Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research report

Semantic fluency: Cognitive basis and diagnostic performance in focal dementias and Alzheimer's disease

Carlo Reverberi[a,*], Paolo Cherubini[a], Sara Baldinelli[b] and Simona Luzzi[b]

[a] Psychology Department, Università Milano – Bicocca, Milano, Italy
[b] Department of Clinical and Experimental Medicine, Polytechnic University of Marche, Ancona, Italy

ARTICLE INFO

Article history:
Received 9 August 2013
Reviewed 28 November 2013
Revised 28 November 2013
Accepted 6 February 2014
Action editor Michael Kopelman
Published online 22 February 2014

Keywords:
Semantic fluency
Multivariate analysis
Control functions
Executive functions
Semantic memory
Semantic dementia
Primary progressive aphasia
Fronto-temporal dementia
Alzheimer's Disease

ABSTRACT

Semantic fluency is widely used both as a clinical test and as a basic tool for understanding how humans extract information from the semantic store. Recently, major efforts have been made to devise fine-grained scoring procedures to measure the multiple cognitive processes underlying fluency performance. Nevertheless, it is still unclear how many and which independent components are necessary to thoroughly describe performance on the fluency task. Furthermore, whether a combination of multiple indices can improve the diagnostic performance of the test should be assessed.

In this study, we extracted multiple indices of performance on the semantic fluency test from a large sample of healthy controls ($n = 307$) and patients ($n = 145$) suffering from three types of focal dementia or Alzheimer's Disease (AD). We found that five independent components underlie semantic fluency performance. We argue that these components functionally map onto the generation and application of a search strategy (component 2), to the monitoring of the overall sequence to avoid repetitions (component 3) and out-of-category items (component 4), and to the full integrity of the semantic store (component 5). The integrated and effective work of all these components would relate to a "general effectiveness" component (component 1). Importantly, while all the focal dementia groups were equally impaired on general effectiveness measures, they showed differential patterns of failure in the other components. This finding suggests that the cognitive deficit that impairs fluency differs among the three focal dementia groups: a semantic store deficit in the semantic variant of primary progressive aphasia (sv-PPA), a strategy deficit in the non-fluent variant of primary progressive aphasia (nfv-PPA), and an initiation deficit in the behavioural variant of fronto-temporal dementia (bv-FTD). Finally, we showed that the concurrent use of multiple fluency indices improves the diagnostic accuracy of semantic fluency both for focal dementias and for AD. More generally, our study suggests that a formal evaluation of fine-grained patterns of performance would improve the diagnostic accuracy of neuropsychological tests.

[*] *Corresponding author.* Psychology Department, Università Milano – Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano, Italy.
E-mail address: carlo.reverberi@unimib.it (C. Reverberi).

http://dx.doi.org/10.1016/j.cortex.2014.02.006
0010-9452/© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

The semantic fluency test (Bousfield & Sedgewick, 1944) requires a subject to generate as many words as possible from a given semantic category (e.g., "fruits") within a limited time, usually 1 min. Semantic fluency is a simple yet powerful test. For clinical purposes, the test has the benefit of being very fast and easy to administer, while still being highly sensitive to detecting cognitive impairment and brain damage (Henry & Crawford, 2004). For research purposes, semantic fluency provides a unique window on the fundamental ability to explore and extract information from our semantic and lexical store.

Several cognitive functions may be involved in semantic fluency. First, when performing the task, subjects should generate and then follow a *strategy* to explore the semantic store efficiently. Typically, healthy subjects explore a semantic category (e.g., fruits) by exploiting its internal organisation. They tend to generate items by clustering them into subcategories, e.g., soft fruits, dry fruits, citrus fruits. Second, subjects need to *flexibly switch* between different subcategories or items and *select* a target item among alternative competitors. Third, subjects need to extract entries from semantic memory, and fourth, they need to monitor and check the output to avoid producing repetitions or items out-of-category. Finally, they need to keep an "active" state during task execution to cope with the limited amount of time available for production (Gruenewald & Lockhead, 1980; Reverberi, Laiacona, & Capitani, 2006; Rosen & Engle, 1997; Unsworth, Spillers, & Brewer, 2011; Wixted & Rohrer, 1994).

A deficit in any of the above-mentioned cognitive abilities may produce an impaired fluency performance. As a consequence, the raw performance in semantic fluency is not cognitively specific. Thus, it is unsurprising that an impairment on the semantic fluency task has been reported following damage to several brain structures: the left temporal lobe, for its role in semantic memory storage; the lateral frontal cortex, for its role in strategy generation, flexibility, and selection; and the medial frontal cortex, for its role in behavioural initiation and activation (Baldo & Shimamura, 1998; Henry & Crawford, 2004; Laisney et al., 2009; Robinson, Shallice, Bozzali, & Cipolotti, 2012; Stuss et al., 1998; Troyer, Moscovitch, Winocur, Alexander, & Stuss, 1998).

