec, which is a group of related models developed by Tomas Mikolov and his colleagues (Mikolov et al., 2013a,b). Word2vec models are neural network models that were proposed for computing vector representations and semantic space models of large data sets of words. These models are based on the assumption that words can be ‘similar’ in multiple ways. In contrast to LSA and similar approaches, word2vec models can measure both semantic (meaning) and syntactic (grammatical) regularities (Glasgow et al., 2016; Villegas et al., 2016). Neural networks outperform LSA in preserving linear relations between words, while they are less computationally expensive in large data sets (Mikolov et al., 2013c). In particular word2vec has been shown to perform better than LSA in multiple studies, especially in capturing grammatical relations between words (Glasgow et al., 2016; Villegas et al., 2016). To the best of our knowledge, only one study (Prud'hommeaux et al., 2017) used word2vec as a tool to distinguish patients from healthy controls, in addition to LSA. Word2vec showed slightly lower accuracy scores compared to their LSA models, however, these analyses were performed on single word responses on verbal fluency tasks. Future studies should also test the efficacy of word2vec on full-sentences, especially given the syntactic properties of these models.

4.3. Conclusion

Semantic space models, such as LSA or word2vec potentially are a valuable diagnostic and prognostic tool in a variety of neurological and psychiatric disorders. To date, their use has mostly been tested to compare patients to healthy controls. Future studies should focus on testing these models' specificity to compare different disorders and thus their potential to be used as a differential-diagnostic tool.

Conflict of interest

The authors declare no conflict of interest regarding the topic and contents of this article.

Acknowledgements

Special thanks to Dr. Clark, Dr. Luo, Dr. Prud'hommeaux, and Dr. Rosenstein for providing additional data necessary for performing the current meta-analysis.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.neubiorev.2018.06.008.

References

Altmann, L.J.P., Troche, M.S., 2011. High-level language production in Parkinson's disease: a review. Park. Dis. 2011.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub.

Appell, J., Kertesz, A., Fisman, M., 1982. A study of language functioning in Alzheimer patients. Brain Lang. 17, 73–91. http://dx.doi.org/10.1016/0093-934X(82)90006-2.

Bedi, G., Carrillo, F., Cecchi, Ga., Slezak, D.F.F., Sigman, M., Mota, N.B.N.B., Ribeiro, S., Javitt, D.C., Copelli, M., Corcoran, C.M., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 1, 15030. http://dx.doi.org/10.1038/npjschz.2015.30.

Borenstein, M., Hedges, L., Higgins, J., Rothstein, H., 2005. Comprehensive Meta-Analysis Version 2. NJ Biostat, Englewood 104.

Carruthers, P., 2002. The cognitive functions of language. Behav. Brain Sci. http://dx.doi.org/10.1017/S0140525X02000122.

Cecchi, G., 2016. A computational linguistics approach for prodromal psychosis. Neuropsychopharmacology 41, S97–S98. http://dx.doi.org/10.1038/npp.2016.239.

Clark, D.G., Mclaughlin, P.M., Woo, E., Hwang, K., Hurtz, S., Ramirez, L., Eastman, J., Dukes, R.-M.R.-M., Kapur, P., Deramus, T.P., Apostolova, L.G., 2016. Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. Alzheimer's Dement. Diagn. Assess. Dis. Monit. 2, 113–122. http://dx.doi.org/10.1016/j.dadm.2016.02.001.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. NJ Lawrence Earlbaum Assoc., Hilsdale 2.

Cohen, A.S., Elvevåg, B., 2014. Automated computerized analysis of speech in psychiatric disorders. Curr. Opin. Psychiatry 27, 203.

Cohen, A., Mitchell, K., Elvevåg, B., 2014. What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments. Schizophr. Res. 159, 533–538. http://dx.doi.org/10.1016/j.schres.2014.09.013.

Condray, R., Steinhauer, S.R., van Kammen, D.P., Kasparek, A., 1996. Working memory capacity predicts language comprehension in schizophrenic patients. Schizophr. Res. 20, 1–13.

Corcoran, C.M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D.C., Bearden, C.E., Cecchi, G.A., 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry 17, 67–75. http://dx.doi.org/10.1002/wps.20491.

Dell, G.S., Burger, L.K., Svec, W.R., 1997. Language production and serial order: a functional analysis and a model. Psychol. Rev. 104, 123–147. http://dx.doi.org/10.1037/0033-295X.104.1.123.

Dodge, H., Mattek, N., Gregor, M., Bowman, M., Seelye, A., Ybarra, O., Asgari, M., Kaye, J.A., 2015. Social markers of mild cognitive impairment: proportion of word counts in free conversational speech. Curr. Alzheimer Res. 12, 513–519.

