Mapping semantic networks to Dutch word embeddings as a diagnostic tool for cognitive decline

Anonymous ACL submission

Abstract

We explore the possibility of semantic networks as a diagnostic tool for cognitive decline by using Dutch verbal fluency data to investigate the relationship between semantic networks and cognitive health. We constructed these networks using the Dutch BERTje contextual word embedding model. In psychology, semantic networks serve as abstract representations of the semantic memory system. Semantic verbal fluency data can be used to estimate said networks. Traditionally, this is done by counting the number of raw items produced by participants in a verbal fluency task. We used BERTje to connect the elicited words through semantic similarity scores, and extracted three network distance metrics in addition to the traditional approach of counting. We then tested how well these metrics predict participants' cognitive health scores on the Mini-Mental State Examination (MMSE). Using a data-driven variable selection procedure, we found that the centroid-diameter metric is the only significant predictor of cognitive health scores in our data, with the traditional scoring method not reaching significance.

1 Introduction

A universally used neuropsychological method to assess cognitive state in people with cognitive impairment or disorders is the semantic, or categorical, verbal fluency task (Chi et al., 2014; Maseda et al., 2014; Quaranta et al., 2019). However, there is a growing body of studies questioning the granularity of this task's assessment method (Linz et al., 2018; March and Pattison, 2006; Shao et al., 2014). A more granular approach to interpreting the verbal fluency data is through semantic networks (Chan et al., 1993; Goñi et al., 2011; Martínez-Nicolás et al., 2019). Semantic networks serve as abstract representations of the semantic memory system and therefore enable valuable insights into higher cognitive concepts, including cognitive health. The body of research establishing a link between semantic networks and cognitive health is steadily growing (Chan et al., 1993; Lerner et al., 2009; Martínez-Nicolás et al., 2019).

Distributional semantic models are a related class of semantic memory models: these language models encode meaning representations, predicated on the hypothesis that the statistical distribution of linguistic items within a given context significantly influences and defines their semantic attributes (Landauer and Dumais, 1997). Although distributional semantic models do not directly represent semantics, they do represent associations that reflect word semantic similarity (Hill et al., 2015; McRae et al., 2012). Such models use word embeddings to quantify this semantic similarity. By combining distributional semantic models with network-based models, it is possible to reinforce semantic network-based accounts through the application of machine learning techniques (Steyvers and Tenenbaum, 2005; Utsumi, 2015).

As far as we are aware, semantic networks derived from distributional semantic modelling have never been used to study cognitive health. In this study, we investigate whether semantic networks generated from verbal fluency data using contextual word embedding models can function as a reliable or generally better diagnostic tool for assessing cognitive impairment in comparison to the use of conventional scoring methods. In particular, we use verbal fluency data in the Dutch language to investigate the connection between semantic networks and cognitive health. Our experiment contrasts novel scoring metrics with traditional scoring to determine which is a better predictor of cognitive health. Furthermore, the majority of discussion on semantic networks and cognitive health, as well as NLP applications in general, is centered around the English language. By working with Dutch, we aim to show that our approach is useful for mid-resource languages as well.

2 Background

Verbal fluency tasks are a common component of standardized assessments used for screening mild cognitive impairment and evaluating cognitive state of various populations, including those with schizophrenia (Frith et al., 1995), Alzheimer's disease (Clark et al., 2016; Monsch et al., 1992; Troyer et al., 1998), and Parkinson's disease (Piatt et al., 1999; Troyer et al., 1998). Throughout this task, participants are asked to generate as many items as possible within a specific category, all within a fixed time frame. Categories can be semantic, such as 'animals', or phonemic such as 'words starting with an a' (Bousfield and Sedgewick, 1944). The score is determined by tallying the number of correctly generated words; with correctly generated in this context referring to items fitting within the given category. The verbal fluency task holds its popularity in neuropsychological assessments for several reasons: it is relatively brief and taps into both executive control and semantic memory retrieval processes. It is not only effective within a larger battery of assessments, but even on its own, the verbal fluency task scores are capable of separating people with cognitive impairment from healthy controls (Chi et al., 2014; Clark et al., 2009, 2016; Henry et al., 2004; McDonnell et al., 2020).

However, several studies have questioned the granularity of traditional verbal fluency assessment methods, which predominantly rely on the raw count of word responses (March and Pattison, 2006; Shao et al., 2014). This scoring method overlooks relationships among responses, neglecting potential insights into the structure of an individual's knowledge representations and underlying cognitive processing mechanisms (Linz et al., 2018; Troyer et al., 1997). In other words, the scoring system of the semantic verbal fluency task, which revolves around a semantic category, does not take any semantic information into account. To address this shortcoming, various qualitative metrics have been introduced to complement the raw word count.

One advance in verbal fluency data analysis involves the identification of temporal and semantic clusters within the verbal fluency dataset (clustering) and the moment of jumping between these clusters (switching) (Troyer et al., 1997). Typically, word production in a fluency task follows a pattern of organized bursts of retrieval of words within clusters, referred to as temporal clusters, interrupted by pauses suggesting a lexical search for transitions among clusters. Consequently, the words within these temporal clusters are inclined to share linguistic features. In effect, considering the mean size of clusters and the number of cluster switches offers insights into the strategies participants employ during word searches. The two components complement each other nicely since clustering is associated with local explorations of semantic memory while switching is an indication of the flexibility to start looking for words in a new semantic cluster. Both components rely on distinct brain regions: switching predominantly involves areas within the fronto-parietal network associated with executive functioning, while clustering depends on the activation of regions within the temporal cortex, along with the hipoccampus and the cerebellum, which are crucial for semantic memory retrieval and categorization of information (Goñi et al., 2011; Troyer et al., 1998; Ahn et al., 2022; Lundin et al., 2023). Because of this, the scoring of the verbal fluency task with these two additional components has been substantiated as a more detailed indicator of cognitive state. Clustering and switching, however, are sensitive to human errors and subjectivity, as the cluster boundaries are decided by the scorer. To mitigate the influence of human errors and subjective judgment, researchers have devised automated scoring methods. These methods encompass not only the automated analysis of clustering (König et al., 2018; Linz et al., 2017a, 2018; Kim et al., 2019) but also data-driven assessments of the semantic properties of verbal fluency data through semantic networks (Martínez-Nicolás et al., 2019).