The cognitive opacity of the overall production score curtails the usefulness of the semantic fluency test in clinical settings and hinders the understanding of the cognitive and neural basis of semantic fluency. In recent years, several proposals were made for improving the specificity of the indices used to assess fluency performance, mainly by devising scoring procedures grounded on more detailed cognitive models of the task. Ideally, these new indices should be able to selectively measure each of the cognitive processes underlying semantic fluency. "Clustering" and "switching" are well-known examples of indices devised for this purpose (Moscovitch, 1992; Troyer, Moscovitch, & Winocur, 1997; Troyer et al., 1998). The clustering index is the average size of the produced subcategory clusters, while the switching index counts the number of times a subject switches from an old subcategory to a new subcategory. Clustering is thought to specifically measure the integrity of the semantic store, a measure related to the temporal lobe functions, while switching measures the integrity of strategic search processes, cognitive flexibility, and shifting, i.e., cognitive functions related to the frontal lobes (Troyer et al., 1998). Other laboratories have further contributed with alternative theoretical interpretations of these fluency indices by testing the new indices in several different pathological groups, or by proposing different indices, such as the "order index" and the "number of subcategories index" (e.g., Abwender, Swan, Bowerman, & Connolly, 2001; Fagundo et al., 2008; Ho et al., 2002; Mayr, 2002; Price et al., 2012; Reverberi et al., 2006; Troster et al., 1998).

Notwithstanding this major theoretical and empirical effort, whether these new indices indeed provide more information compared with the mere total number of new words still needs to be formally evaluated. Specifically, it remains unclear whether each of the above-mentioned indices measures one (or a few) underlying cognitive processes, whether the new indices are more accurate in distinguishing healthy controls from patients or in distinguishing among different patient groups, and whether the combined use of two or more indices can increase the amount of information conveyed by each index when used alone.

To further investigate and evaluate these issues, this study pursued three related aims. First, we wanted to assess how many and which independent components would be necessary to thoroughly describe semantic fluency performance. Second, we wanted to assess whether considering this larger set of fluency indices would allow for a better understanding of the cognitive basis of fluency impairment in three focal dementias. Third, we systematically evaluated nine fluency indices to understand whether any of them, alone or in combination with others, can better discriminate between patients and healthy controls and among different patient groups.

In this study, we considered two main patient groups: focal dementias and Alzheimer's Disease (AD). First, we focused on relatively rare but cognitively focal syndromes: semantic dementia, which has been recently reclassified as the semantic variant of primary progressive aphasia (sv-PPA; Gorno-Tempini et al., 2011); the behavioural variant of fronto-temporal dementia (bv-FTD); and non-fluent variant of primary progressive aphasia (nfv-PPA). These clinical syndromes constitute excellent models to study semantic fluency because they are characterised by a focal impairment on one or more of the cognitive functions involved in semantic fluency. Sv-PPA presents with a progressive loss of general semantic knowledge, involving multiple modalities (e.g., words, visual percepts, sounds, and tastes). Sv-PPA patients generally show language problems consisting of anomia and semantic paraphasias, which are associated with an impairment in understanding single words, in the absence of phonological and syntactical problems (Hodges, Patterson, Oxbury, & Funnell, 1992; Patterson, Nestor, & Rogers, 2007; Snowden, Goulding, & Neary, 1989; Warrington, 1975). Nfv-

PPA is a language disorder associated with atrophy of the perisylvian areas of the left hemisphere and is characterised by an impairment of the phonological and syntactical aspects of oral and written language, with sparing at the semantic level. Nfv-PPA patients show anomia with phonological paraphasia, agrammatism, and speech apraxia in the absence of single word comprehension deficits (Grossman, 2010; Neary et al., 1998). Bv-FTD refers to a disorder predominantly of behaviour, with the presence of a dysexecutive syndrome that is the sole problem in most cases. In a few cases, the behavioural and executive problems can be associated with impaired language, and mild semantic problems can sometimes be detected during the neuropsychological examination. The prototypical presentation of bv-FTD is with behavioural problems (from apathy to disinhibition), which involve both personal (e.g., loss of hygiene, hyperorality) and social (e.g., loss of social awareness) aspects. Formal tests of bv-FTD usually detect a dysexecutive syndrome characterised by perseverations, problems in categorisation, abstract thinking, and judgement, and loss of attention and poor working memory (Neary et al., 1998; Neary, Snowden, & Mann, 2005; Salmon et al., 2003; Snowden, Neary, & Mann, 1996). In addition to focal dementias, we also examined AD patients. AD patients have impairments in multiple cognitive functions (e.g., Snowden et al., 2011). Their study can provide limited information on the cognitive basis of semantic fluency. However, a formal assessment of the diagnostic power of alternative semantic fluency indices in AD would provide important clinical information for a more prevalent degenerative disease. Finally, in this study, all patient groups were compared with a large group of healthy controls ($n = 307$).