Ekberg, K., Reuber, M., 2015. Can conversation analytic findings help with differential diagnosis in routine seizure clinic interactions? Commun. Med. 12, 13–24. http://dx.doi.org/10.1558/cam.v12i1.26851.

Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr. Res. 93, 304–316. http://dx.doi.org/10.1016/j.schres.2007.03.001.

Elvevåg, B., Foltz, P.W., Rosenstein, M., DeLisi, L.E., 2010. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. J. Neurolinguist. 23, 270–284. http://dx.doi.org/10.1016/j.jneuroling.2009.05.002.

Emmorey, K., 2001. Language, Cognition, and the Brain: Insights from Sign Language Research. Psychology Press.

Firth, J.R., 1957. A Synopsis of Linguistic Theory, 1930-1955.

García, A.M., Carrillo, F., Orozco-Arroyave, J.R., Trujillo, N., Vargas Bonilla, J.F., Fittipaldi, S., Adolfi, F., Nöth, E., Sigman, M., Fernández Slezak, D., Ibáñez, A., Cecchi, G.A., 2016. How language flows when movements don't: an automated analysis of spontaneous discourse in Parkinson's disease. Brain Lang. 162, 19–28. http://dx.doi.org/10.1016/j.bandl.2016.07.008.

Gavilán, J.M., García-Albea, J.E., 2011. Theory of mind and language comprehension in schizophrenia: poor mindreading affects figurative language comprehension beyond intelligence deficits. J. Neurolinguist. 24, 54–69.

Glasgow, K., Roos, M., Haufler, A., Chevillet, M., Wolmetz, M., 2016. Evaluating Semantic Models With Word-Sentence Relatedness. arXiv, pp. 1–8.

Godino-Llorente, J.I., Shattuck-Hufnagel, S., Choi, J.Y., Moro-Velázquez, L., Gómez-García, J.A., 2017. Towards the identification of Idiopathic Parkinson's disease from the speech. New articulatory kinetic biomarkers. PLoS One 12, e0189583.

Hajian-Tilaki, K., 2013. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Casp. J. Intern. Med. 4, 627.

Higgins, E.T., Idson, L.C., Freitas, A.L., Spiegel, S., Molden, D.C., 2003. Transfer of value from fit. J. Pers. Soc. Psychol. 84, 1140.

Hoffman, P., Lambon Ralph, M.A., Rogers, T.T., 2013. Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. Behav. Res. Methods 45, 718–730. http://dx.doi.org/10.3758/s13428-012-0278-x.

Hoffman, P., Meteyard, L., Patterson, K., 2014. Broadly speaking: vocabulary in semantic dementia shifts towards general, semantically diverse words. Cortex 55, 30–42. http://dx.doi.org/10.1016/j.cortex.2012.11.004.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458.

Jenkins, L., Cosgrove, J., Chappell, P., Kheder, A., Sokhi, D., Reuber, M., 2016. Neurologists can identify diagnostic linguistic features during routine seizure clinic interactions: results of a one-day teaching intervention. Epilepsy Behav. 64, 257–261. http://dx.doi.org/10.1016/j.yebeh.2016.08.008.

Landauer, T., Dumais, S., 1997. A solution to Plato's problem : the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104, 211–240. http://dx.doi.org/10.1037/0033-295X.104.2.211.

Landauer, T., Foltz, P., Laham, D., 1998. An introduction to latent semantic analysis. Discourse Process. 25, 259–284. http://dx.doi.org/10.1080/01638539809545028.

Lee, M., Martin, G.E., Hogan, A., Hano, D., Gordon, P.C., Losh, M., 2017. What's the story? A computational analysis of narrative competence in autism. Autism. http://dx.doi.org/10.1177/1362361316677957. 136236131667795.

Logemann, J.A., Fisher, H.B., Boshes, B., Blonsky, E.R., 1978. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. J. Speech Hear. Disord. 43, 47–57.

Losh, M., Gordon, P.C., 2014. Quantifying narrative ability in autism spectrum disorder: a computational linguistic analysis of narrative coherence. J. Autism Dev. Disord. J. Autism Child. Schizophr. 44, 3016–3025. http://dx.doi.org/10.1007/s10803-014-2158-y.

Luo, S.X., Shinall, J.A., Peterson, B.S., Gerber, A.J., 2016. Semantic mapping reveals distinct patterns in descriptions of social relations in adults with autism spectrum disorder. Autism Res. 9, 846–853. http://dx.doi.org/10.1002/aur.1581.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient Estimation of Word Representations in Vector Space. Arxiv, pp. 1–12. http://dx.doi.org/10.1162/153244303322533223.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. pp. 3111–3119.