2.1 Semantic networks

The concept of semantic networks originates in network theory, a field that has broadened the understanding of a wide variety of systems, including language and semantics (Borge-Holthoefer and Arenas, 2010; Cong and Liu, 2014; Ke and Yao, 2008). Remarkably, for specific cognitive concepts like creativity, these networks have proven to be superior to traditional methods, effectively eliminating human errors in assessment (Beaty and Johnson, 2021). A creativity assessment such as the Alternate Uses Task (AUT; Benedek et al., 2013) bears a striking resemblance to the verbal fluency task in measuring associative cognition. AUT measures creativity by asking individuals to generate as many alternative uses for a common object or item as possible within a fixed time frame. This task considers both the quality and originality of responses when

scoring, distinguishing it from fluency tasks that do not assess response quality (Beaty and Johnson, 2021; Johnson et al., 2023). Frith et al. (2021) have found a shared cognitive and neural foundation of intelligence and creativity by examining verbal fluency data for the former and AUT data for the latter. Previous work has shown that individual-level semantic networks can be estimated through verbal fluency data (Zemla and Austerweil, 2017, 2018; Zemla et al., 2016, 2020).

During a semantic association task like the (semantic) verbal fluency task, it is expected that associations exhibit specific semantic connections or overlays. The evaluation of the quality or strength of these connections is a central aspect of semantic networks. Thus, by interpreting fluency data as a network, the potential for expanding the assessment's scope becomes evident. Considering the proven role of semantic networks in assessing complex cognitive constructs such as creativity, it is reasonable to expect that they could also provide more precise insights into cognitive health.

Semantic networks can be seen as an abstract representation of the semantic memory system. In these networks the nodes represent concepts and the edges represent similarity, co-occurrence, or strength of associations among semantic or lexical units (Christensen and Kenett, 2023a; Lerner et al., 2009; Muhammad et al., 2019). According to the spreading activation model, nodes within a semantic network represent concepts from semantic memory, and their interconnections depend on the degree of semantic similarity. Concepts that are semantically more similar are positioned closer together and linked with stronger connections than concepts that are less semantically related. By this model, when a concept is activated, its mental representation spreads to interconnected concepts, gradually diminishing as the distance between them increases (Kenett et al., 2017). It is the capacity of semantic networks to mirror the structure of semantic memory that makes them valuable in the exploration of higher cognitive functions. Semantic networks have been utilized to evaluate different aspects of cognition, with the majority of studies focusing on the English language, such as cognitive aging (Cosgrove et al., 2021; Wulff et al., 2022), Alzheimer's disease (Chan et al., 1993; Lerner et al., 2009; Martínez-Nicolás et al., 2019), cognitive processes (Christensen and Kenett, 2023b), associative abilities (He et al., 2021) and creativity (He et al., 2021; Kenett and Faust, 2019).

2.2 Distributional semantics

Another important class of lexical representation models, in addition to network-based models, are distributional semantic models: computational models of human semantic memory (Utsumi, 2015). Distributional semantic models fall under the category of language models based on the distributional hypothesis (Landauer and Dumais, 1997). This hypothesis posits that the semantics of a word are intrinsically linked to its contextual usage. As such, the statistical distribution of words within a language corpus significantly informs their meaning (Erk, 2012). In essence, these language models operationalize meaning by analyzing the distributional patterns of linguistic items in large language corpora, extracting associations between words from statistical regularities and co-occurrence frequencies.

Informed by the distributional hypothesis, they model the semantics of a word as intrinsically linked to its contextual usage (Erk, 2012). Subsequent word embedding approaches such as Word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2018) are based on this paradigm. These models do not explicitly model semantics, but they do model associations that correspond to semantic similarity between words (Hill et al., 2015; McRae et al., 2012). Language models have been evaluated on both of these linguistic concepts: association and similarity. In these models, we quantify semantic similarity by computing the cosine similarity between vector representations of the two words. These similarity scores can then be correlated with human similarity rating datasets. A prominent association, or relatedness, dataset is WordSim-353 (Finkelstein et al., 2001), whereas SimLex-999 (Hill et al., 2015) is a widely used similarity dataset. These benchmarks are language-specific and, in this case, both evaluate English language models. Alternatives such as Multi-Simlex (Vulić et al., 2020) have been developed for other languages, and for Dutch, Brans and Bloem's (2024) Dutch SimLex-999 was used to evaluate the BERT-based language models BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020). Models that correlate well with human word pair similarity scores can be harnessed for semantically interpreting language data.

It is uncommon in the literature to find connections between distributional semantic models and network-based models, as they are separate classes

of models rooted in different fields of study (for a critical review, see Kumar et al., 2022), but a few studies have used the cosine similarity between words as a threshold in constructing paths in semantic networks (Steyvers and Tenenbaum, 2005; Utsumi, 2015). The approach has not been used for diagnostic purposes as far as we are aware.

3 Methods

We aim to test whether semantic networks derived from verbal fluency data can serve as a robust or overall greater diagnostic tool for assessing cognitive impairment, relative to the use of the most commonly employed traditional scoring method: the overall number of responses. Specifically, we explore the relationship between semantic networks and cognitive health using Dutch verbal fluency data.

3.1 Data

We leverage a rich existing dataset, which was originally collected by Konijnenberg et al. (2018). It includes verbal fluency data from participants with mild cognitive impairment, along with their results on the Mini-Mental State Examination (MMSE; Folstein et al., 1975). The MMSE is one of the best-known cognitive screenings for providing an overall measure of cognitive impairment.