2. Methods

2.1. Subjects

A total of 455 subjects took part in the study (Table 1). Two main clinical groups formed the patient cohort, with each group having a different role in the study. The first clinical group comprised 70 individuals attending a specialist neurological clinic for early onset dementia; these individuals had a clinical diagnosis of bv-FTD (39 patients), semantic dementia (recently reclassified as sv-PPA, 15 patients), or nfv-PPA (16 patients). Six bv-FTD patients were excluded from the main analyses because they produced less than three valid items on the fluency test. The second clinical group comprised 78 patients diagnosed with AD. Three AD patients were excluded from the main analyses because they produced less than three valid items. The diagnosis was based on the clinical history and neurological and neuropsychological examinations, and it was further supported by structural and functional imaging. Routine blood screening tests excluded secondary causes of dementia. No patient had a history for cerebrovascular disease. All patients had been followed for at least one year, confirming the progressive nature of their disorder. Patients fulfilled the currently accepted diagnostic criteria for bv-FTD, sv-PPA, nfv-PPA or AD (Gorno-Tempini et al., 2011; McKhann et al., 2011; Neary et al., 1998, 2005; Rascovsky et al., 2011; Snowden et al., 1996). The fluency test was always administered during the first visit at the clinical service. Thus, most of the patients were in the relatively early stages of the disease. Clinical severity measures based on the Activities of Daily Living (ADL) scale (Katz, Downs, Cash, & Grotz, 1970) are reported in Table 1. An additional 307 healthy controls served as the reference group. Data from a subset of the control subjects were also used in previously published studies (Capitani, Laiacona, & Barbarotto, 1999; Capitani, Rosci, Saetti, & Laiacona, 2009). All participants gave their written informed consent, and the study was approved in accordance with the Helsinki Declaration.

Table 1 – Demographic information for all subject groups. Subjects not included in the main analyses (see Methods) have been removed.

	nfv-PPA	bv-FTD	sv-PPA	AD	CTL
N subjects	16	33	15	75	307
Age (M SD)	73.6 (3.4)	67.0 (6.1)	67.9 (6.5)	77.3 (6.2)	54.9 (17)
Education (M SD)	7 (4.6)	8.6 (4.4)	9.3 (4.9)	6.8 (4.1)	9.6 (5)
ADL (M SD)	95.5 (11.7)	88.9 (20.7)	100 (0)	91.1 (16.3)	–

nfv-PPA: non-fluent variant of Primary Progressive Aphasia; bv-FTD: behavioural variant of Fronto-Temporal Dementia; sv-PPA: semantic variant of Primary Progressive Aphasia; AD: Alzheimer's Disease; CTL: Control Group; ADL: Activities of Daily Living scale.

2.2. Procedure

The semantic fluency test was administered to all subjects on an individual basis. The subjects were instructed to produce as many different words as possible belonging to the category "fruit". One minute was granted. All words produced and their exact sequence were recorded. Phonological errors on otherwise clearly recognisable items were not recorded.

2.3. Scoring and experimental measures

For each fluency test, we computed the following indices:

(i) The number of new words produced in 1 min. Items out of the "fruit" category, and repetitions were excluded.
(ii) Repetitions: The number of repeated words.
(iii) Out-of-category words: The number items not belonging to the target category fruit: e.g., "salad" or "red".
(iv) Average familiarity. The familiarity index conveys, on a scale from 1 to 10, how common experience with a particular fruit is. Ten corresponds to the maximum familiarity (e.g., "apple") and one to the minimum familiarity (e.g., "kumquat"). We used the familiarity scores made available in a previous study by our group (Reverberi, Capitani, & Laicona, 2004). The average familiarity was computed over all items within the category fruit (out-of-category words were excluded).
(v) The number of subcategories (e.g., citrus fruits) to which the produced items belong. The subcategories were defined following the procedure described in our previous works on the topic (Reverberi et al., 2004, 2006). Overall, we considered 15 subcategories of fruits. The

subcategory index corresponds to the number of subcategories for which each subject produced at least one word.

(vi) The number of switches between subcategories. A switch was defined as any transition between two words that belonged to different subcategories.

(vii) The relative switching or average cluster size. The relative switching is the ratio of the number of switches to the total number of words generated minus 1, including repetitions. Although we report the relative switching index in the manuscript, the relative switching is roughly equivalent to the reciprocal of the average cluster size (see Supplementary Online Material in Reverberi et al., 2006). A cluster is defined as a group of words belonging to the same subcategory uttered sequentially. The cluster size is the number of words that form a cluster, counted starting from the second word of the series (Troyer et al., 1998).