Mikolov, T., Yih, W., Zweig, G., 2013c. Linguistic regularities in continuous space word representations. Proc. NAACL-HLT. pp. 746–751.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 6, e1000097.

Nicodemus, K.K., Elvevag, B., Foltz, P.W., Rosenstein, M., Diaz-Asper, C., Weinberger, D.R., Elvevåg, B., Foltz, P.W., Rosenstein, M., Diaz-Asper, C., Weinberger, D.R., 2014. Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. Cortex J. Devoted Stud. Nerv. Syst. Behav. 55, 182–191. http://dx.doi.org/10.1016/j.cortex.2013.12.004.

Orimaye, S.O., Wong, J.S.-M., Golden, K.J., Wong, C.P., Soyiri, I.N., 2017. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. BMC Bioinform. 18, 34. http://dx.doi.org/10.1186/s12859-016-1456-0.

Padó, S., Lapata, M., 2007. Dependency-based construction of semantic space models. Comput. Linguist. 33, 161–199. http://dx.doi.org/10.1162/coli.2007.33.2.161.

Pakhomov, S.V.S., Hemmy, L.S., 2014. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. Cortex J. Devoted Stud. Nerv. Syst. Behav. 55, 97–106. http://dx.doi.org/10.1016/j.cortex.2013.05.009.

Pakhomov, S.V.S., Hemmy, L.S., Lim, K.O., 2012. Automated semantic indices related to cognitive function and rate of cognitive decline. Neuropsychologia 50, 2165–2175. http://dx.doi.org/10.1016/j.neuropsychologia.2012.05.016.

Pakhomov, S.V.S., Jones, D.T., Knopman, D.S., 2015. Language networks associated with computerized semantic indices. Neuroimage 104, 125–137. http://dx.doi.org/10.1016/j.neuroimage.2014.10.008.

Pinker, S., 2007. The Stuff of Thought: Language as a Window into Human Nature. Penguin.

Prud'hommeaux, E., Roark, B., 2015. Graph-based word alignment for clinical language evaluation. Comput. Linguist. 41, 549.

Prud'hommeaux, E., van Santen, J., Gliner, D., 2017. Vector space models for evaluating semantic fluency in autism. Proc. 55th Annu. Meet. Assoc. Comput. Linguist. http://dx.doi.org/10.18653/v1/P17-2006. (Volume 2 Short Pap. 32–37).

Rosenstein, M., Diaz-Asper, C., Foltz, P.W., Elvevag, B., 2014. A computational language approach to modeling prose recall in schizophrenia. Cortex J. Devoted Stud. Nerv. Syst. Behav. 55, 148–166.

Rouhizadeh, M., Sproat, R., van Santen, J., 2015. Similarity measures for quantifying restrictive and repetitive behavior in conversations of autistic children. Proc. Conf. Assoc. Comput. Linguist. North Am. Chapter. Meet. 2015. pp. 117–123.

Tagamets, M.A., Cortes, C.R., Griego, J.A., Elvevag, B., Elvevåg, B., 2014. Neural correlates of the relationship between discourse coherence and sensory monitoring in schizophrenia. Cortex 55, 77–87. http://dx.doi.org/10.1016/j.cortex.2013.06.011.

Villegas, M.P., José, M., Ucelay, G., Fernández, J.P., Álvarez-Carmona, M.A., Errecalde, M.L., Cagnina, L.C., 2016. Vector-based word representations for sentiment analysis: a comparative study. XXII Congr. Argentino Ciencias la Comput. (CACIC 2016). pp. 785–793.

Walter, S.D., 2002. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat. Med. 21, 1237–1256.

White, H.A., Shah, P., 2016. Scope of semantic activation and innovative thinking in college students with ADHD. Creat. Res. J. 28, 275–282. http://dx.doi.org/10.1080/10400419.2016.1195655.

Whiting, P.F., Rutjes, A.W.S., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B., Leeflang, M.M.G., Sterne, J.A.C., Bossuyt, P.M.M., 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern. Med. 155, 529–536.

Zamora, J., Abraira, V., Muriel, A., Khan, K., Coomarasamy, A., 2006. Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med. Res. Methodol. 6, 31.

Zhang, J., Pan, Z., Gui, C., Xue, T., Lin, Y., Zhu, J., Cui, D., 2018. Analysis on speech signal features of manic patients. J. Psychiatr. Res. 98, 59–63.