3.1.1 Participants

Participants were recruited as part of the EMIF-AD PreclinAD study, a project on cognitive impairment and dementia as a consequence of amyloid pathology (Konijnenberg et al., 2018). The subset of their data that the present study reuses specifically includes monozygotic twins, recruited through the Netherlands Twins Register. They invited 517 participants, from which 313 were excluded from participation (See Appendix A for the exclusion flow chart). After the application of the same exclusion criteria as the original study, a total of 204 cognitively normal participants (119 females) between 60 and 94 years old (mean age: 70.8 ± 7.8 years old) remain. All participants gave written informed consent for participation in the original study.

3.1.2 Procedure

All participants completed a categorical verbal fluency task (semantic category: animals) as part of a larger battery of neuropsychological tests. They got the following instruction (translated to English from Dutch): "You will have two minutes to list as many animals as possible. Anything is fine as long as it is an animal. When the time is up, I will say stop. You may begin." While administering the test, the generated items were directly transcribed to track the number of items. Along with these data, we have access to the final scores on the Mini-Mental State Examination (MMSE), which range from 0 to 30 (mean score: 28.9 ± 1.2). The MMSE is divided into two sections. The first assesses orientation, memory, and attention (maximum score: 21); the second measures the ability to name, obey written and verbal instructions, compose a sentence on the spot, and replicate a challenging polygon that resembles the Bender-Gestalt Figure (maximum score: 9) (Folstein et al., 1975).

3.1.3 Preprocessing

We spell-checked the responses, and any irrelevant responses (intrusions) such as "aardbei" (strawberry) and "glas" (glass) were identified and removed. We did however decide to admit duplicate responses (perseverations). The omission of both intrusions and perseverations is routine to the verbal fluency assessment (Christensen and Kenett, 2023a). However, the omission of perseverations would skew our metric since we are interested in the direct order of the items and the semantic similarity between them. Because the task is inherently based on a semantic category, it is the only restriction that participants get, we did decide to omit intrusions. If not for the semantic category, it would be a free association task.

3.2 Model

We use the Dutch contextual embedding model BERTje (de Vries et al., 2019) to build individual semantic networks from the verbal fluency data. BERTje is a Dutch sibling of BERT (Devlin et al., 2018), which has been used to investigate semantic representations and associative cognition in English (Johnson et al., 2023). Instead of using cosine similarity as a threshold to construct paths, we connect every node to every other node and use cosine similarity as a notion of distance that reflects the semantic similarity between the words.

3.2.1 Extracting embeddings

Each participant's verbal fluency data is embedded using BERTje, where each item gets translated into a 768-dimensional embedding vector. Within this space, cosine similarity is a measure of semantic similarity: greater proximity between two embed-

dings reflects a stronger semantic similarity (Linz et al., 2017a).

BERTje is a contextual embedding model. In the case of a fluency task, however, the participant's responses are divorced from their context, since the input provided by the participant only allows for single-word responses. Taking this into account, we use only input layer embeddings, which do not add any contextual information to the embedding (Ethayarajh, 2019). This is similar to static embeddings, but with the advantage of being able to generate embeddings for out-of-vocabulary words through BERTje's WordPiece tokenizer. If our items are split up by BERTje's tokenizer, we construct an embedding by averaging the embeddings of the word's subtokens. We implemented the same procedure for items consisting of multiple words, e.g., "orang-oetan" (orangoutang), "bruine beer" (brown bear), and "Vlaamse gaai" (jay).

3.3 Analysis

To investigate the relation between the semantic verbal fluency results and MMSE scores, we captured different aspects of the fluency results in the following four metrics.

Each participant's vector space that holds all the embedded items (the embedding space) can be interpreted as a semantic network in the sense that all the items can be connected through cosine similarity scores. The embedded items (words) can be seen as the nodes or objects, while the edges represent the semantic similarity between them. The way we are considering individual semantic networks based on one verbal fluency task is identical to cluster analysis on just one cluster: our network does not have weighted nodes or directed edges and can therefore be seen as one cluster in which all the items are connected through semantic similarity. Because of this, cluster analysis techniques can be applied to our data, specifically, intracluster distances (distances between objects belonging to the same cluster) are of interest to the current analysis. Some of these metrics were inspired by Oortwijn et al.'s (2021) use of cluster analysis methods for target terms in embedding models.

To explore different aspects of the participant's network quality, we considered the following four metrics based on the properties of cosine similarity:

Number of words The number of generated items in the verbal fluency task. This is how the verbal fluency task is classically scored and its relation to MMSE scores has been investigated in previous research (Linz et al., 2017b).

Path length The sum of the similarity scores between every sequential embedding pair. Following the order in which the participant-generated the items, we sum all scores between the items. The path length is one of the main measures of networks, though one must bear in mind that in our case this score does not represent the number of steps needed to get from one word in the network to another (Kenett et al., 2017), but rather the cosine similarity between two words. As it is common to measure the amount of semantic clustering in a verbal fluency task, this is somewhat represented by this metric: words within the same semantic cluster will be more similar and thus have a higher semantic similarity; words from different semantic clusters will be less similar and thus have a lower semantic similarity. In a way, this metric bears some similarity to that of forward flow. This is a measure, formally home to creativity research, that captures associative cognition by evaluating how far participants traverse their semantic space in a free association task (Gray et al., 2019).