(viii) The order index. In a fluency test, the number of switches, the number of subcategories, and the number of produced words are not independent. For example, the number of subcategories and the number of produced words together define the range of the possible number of switches. Thus, the comparability of the number of switches of different subjects could be undermined if these subjects presented different amounts of produced words and subcategories. The order index represents a way to make the number of switches from different participants comparable. This index is computed as the discrepancy between the theoretical maximum number of switches and the observed number of switches, divided by the theoretically admitted maximum minus the theoretically admitted minimum number of switches. The conceptual meaning of the order index is analogous to that of the relative switching/average cluster size. Specifically, the order index measures how semantically ordered a series of items is. Compared with clustering, the order index has the advantage of avoiding the bias that arises from the interdependence of the variables number of switches, number of subcategories, and number of produced words (Reverberi et al., 2006).

(ix) The average semantic proximity between each successive pair of produced words. We relied on the corpus of proximity scores made available for the category "fruit" and for the Italian language in previous studies by our group (Reverberi et al., 2004, 2006). The semantic proximity rates the similarity between a pair of fruits on a continuous scale from minimum of one (e.g., pinekernel and watermelon) to ten (e.g., fig and early fig). Conceptually, semantic proximity is similar to the order index and relative switching. Importantly, however, it has a finer grained level of description because both order index and relative switching dichotomise the real proximity of a pair of items, classifying it as relatively high (i.e., within a subcategory) or relatively low (i.e., between subcategories). By contrast, semantic proximity uses a continuous scale. Thus, for example, orange-lime and orange-tangerine are both pairs of fruits belonging to the same subcategory. However, the latter pair clearly has a higher semantic proximity than the former pair.

2.4. *Statistical analyses*

The analyses were divided into two main stages, each with partially different aims and involving different patient groups. In the first stage, the analyses focused on focal dementia syndromes (nfv-PPA, bv-FTD, and sv-PPA), which produce selective impairments in the cognitive functions involved in semantic fluency. The first stage of analyses aimed at clarifying (i) whether the fluency indices measure different cognitive processes, (ii) whether the fluency indices can discriminate patients from healthy controls and among the different focal patient groups, and (iii) whether a combination of more indices provides more information than the same indices used alone.

In the second stage, the analyses focused on AD. AD produces impairments in multiple cognitive functions; thus, it can provide only limited information on the cognitive basis of semantic fluency. On the other hand, AD is much more common than the other focal dementia syndromes. The second stage of analyses aimed at evaluating whether the combination of multiple fluency indices can improve the discriminability of AD compared with controls and the other focal dementia syndromes.

2.4.1. *Stage 1: analyses on focal dementias*

2.4.1.1. Severity of disease. The fluency test was always administered at the first visit to the clinical service; thus, the severity of the disease should be roughly comparable across groups. Additionally, the clinical severity at presentation was also evaluated by means of the ADL scale (Katz et al., 1970). We compared the average severity across groups. To further evaluate whether a relationship (irrespective of patient group) existed between disease severity and any of the fluency indices, we tested whether any fluency index correlated with the ADL across all patients. Given the ordinal nature of the ADL scale, we used the non-parametric Spearman rho.

2.4.1.2. Correlation and principal component analysis. The variables considered in this study are expected to show large correlations. There are two main reasons for this expectation. First, some of the variables determine the range of variation of the other variables. For example, the number of words and the number of subcategories determines the range (minimum and maximum) of the switching index. Second, some of the variables considered may measure (partially) overlapping sets of cognitive functions. For example, the order index, the relative switching and the average proximity should all measure the amount of semantic order in a list. The aim of the following analyses is to measure how much variance is shared across variables in a large sample of subjects. Furthermore, by using principal component analysis, we aim to identify both *how many* and *which* components are sufficient to describe performance on the semantic fluency task. In addition to evaluating the healthy subjects, this set of analyses was also applied to the patient group for two main reasons. First, the patient group acted as an independent replication of the healthy control group, allowing us to obtain an estimate of the reliability of the results. Second, and most critical, the variance exploited by the patient group PCAs is partially different. In healthy controls, the PCA only relies on the variability between subjects in performing the fluency task. In the patient group, besides this "natural" variability, the PCA also uses the large variability introduced by the different types of damage to the cognitive system (and, therefore, different fluency performance) associated with the three different focal syndromes (nfv-PPA, bv-FTD, and sv-PPA). Two measures grouped in the same component by the PCA on healthy subjects could be assigned to different components in a patient group because only one of the two measures is sensitive to the deficit present in that patient group, e.g., a semantic memory deficit. In contrast, the assignment of two variables to the same component in patients should be interpreted as strong evidence that the two variables measure the same underlying cognitive component.

We performed a PCA and evaluated the Pearson's linear correlations between all variables. The relevant number of components was determined by relying on the scree test criterion (Cattell & Vogelmann, 1977). The resulting components were rotated to maximise interpretability with oblique (oblimin) rotation. The PCA was performed using SPSS 20, while the correlation was calculated using MATLAB 2012b.

2.4.1.3. Group comparisons. First, all variables were preprocessed to remove variance due to differences in age and level of education. For each of the nine variables considered, we ran a linear regression analysis with the relevant variable as the dependent factor and with age and education as regressors. We estimated the regression coefficients for age and education by only considering healthy subjects to avoid any potential bias in the estimates due to brain damage. The regression coefficients were then used to compute the residuals for each variable and for all subjects. The average for each variable and each patient group was compared with the respective average in the control group by means of a two-sample *t*-test, as implemented in MATLAB 2012b. For each experimental question (e.g., "Is the average of any variable in this specific patient group different from that of healthy controls?"), we considered significant only those tests generating a *p*-value lower than .05, which were two-tailed and Bonferroni corrected for the number of comparisons performed for the experimental question under scrutiny. However, the tests composing the Bonferroni set are far from being mutually independent (see Tables 2 and 4 in the Supplementary Online Material), hence the applied Bonferroni correction is overly conservative in our case (e.g., Bland & Altman, 1995). Furthermore, the applied Bonferroni correction produces a low statistical power even for medium size effects in some groups. Thus, we reported the effect size, the *t* value, the non-corrected *p*-value and the Bonferroni-corrected alpha level for each group comparison. Observations that are only significant at a non-corrected alpha level may be discussed, but they are clearly singled out as exploratory observations that need to be confirmed by future observations.

We also performed the same-group comparisons described above on data from which we removed the variance due to the number of new words produced. The number of new words index shares a large proportion of variance with many other variables. Removing this non-specific source of variance may allow us to detect differential patterns specific to each pathological group. For example, the average familiarity score should increase in subjects producing fewer words (with all other factors remaining constant) because the most familiar items tend to be produced first. Thus, if the average familiarity remained high even after factoring out the new words' variance, then this finding would suggest a specific cognitive impairment affecting only rare items. Importantly, as in the case of age and education, the regression parameters were estimated using only the healthy control group. In this way, any correlation that is generated only by cognitive impairment would not bias the regression estimates.

Finally, to further ensure that our procedure effectively removed the effects of age and education, the same comparisons were also run after selecting a subgroup of healthy controls that best matched the patients for age and education (Supplementary Online Material).

2.4.1.4. Classification analysis. The first aim of these analyses was to determine which variables, either alone or in combination with other variables, would best allow us to classify a subject as being a patient or a healthy control. The second aim was to determine which variable, or combination of variables, provided the best information to determine the specific pathological group of each patient (i.e., nfv-PPA, bv-FTD, or sv-PPA).

As in the group comparisons, we removed the variance due to differences in age and education level (but see also the Supplementary Online Material for the classification analysis performed on matched groups). The implemented procedure is described in the Section 2.4.1.3. The residuals for all nine variables were submitted to two sets of classification analyses: the first considering patients against healthy subjects and the second considering the three groups of patients against each other. For the patients versus controls analysis, we ran a classification analysis for each of the 511 possible combinations of the nine variables. For each combination, we performed a cross-validated classification by means of a linear support vector machine (SVM), as implemented in LIBSVM 3.17 (Chang & Lin, 2011). We utilised a leave-one out cross validation: the subjects were split in two independent groups, one for training the classifier and one for testing it. The training group comprised all subjects ($n = 370$) except the one in the testing group ($n = 1$). The parameters of the linear SVM were estimated using only the subjects in the training group. Then, the estimated parameters were used to classify the remaining subject. The accuracy of the classification was then evaluated. The cross-validation procedure was repeated 371 times, leaving out a different subject during each iteration. Cross-validation allowed us to avoid overfitting the data and the consequent upward bias in accuracy estimation. Overall, we obtained one accuracy value per subject. The accuracy values in the patient group (sensitivity) and the healthy control group (specificity) were computed separately and then averaged into one single value. This two-step process was utilised to give equal weight to the performance of the classifier in both patients and controls, i.e., to equally weight the classifiers sensitivity and specificity. The average accuracy values were assigned to the specific combination of variables being evaluated. The whole procedure was repeated for each of the 511 combinations of variables. The 511 combinations of variables were split into 9 groups depending on the number of variables considered at the same time (range 1–9). Finally, the variable combinations in each group were sorted and evaluated for accuracy. A similar approach was applied for the second classification analysis, aimed at classifying patients in each specific pathological group. The only differences were that we only considered patients, and the classification included three groups (nfv-PPA, bv-FTD, and sv-PPA) instead of two (patients and controls). Similar analysis approaches are already used in related fields (Klöppel et al., 2008; Pereira, Mitchell, & Botvinick, 2009; Reverberi, Görgen, & Haynes, 2012a, 2012b).