Average diameter distance The average distance between all words generated by the participant – this is an intracluster distance from the field of cluster analysis (Bolshakova and Azuaje, 2003). It is defined as follows, where S is a semantic network; $d(x, y)$ defines the distance between any two items, x and y, belonging to S; and $|S|$ represents the number of items included in network S.

$$\Delta_A(S) = \frac{1}{|S| \cdot (|S| - 1)} \sum_{\substack{x,y \in S \\ x \neq y}} d(x, y) \quad (1)$$

Centroid diameter distance Twice the average distance between all items and the centroid of that participant's network. To determine the centroid of each participant's cluster, we calculated the mean of the normalized embedding vectors for each participant. This too is an intracluster distance derived from the field of cluster analysis (Bolshakova and Azuaje, 2003). In the equations below, S is a semantic network; $d(x, \bar{v})$ defines the distance between any item x belonging to S and the centroid $\bar{v}$; and $|S|$ represents the number of items included in network S.

$$\Delta_C(S) = 2 \left(\frac{\sum_{x \in S} d(x, \bar{v})}{|S|} \right) \quad (2)$$

where

$$\bar{v} = \frac{1}{|S|} \sum_{x \in S} x \tag{3}$$

We compare these four metrics and apply a data-driven approach to determine which variable proves to be most effective in predicting cognitive health. In general, we expect measures indicating a broader semantic network to be related to higher scores on the MMSE cognitive battery. Concerning our second objective, we hypothesize that path length and centroid diameter distance will be the best predictors of cognitive decline

3.3.1 Statistics

We employ Generalized Additive Models (GAM) to investigate the impact of the aforementioned metrics representing network properties (the number of words, path length, average diameter distance, and centroid diameter distance) on the MMSE scores. The data analyses were performed using R (v4.1.2; R Core Team, 2021). We made use of the gam() function from the mgcv R package (1.8.42; Wood, 2011). GAMs offer a means of capturing non-linear associations by estimating a curve based on penalized smoothing splines. These splines are functions composed of simpler functions that create a resultant spline that is the sum of evaluated basis functions for each of the fitted variables. This creates possibilities for curves in the model fit where a non-linear relationship more accurately describes the variance within the observed data. To ensure robust generalizability of our model, efforts were made to prevent overfitting, where the model becomes overly tailored to individual data points, potentially compromising its performance on new data.

This method optimizes the trade-off between the model's ability to fit the data and its complexity, or "wiggliness" (Wood, 2011). A smooth curve depicts a more fundamental, straight-line function, whereas a "wiggly" curve denotes a more complicated one. To compensate for the amount of "wiggliness", a penalty is added since a greater distortion of the curve corresponds to an increasing likelihood of overfitting the model. This penalty preserves the balance between the function's complexity and fit. Non-linearity is thus only introduced in a model if the variance explained by a wiggly line outweighs the penalty that comes with it (Wood, 2020). When there is not enough evidence available to support a curved function, GAMs will default to modeling a linear relationship. The crucial metric used to check the complexity is the effective degrees of freedom (edf), which serves as an indicator of whether the predictor variable maintains a non-linear (edf > 1) or linear (edf = 1) relation with the dependent variable.

In our first-level analysis, we fitted GAMs with MMSE scores as the outcome variable and all metrics as predictors: number of words, path length, average diameter distance, and centroid diameter distance. In this step, we employed a variable selection procedure with two goals: (1) to optimize our model enhancing its predictive accuracy by determining which available variables were the most influential predictors; (2) to address whether the traditional scoring method (i.e., tallying number of words) is a better predictor of the MMSE-scores than the metrics derived from semantic networks. We did this by incorporating the double penalty approach into our GAMs (Marra and Wood, 2011). This is a variable selection method that is data-driven and uses an empirical Bayes procedure to identify any predictor variables that do not have a significant effect on the outcome variable. This method is implemented by including the argument select=TRUE in the gam() function in R. Subsequently, we ran GAM for our second-level analysis, in which we incorporated only those predictors that emerged as significant in the first-level analysis.

4 Results

Performing the first-level analysis with GAMs and the double penalty approach allowed us to determine which variables did not influence the MMSE-score prediction. The results of our analysis showed that the average diameter distance scores, path length, and notably also the word count did not significantly affect the MMSE scores (all p-values > 0.071). Figure 1 shows the predicted partial effects of the traditional number of words metric, which is not statistically significant.

As a result, the model for our second-level analysis did not include these factors. The centroid diameter distance was the only predictor taken into the second-level analysis.

The second-level analysis thus tests whether centroid diameter distance serves as a significant predictor of MMSE scores. This is the only variable that remained after our first-level analysis. The results revealed that the centroid diameter distance did emerge as a highly significant positive predic-

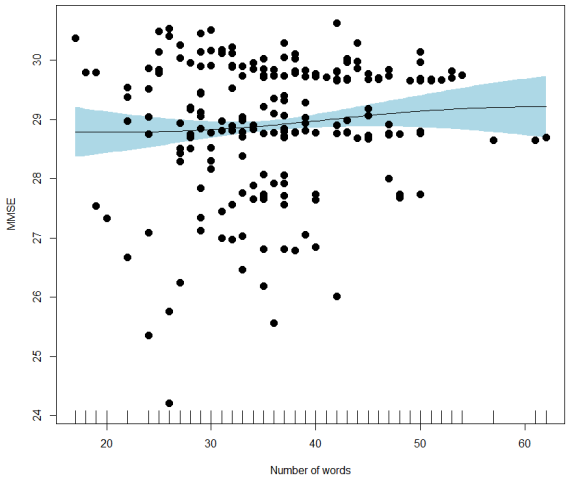

Figure 1: Effects of number of words on the MMSE-scores in the first-level analysis. GAMMs with 95% confidence interval (blue color) illustrates a small effect. The y-axis shows partial effects (for 1 of 4 variables) thus the values are not whole numbers.