To understand whether the accuracy values found in the classification analyses were different than what would be expected by chance alone, we ran a permutation test. The group labels of the subjects (e.g., "control" or "patient") were randomly permuted before replicating the exact same procedure as described above. The permutation procedure was repeated 2000 times, each time with a new random permutation. In this way, we obtained 2000 accuracy values for each of the 511 combinations of variables (1,022,000 accuracy values overall). As we did for the real data, the 511 combinations of variables were split into 9 groups. Finally, for each of the 2000 permutations, we identified the maximum performance in each of the 9 variable groups. This procedure allowed us to obtain the distribution of 2000 accuracy maxima for each variable group. We considered a finding on the real data to be significant if its accuracy was higher than the 95th percentile of the relevant variable group. Notably, our permutation procedure considered the small positive selection bias introduced by the choice of the best predicting variable set.

A further related question that we wanted to assess with a permutation test was whether the best combinations in a variable group granted a significant increase of accuracy compared with all variable groups relying on fewer variables. For each of the 2000 permutations, we computed the difference between the maximum accuracy in the target group and the maximum accuracy in all groups relying on fewer variables than the target group. This calculation provided us with the distribution of 2000 accuracy differences for each target group. We considered an accuracy difference in the real data to be significant if it was higher than the 95th percentile of the relevant target group.

2.4.2. *Stage 2: analyses on AD patients*

2.4.2.1. Severity of disease. The clinical severity at presentation was evaluated by means of the ADL scale in Alzheimer's patients (Katz et al., 1970). We compared the average severity between the AD and the focal dementia groups. To further evaluate the presence of a fixed relation (irrespective of patient group) between disease severity and any of the fluency indices, we evaluated whether any fluency index correlated with ADL across all patients.

2.4.2.2. Group comparisons. AD patients were compared with healthy controls using the same methodology applied for focal dementia syndromes (Section 2.4.1.3). As above, we ran two main sets of analyses comparing the averages of all indices in AD and in healthy controls: in the first analysis, we removed the variance due to differences in age and level of education from averages, while in the second analysis, we also removed the variance due to the number of new words produced. Because the regression parameters for age, education and number of words were only estimated using the healthy control group, the same corrections used for the other pathology groups were applied also to AD patients. The same comparisons were also run after selecting a subgroup of healthy controls that best matched the AD patients for age and education (Supplementary Material Online). Finally, the same procedure described in 2.4.1.3 was used to correct the alpha level for multiple comparisons.

2.4.2.3. Classification analysis. We ran a classification analysis on AD patients using the same procedure described for focal dementia patients. We aimed at determining which variables, either alone or in combination with other variables, would best allow us to classify a subject as being (i) an AD patient versus a healthy control and (ii) an AD versus a focal dementia patient.

3. Results

3.1. *Focal dementias*

3.1.1. *Clinical severity of disease*

The three clinical groups were not significantly different for ADL [$F(2,61) = 2.67$, $p = .08$]. Furthermore, ADL showed very low correlations with all fluency variables (mean of the absolute rho = .08; range: .002–.24). None of the correlations was significantly different from 0, also at $p < .05$ and not corrected for multiple comparisons.

3.1.2. *Correlation and principal component analysis*

Our first goal was to establish *how many* components are needed to describe performance on the semantic fluency task. An exploration of the matrix of correlation (Tables 2, 4, 5, and 6 in Supplementary Online Material) confirms that several variable pairs shared a large proportion of the variance. To better understand the component structure underlying this matrix of correlations, we performed a PCA considering all nine variables. The PCA was applied first to healthy controls and then to patients. To identify the number of components, we relied on the Cattell scree plot criterion. We thus generated the scree plot of the analysis (Fig. 1): a major elbow in the eigenvalue drop was found between four and five components, suggesting that four[1] would be the appropriate number of components to consider. Four components explained 86% of the total variance. The proportion of variance in each variable explained by the four components ranged from 65% for familiarity to 99% for out-of-category words. After an oblique

[1] Given the scree plot in Fig. 1, one may also consider only 2 components. However, in this case the proportion of variance explained for out-of-category words and repetitions would drop to <10%.