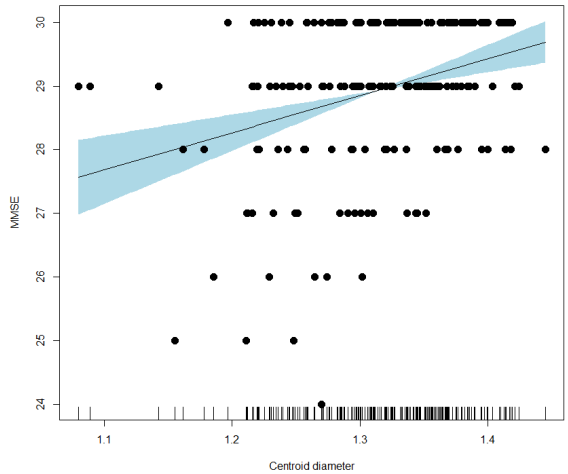

Figure 2: Effects of centroid diameter distance on the MMSE-scores in the second-level analysis. GAMMs with 95% confidence interval (blue color) illustrates a significant effect.

tor of the MMSE scores (p-value < 0.001; edf $= 1.001$). These results are illustrated in Figure 2. With the estimated degrees of freedom close to 1, this relationship is linear. Additionally, it tells us the centroid diameter term is not overly smoothed. To assess the model fit, we utilized the gam.check() function from the mgcv package in R. This function confirmed full convergence of the model and indicated that the Hessian matrix is positive definite, suggesting good stability ($k\prime = 9$; k-index $= 0.94$; p-value $= 0.2$). It also verifies that the number of basis functions (k) for the centroid diameter term is appropriate for this model. The k-index being close to 1 suggests that the model is effectively using the degrees of freedom available, indicating a good balance between fit and complexity. It points to an appropriate level of smoothness in the model's terms, contributing to its overall robustness and reliability. Additionally, a p-value greater than 0.05 indicates that the residuals align well with the model assumptions, meaning there are no significant deviations or patterns in the residuals that would indicate a poor fit for the variable *centroid diameter*. Collectively, the results from the gam.check() function affirm that the smoothing term has been chosen appropriately for the data.

5 Discussion

The (semantic) verbal fluency task is a popular diagnostic tool to determine cognitive decline. In the present study, we contrasted the canonically used scoring system based on the number of responses with automated scoring metrics derived from contextual word embeddings by means of semantic network analysis. This incorporates the notion of semantic similarity into the metrics. Our second goal was to test whether these semantic networks computed from Dutch verbal fluency data can be effectively applied to assess cognitive state.

Among our three proposed novel metrics, we have found that the centroid diameter distance is the most effective and only significant predictor of MMSE scores, which index cognitive health. Intuitively, this metric describes the average broadness of the cluster: it is defined as double the average distance between every item and the center of the network (Bolshakova and Azuaje, 2003). Our proposed metric is an automated, consistently determined score of the semantic verbal fluency task that was significantly more accurate in predicting MMSE scores than the traditional scoring. This suggests that semantic networks offer a more accurate way of scoring the verbal fluency task than mere tallying of correct responses. Second, it confirms a relationship between semantic networks and the cognitive state for the Dutch language.

The fact that we did not find evidence for an effect of the traditional scoring method is of course an interesting observation, as this method is widely used. However, it is worth mentioning that the verbal fluency task is a small part of the MMSE battery, and thus any metric based on this task cannot be expected to fully account for MMSE scores.

This could be a reason why we do not find a significant result for this metric. A previous study by Linz et al. (2017b) found that the semantic verbal fluency task (with the traditional word count metric) can predict MMSE scores. They incorporated multiple computed features in their regression models, among others word count. Before training their models, they considered the correlation between these features and MMSE scores, which concluded the correlation between word count and MMSE score to be 'moderate' on their scale. The feature combination that their model was trained on and that resulted in promising results, consisted of statistical clustering and switching, word frequency, and the vocal feature pause length. An interesting follow-up to the present research would be to examine the combination of metrics in the present study in an expanded multifactorial regression model incorporating these factors as controls.

Similarly, we did not find substantial results for the path length feature. We had hypothesized that this would be a good predictor next to the centroid diameter, since it does represent the span of one's semantic network in a way. It is also analogous to the forward flow measurement in creativity research, which is used to measure associative cognition. The semantic verbal fluency task could be argued to be an association task as well, though it would not be completely free association given the fact that the participant is bound to a semantic category. It could be that our path length feature therefore did not as accurately reflect "how far" participants traversed their semantic space.

Our findings indicate that we might want to reevaluate how we assess the verbal fluency task. Especially since it is such a widely used test, improving its accuracy will be beneficial. As a screening for Alzheimer's disease, the semantic verbal fluency task is even an attractive alternative to the MMSE (Chi et al., 2014). Adding automated and more accurate scoring to the mix makes it an even more appealing screening method. Furthermore, analyzing verbal fluency data through semantic networks could broaden the understanding of our semantic memory system and enhance our knowledge of handling tasks like categorical listing. This could be a good opportunity since there are many existing verbal fluency datasets to be reanalyzed. On top of that, exploring these datasets further through semantic networks might lead to additional, complementary scoring methods and therefore enhance such healthcare screenings. Another way of broadening the results would be through incorporating EEG data into the analysis, if available.

Given the promising results of this approach, it would be interesting to further explore language model-derived metrics for the verbal fluency task, potentially with other models as well. While we worked with Dutch data, for English there is a wider range of state-of-the-art models available to explore for this purpose. The fact that we observed these findings for Dutch also opens up possibility of making a cross-linguistic comparison of semantic networks. Lastly, our semantic network-based metrics could also be applied to other use cases, such as quantifying language acquisition rather than cognitive decline.

6 Conclusion

We explored the possible utility of semantic networks in evaluating cognitive health by examining this relationship in the Dutch language and reported the following findings: (a) Measurements of semantic networks, constructed from distributional semantic model BERTje, outperform the conventional method of counting words alone – specifically the centroid diameter distance metric. This approach can circumvent human error and judgment in the semantic verbal fluency task. Therefore, this study lays a foundation for improving diagnostic tools using computational methods; (b) There is a confirmed relationship between semantic networks and cognitive health for the Dutch language. To the best of our knowledge, this is the first study to test the relationship between semantic networks and cognitive state for Dutch. With evidence that semantic networks indeed relate to cognition for the Dutch language, this research paves the way for investigating Dutch semantic networks in other linguistic contexts. Moreover, this is the first study to use distributional semantic models to create semantic networks in this manner, thus providing an original contribution to research on linking the two. This suggests that distributional semantic models provide a new method for exploring the properties and structures of semantic networks.

7 Limitations

The presented results should be considered in the context of this study's limitations. As we have made use of the language model BERTje to determine the semantic similarity of word pairs, our results are only as good as BERTje's semantic rep-

resentations. We made use of the static embeddings in layer 0 since these hold no contextual information, just like our input data is generated without any context. However, the verbal fluency task is mildly contextual, as participants are asked to stick to a prompt. As BERTje can include such contextual information in layers later on, some relevant associations might be lacking in layer 0. The static embeddings may be too abstract – especially for words that tend to differ in meaning depending on the context. Contextual information could thus be helpful even in word-level tasks. Chronis and Erk (2020) addressed this by tuning models to identify the meaning of words. Their results confirm that static embeddings are unable to simultaneously surface every component of lexical semantic meaning and their embeddings preserve contextual information that is essential for some word-level tasks. They show that similarity estimation benefits from this contextual knowledge. Their research concerns English language models, but since the present study is centered around a word-level task and uses static embeddings, our results might improve with this same type of tuning for Dutch language models, or by providing the context of the verbal fluency task prompt and extracting contextual embeddings. Including the order of word generation might be another way to provide context.

Another limitation of the approach is that extracting embeddings from a large-ish model is far more resource-intensive than the previous metric of counting responses, and not available for very low-resourced languages.

References

Hyejin Ahn, Dahyun Yi, Kyungjin Chu, Haejung Joung, Younghwa Lee, Gijung Jung, Kiyoung Sung, Dongkyun Han, Jun Ho Lee, Min Soo Byun, et al. 2022. Functional neural correlates of semantic fluency task performance in mild cognitive impairment and alzheimer's disease: An fdg-pet study. *Journal of Alzheimer's Disease*, 85(4):1689–1700.

Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behav. Res. Methods*, 53(2):757–780.

Mathias Benedek, Caterina Mühlmann, Emanuel Jauk, and Aljoscha C Neubauer. 2013. Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity. *Psychol. Aesthet. Creat. Arts*, 7(4):341–349.

N Bolshakova and F Azuaje. 2003. Cluster validation techniques for genome expression data. *Signal Processing*, 83(4):825–833.

Javier Borge-Holthoefer and Alex Arenas. 2010. Semantic networks: Structure and dynamics. *Entropy (Basel)*, 12(5):1264–1302.

W A Bousfield and C H W Sedgewick. 1944. An analysis of sequences of restricted associative responses. *J. Gen. Psychol.*, 30(2):149–165.

Lizzy Brans and Jelke Bloem. 2024. SimLex-999 for Dutch. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14832–14845, Torino, Italia. ELRA and ICCL.

A S Chan, N Butters, J S Paulsen, D P Salmon, M R Swenson, and L T Maloney. 1993. An assessment of the semantic network in patients with alzheimer's disease. *J. Cogn. Neurosci.*, 5(2):254–261.

Yeon Kyung Chi, Ji Won Han, Hyeon Jeong, Jae Young Park, Tae Hui Kim, Jung Jae Lee, Seok Bum Lee, Joon Hyuk Park, Jong Chul Yoon, Jeong Lan Kim, Seung-Ho Ryu, Jin Hyeong Jhoo, Dong Young Lee, and Ki Woong Kim. 2014. Development of a screening algorithm for alzheimer's disease using categorical verbal fluency. *PLoS One*, 9(1):e84111.

Alexander P Christensen and Yoed N Kenett. 2023a. Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. *Psychol. Methods*, 28(4):860–879.

Alexander P Christensen and Yoed N Kenett. 2023b. Semantic network analysis (semna): A tutorial on preprocessing, estimating, and analyzing semantic networks. *Psychological Methods*, 28(4):860.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Glenn Clark, Paula M McLaughlin, Ellen Woo, Kristy Hwang, Sona Hurtz, Leslie Ramirez, Jennifer Eastman, Reshil-Marie Dukes, Puneet Kapur, Thomas P DeRamus, and Liana G Apostolova. 2016. Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimers Dement. (Amst.)*, 2(1):113–122.

Linda J Clark, Margaret Gatz, Ling Zheng, Yu-Ling Chen, Carol McCleary, and Wendy J Mack. 2009. Longitudinal verbal fluency in normal aging, preclinical, and prevalent alzheimer's disease. *Am. J. Alzheimers. Dis. Other Demen.*, 24(6):461–468.

Jin Cong and Haitao Liu. 2014. Approaching human language with complex networks. *Phys. Life Rev.*, 11(4):598–618.

Abigail L Cosgrove, Yoed N Kenett, Roger E Beaty, and Michele T Diaz. 2021. Quantifying flexibility in thought: The resiliency of semantic networks differs across the lifespan. *Cognition*, 211(104631):104631.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A dutch BERT model.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: A dutch RoBERTa-based language model.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Lang. Linguist. Compass*, 6(10):635–653.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context. In *Proceedings of the 10th international conference on World Wide Web*, New York, NY, USA. ACM.

Marshal F Folstein, Susan E Folstein, and Paul R McHugh. 1975. "mini-mental state". *J. Psychiatr. Res.*, 12(3):189–198.

C D Frith, K J Friston, S Herold, D Silbersweig, P Fletcher, C Cahill, R J Dolan, R S J Frackowiak, and P F Liddle. 1995. Regional brain activity in chronic schizophrenic patients during the performance of a verbal fluency task. *Br. J. Psychiatry*, 167(3):343–349.

Emily Frith, Daniel B Elbich, Alexander P Christensen, Monica D Rosenberg, Qunlin Chen, Michael J Kane, Paul J Silvia, Paul Seli, and Roger E Beaty. 2021. Intelligence and creativity share a common cognitive and neural basis. *J. Exp. Psychol. Gen.*, 150(4):609–632.