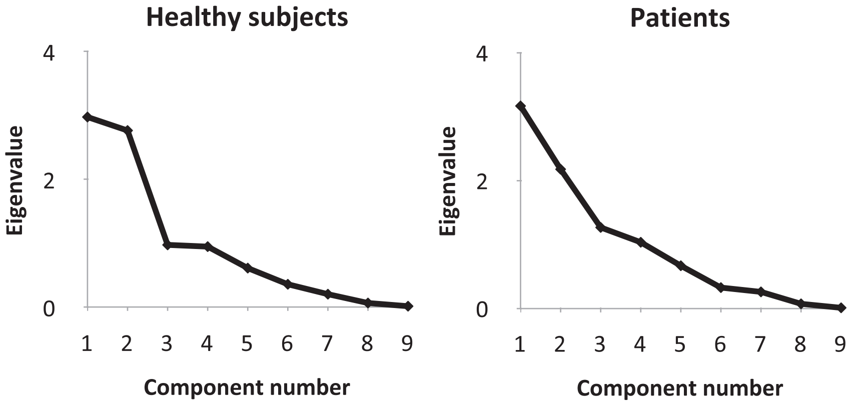

Fig. 1 – Principal component analysis. Scree plots for PCA of healthy subjects (left) and patients (right).

rotation, all variables showed a high loading on only one component (Table 2). Four variables showed high loadings (all >.75) on the first component: the number of new words, the number of new categories, the number of switches and familiarity. Three variables showed high loadings (all >.85) on the second component: proximity, relative switching, and order index. Finally, only one variable showed an extremely high loading on the third and fourth component (>.99), making these components equivalent to the corresponding variables. The correlation matrix between the rotated components shows overall low values (<.2), indicating that the rotated components are largely independent (Table 3).

Five components were identified by means of the Cattell scree plot criterion in patients (Fig. 1). The five components explained 92% of the total variance. The communalities ranged from 80% for proximity to 99% for out-of-category words. All variables showed a high loading on only one component (Table 4). Three variables had high loadings (all >.78) on the first component: the number of new words, the number of new categories, and the number of switches. Three variables showed high loadings (all >.84) on the second component: proximity, relative switching, and order index. The components from three to five showed a high loading on only one variable each. As in healthy controls, repetitions showed a high loading on the third component, while out-of-category words showed a high loading on the fourth component. Unlike in the healthy controls, familiarity was removed from the first component and assigned to an independent fifth component. The correlation matrix between the rotated components shows low values (<.23), indicating that the rotated components are largely independent (Table 5). We also tried to force the use of the same number of components that were identified in healthy subjects in patients. The overall variance that was explained dropped to 85%, mainly because of a reduction in the communality of familiarity and out-of-category words (respectively, .72 and .73). Interestingly, in this four-component model, familiarity was not assigned to the first component as in healthy subjects, but to the fourth, i.e., the same as out-of-category words.

Table 2 – Principal component analysis: variable loadings of each variable on each component in healthy subjects. Variable loadings higher than .75 are reported in bold.

	Components			
	1	2	3	4
Familiarity	**−.760**	−.172	.226	.077
N category	**.798**	−.125	.100	.025
Proximity	.132	**.894**	.077	.071
Switching	**.847**	−.377	.163	.018
Rel. switching	−.071	**−.968**	−.026	.037
Order	−.192	**.867**	−.051	.017
New	**.920**	.256	.024	−.018
Repetitions	−.006	.091	**.994**	−.037
Out	.001	.036	−.036	**.993**

3.1.3. *Comparisons across groups*

We compared the average fluency performance of each pathological group with the performance of healthy controls for each of the nine variables considered. Before running the comparisons, we removed the effects of age and level of education from all comparisons except for those involving the variable out-of-category words. Too few healthy controls produced enough out-of-category words ($n = 4$) to allow a reliable estimate of the regression parameters.

All pathological groups significantly differed from controls on most variables considered (Table 6 and Table 7 in Supplementary Online Material). The only cases in which no differences were found, even at $p < .05$ uncorrected, were repetitions in the nfv-PPA group and order index, repetitions, and out-of-category words in the sv-PPA group.

Next, we explored whether the difference between pathological groups and control subjects on any variable would remain significant even when the effect of the overall reduction in word production was considered. The effect of word production level on other variables was estimated by relying on healthy subjects. These estimates were then used to

Table 3 – Principal component analysis: component correlation matrix in healthy subjects.

Component	1	2	3	4
1	1	.014	.129	−.135
2	.014	1	−.194	.054
3	.129	−.194	1	−.015
4	−.135	.054	−.015	1

Table 4 – Principal component analysis: variable loadings of each variable on each component in patients with focal degenerative syndromes. Variable loadings higher than .75 are reported in bold.