Joaquín Goñi, Gonzalo Arrondo, Jorge Sepulcre, Iñigo Martincorena, Nieves Vélez de Mendizábal, Bernat Corominas-Murtra, Bartolomé Bejarano, Sergio Ardanza-Trevijano, Herminia Peraita, Dennis P Wall, and Pablo Villoslada. 2011. The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. *Cogn. Process.*, 12(2):183–196.

Kurt Gray, Stephen Anderson, Eric Evan Chen, John Michael Kelly, Michael S Christian, John Patrick, Laura Huang, Yoed N Kenett, and Kevin Lewis. 2019. "forward flow": A new measure to quantify free thought and predict creativity. *Am. Psychol.*, 74(5):539–554.

Li He, Yoed N Kenett, Kaixiang Zhuang, Cheng Liu, Rongcan Zeng, Tingrui Yan, Tengbin Huo, and Jiang Qiu. 2021. The relation between semantic memory structure, associative abilities, and verbal and figural creativity. *Think. Reason.*, 27(2):268–293.

Julie D Henry, John R Crawford, and Louise H Phillips. 2004. Verbal fluency performance in dementia of the alzheimer's type: a meta-analysis. *Neuropsychologia*, 42(9):1212–1222.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist. Assoc. Comput. Linguist.*, 41(4):665–695.

Dan R Johnson, James C Kaufman, Brendan S Baker, John D Patterson, Baptiste Barbot, Adam E Green, Janet van Hell, Evan Kennedy, Grace F Sullivan, Christa L Taylor, Thomas Ward, and Roger E Beaty. 2023. Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behav. Res. Methods*, 55(7):3726–3759.

Jinyun Ke and Yao Yao. 2008. Analysing language development from a network approach. *J. Quant. Linguist.*, 15(1):70–99.

Yoed N Kenett and Miriam Faust. 2019. A semantic network cartography of the creative mind. *Trends in cognitive sciences*, 23(4):271–274.

Yoed N Kenett, Effi Levi, David Anaki, and Miriam Faust. 2017. The semantic distance task: Quantifying semantic distance with semantic network path length. *J. Exp. Psychol. Learn. Mem. Cogn.*, 43(9):1470–1489.

Najoung Kim, Jung-Ho Kim, Maria K Wolters, Sarah E MacPherson, and Jong C Park. 2019. Automatic scoring of semantic fluency. *Front. Psychol.*, 10:1020.

Alexandra König, Nicklas Linz, Johannes Tröger, Maria Wolters, Jan Alexandersson, and Phillipe Robert. 2018. Fully automatic speech-based analysis of the semantic verbal fluency task. *Dement. Geriatr. Cogn. Disord.*, 45(3-4):198–209.

Elles Konijnenberg, Stephen F Carter, Mara ten Kate, Anouk den Braber, Jori Tomassen, Chinenye Amadi, Linda Wesselman, Hoang-Ton Nguyen, Jacoba A van de Kreeke, Maqsood Yaqub, Matteo Demuru, Sandra D Mulder, Arjan Hillebrand, Femke H Bouwman, Charlotte E Teunissen, Erik H Serné, Annette C Moll, Frank D Verbraak, Rainer Hinz, Neil Pendleton, Adriaan A Lammertsma, Bart N M van Berckel, Frederik Barkhof, Dorret I Boomsma, Philip Scheltens, Karl Herholz, and Pieter Jelle Visser. 2018. The

EMIF-AD PreclinAD study: study design and baseline cohort overview. *Alzheimers. Res. Ther.*, 10(1).

Abhilasha A Kumar, Mark Steyvers, and David A Balota. 2022. A critical review of network-based and distributional approaches to semantic memory structure and processes. *Top. Cogn. Sci.*, 14(1):54–77.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.*, 104(2):211–240.

Alan J Lerner, Paula K Ogrocki, and Peter J Thomas. 2009. Network graph analysis of category fluency testing. *Cogn. Behav. Neurol.*, 22(1):45–52.

Nicklas Linz, Johannes Tröger, Jan Alexandersson, and Alexandra Konig. 2017a. Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *IWCS 2017 - 12th International Conference on Computational Semantics*, pages 1–7, Montpellier, France.

Nicklas Linz, Johannes Troger, Jan Alexandersson, Maria Wolters, Alexandra Konig, and Philippe Robert. 2017b. Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE.

Nicklas Linz, Johannes Tröger, Hali Lindsay, Alexandra Konig, Philippe Robert, Jessica Peter, and Jan Alexandersson. 2018. Language Modelling for the Clinical Semantic Verbal Fluency Task. In *LREC 2018 Workshop RaPID-2: Resources and ProcessIng of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments*, Miyazaki, Japan.

Nancy B Lundin, Joshua W Brown, Brendan T Johns, Michael N Jones, John R Purcell, William P Hetrick, Brian F O'Donnell, and Peter M Todd. 2023. Neural evidence of switch processes during semantic and phonetic foraging in human memory. *Proceedings of the National Academy of Sciences*, 120(42):e2312462120.

Evrim Gocer March and Philippa Pattison. 2006. Semantic verbal fluency in alzheimer's disease: approaches beyond the traditional scoring system. *J. Clin. Exp. Neuropsychol.*, 28(4):549–566.

Giampiero Marra and Simon N Wood. 2011. Practical variable selection for generalized additive models. *Comput. Stat. Data Anal.*, 55(7):2372–2387.

Israel Martínez-Nicolás, Instituto de Neurociencias de Castilla y León, Universidad de Salamanca, Salamanca, Spain, Juan Carro, Thide E Llorente, Juan José García Meilán, Departamento de psicología básica, psicobiología y metodología de las ciencias del comportamiento, Facultad de Psicología, Universidad de Salamanca, Salamanca, Spain, Instituto de Neurociencias de Castilla y León, Universidad de Salamanca, Salamanca, Spain, Departamento de psicología básica, psicobiología y metodología de las ciencias del comportamiento, Facultad de Psicología, Universidad de Salamanca, Salamanca, Spain, Instituto de Neurociencias de Castilla y León, Universidad de Salamanca, Salamanca, Spain, Departamento de psicología básica, psicobiología y metodología de las ciencias del comportamiento, Facultad de Psicología, Universidad de Salamanca, Salamanca, Spain, Instituto de Neurociencias de Castilla y León, Universidad de Salamanca, Salamanca, Spain, Departamento de psicología básica, psicobiología y metodología de las ciencias del comportamiento, Facultad de Psicología, Universidad de Salamanca, Salamanca, Spain, and Departamento de psicología básica, psicobiología y metodología de las ciencias del comportamiento, Facultad de Psicología, Universidad de Salamanca, Salamanca, Spain. 2019. The deterioration of semantic networks in alzheimer's disease. In *Alzheimer's Disease*, pages 179–191. Codon Publications.

Ana Maseda, Leire Lodeiro-Fernández, Laura Lorenzo-López, Laura Núñez-Naveira, Aránzazu Balo, and Jose C Millán-Calenti. 2014. Verbal fluency, naming and verbal comprehension: three aspects of language as predictors of cognitive impairment. *Aging Ment. Health*, 18(8):1037–1045.

Michelle McDonnell, Lauren Dill, Stella Panos, Stacy Amano, Warren Brown, Shadee Giurgius, Gary Small, and Karen Miller. 2020. Verbal fluency as a screening tool for mild cognitive impairment. *Int. Psychogeriatr.*, 32(9):1055–1062.

Ken McRae, Saman Khalkhali, and Mary Hare. 2012. Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy. In *The adolescent brain: Learning, reasoning, and decision making*, pages 39–66. American Psychological Association, Washington.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

A U Monsch, M W Bondi, N Butters, D P Salmon, R Katzman, and L J Thal. 1992. Comparisons of verbal fluency tasks in the detection of dementia of the alzheimer type. *Arch. Neurol.*, 49(12):1253–1258.

L Muhammad, E Garba, N Oye, and G Wajiga. 2019. Modeling techniques for knowledge representation of expert system: A survey. *J. Appl. Comput. Sci. Math.*, 13(2):39–44.

Yvette Oortwijn, Jelke Bloem, Pia Sommerauer, Francois Meyer, Wei Zhou, and Antske Fokkens. 2021. Challenging distributional models with a conceptual network of philosophical terms. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA. Association for Computational Linguistics.

A L Piatt, J A Fields, A M Paolo, W C Koller, and A I Tröster. 1999. Lexical, semantic, and action verbal fluency in parkinson's disease with and without dementia. *J. Clin. Exp. Neuropsychol.*, 21(4):435–443.

Davide Quaranta, Chiara Piccininni, Alessia Caprara, Alessia Malandrino, Guido Gainotti, and Camillo Marra. 2019. Semantic relations in a categorical verbal fluency test: an exploratory investigation in mild cognitive impairment. *Frontiers in psychology*, 10:2797.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Zeshu Shao, Esther Janse, Karina Visser, and Antje S Meyer. 2014. What do verbal fluency tasks measure? predictors of verbal fluency performance in older adults. *Frontiers in psychology*, 5:772.

Mark Steyvers and Joshua B Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78.

Angela K Troyer, Morris Moscovitch, and Gordon Winocur. 1997. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *neuropsychology*, 11(1):138.

Angela K Troyer, Morris Moscovitch, Gordon Winocur, Larry Leach, and Morris Freedman. 1998. Clustering and switching on verbal fluency tests in alzheimer's and parkinson's disease. *Journal of the International Neuropsychological Society*, 4(2):137–143.

Akira Utsumi. 2015. A complex network approach to distributional semantic models. *PloS one*, 10(8):e0136277.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4):847–897.

Simon N Wood. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36.

Simon N Wood. 2020. Inference and computation with generalized additive models and their extensions. *Test*, 29(2):307–339.

Dirk U Wulff, Simon De Deyne, Samuel Aeschbach, and Rui Mata. 2022. Using network science to understand the aging lexicon: Linking individuals' experience, semantic networks, and cognitive performance. *Topics in Cognitive Science*, 14(1):93–110.

Jeffrey C Zemla and Joseph L Austerweil. 2017. Modeling semantic fluency data as search on a semantic network. In *Cogsci... annual conference of the cognitive science society. cognitive science society (us). conference*, volume 2017, page 3646. NIH Public Access.

Jeffrey C Zemla and Joseph L Austerweil. 2018. Estimating semantic networks of groups and individuals from fluency data. *Computational brain & behavior*, 1:36–58.

Jeffrey C Zemla, Kesong Cao, Kimberly D Mueller, and Joseph L Austerweil. 2020. Snafu: The semantic network and fluency utility. *Behavior research methods*, 52:1681–1699.

Jeffrey C Zemla, Yoed N Kenett, Kwang-Sung Jun, and Joseph L Austerweil. 2016. U-invite: Estimating individual semantic networks from fluency data. In *CogSci*, volume 1, pages 36–58.

A Inclusion requirements

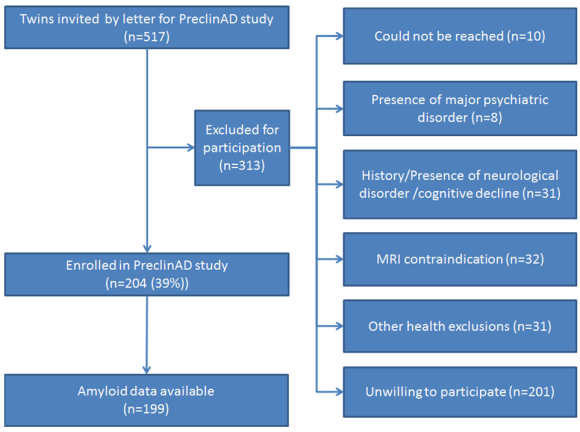

Figure 3: Inclusion flow chart for Amsterdam participants of the EMIF-AD PreclinAD study (Konijnenberg et al., 2018).