	Components				
	1	2	3	4	5
Familiarity	−.020	.021	−.012	−.029	**.995**
N category	**.913**	−.072	.166	.184	.077
Proximity	.100	**.913**	−.124	.007	.099
Switching	**.787**	−.253	−.333	.031	−.032
Rel. switching	.038	**−.939**	.068	.008	.129
Order	−.101	**.842**	.245	.038	.018
New	**.941**	.190	.018	−.148	−.138
Repetitions	−.043	.038	**−.982**	.065	.001
Out	.009	.041	−.044	**.990**	−.028

Table 5 – Principal component analysis: component correlation matrix in patients with focal degenerative syndromes.

Component	1	2	3	4	5
1	1.000	−.141	.161	−.149	−.224
2	−.141	1.000	−.062	.130	−.074
3	.161	−.062	1.000	−.168	.162
4	−.149	.130	−.168	1.000	.091
5	−.224	−.074	.162	.091	1.000

regress out the effect that the number of words had on the other variables, both in healthy subjects and in patients. The resulting scores for each pathological group were then compared with the control group. Several effects were no longer significant (see Table 7 *vs* Table 6, see also Table 8 in Supplementary Online Material). Interestingly, the variables that remained different from healthy controls were different in each pathology group. Nfv-PPA patients produced disorganised sequences (low order index). Bv-FTD patients produced more repetitions. Sv-PPA patients did not show any difference at a corrected alpha level.

Finally, we formally assessed whether the nine variables produced a pattern that was significantly different across pathology groups. We thus ran a 9 × 3 ANOVA with factors variable (9 levels) and pathology group (3 levels). The main effect of variable was significant [$F(8,488) = 36.8$, $p < .001$] while the main effect of group was not [$F(2,61) = .52$, $p = .59$]. Most critically, however, we found a significant interaction variable × group [$F(16,488) = 1.88$, $p = .02$]. This finding means that the pattern of failure of the pathology groups was dissimilar across variables.

3.1.4. *Classification analysis*

In a third set of analyses, we investigated which variables, either alone or in combination with the others, would best allow us to classify whether a subject is a patient or a healthy control. Furthermore, we investigated which variables would best allow us to assess which pathological group a patient belongs to. We systematically explored all possible combinations of the nine variables. The best classification performance for patients versus healthy subjects was found when four variables were considered (Fig. 2). This combination of four variables had an accuracy level that was significantly higher than chance; nevertheless, it was not significantly higher than the accuracy level reached by a classifier using fewer variables ($p = .057$). The combinations of variables present in the best performing classifier were: number of categories, average proximity, new words, and out-of-category words (Table 9, Supplementary Online Material). Notably, the classifier relying only on the number of new words variable showed a remarkable performance, 78%, which was only 6% less than the maximum observed.

We then systematically explored which combinations of the nine variables considered would best perform in classifying patients in their specific pathology group. Three alternative groups were possible (nfv-PPA, bv-FTD, and sv-PPA), corresponding to a chance accuracy level of 33%. The best accuracy (58.3%) was found with two classifiers, one using 5 and the other using 6 variables. The five variables involved in the former classifier are: familiarity, switching, order index, new words, and repetitions. The classifier using 6 variables included out-of-category words in addition to the five previously mentioned variables (Table 9, Supplementary Online Material). The performance of the two classifiers was significantly higher than chance ($p < .05$). Classifiers using one to

Table 6 – Comparisons between focal patient groups and controls for each of the nine variables describing fluency performance. The effect of age and education was removed from all of the comparisons but out-of-category words. For each comparison we report the effect size (ES, Hedges' g), the *t* value, and the uncorrected *p*-value of the two-sample *t*-tests. Degrees of freedom for nfv-PPA, bv-FTD and sv-PPA were respectively 321, 338 and 320. *p*-values <.05 uncorrected for multiple comparisons are reported in bold, *p*-values <.05 corrected for multiple are underlined (Bonferroni critical threshold .005, *n* = 9).

	nfv-PPA			bv-FTD			sv-PPA		
	ES	*t value*	*p value*	ES	*t value*	*p value*	ES	*t value*	*p value*
Familiarity	.67	2.62	**.009**	.85	4.66	**<.001**	1.2	4.49	**<.001**
N category	−.86	−3.37	**.001**	−1.3	−6.94	**<.001**	−1	−3.87	**<.001**
Proximity	−.92	−3.6	**<.001**	−.45	−2.44	**.015**	−.92	−3.48	**.001**
Switching	−.65	−2.52	**.012**	−1	−5.72	**<.001**	−1	−3.78	**<.001**
Rel. switching	.93	3.65	**<.001**	.64	3.48	**.001**	.79	2.99	**.003**
Order	−.85	−3.31	**.001**	−.39	−2.12	**.034**	−.31	−1.16	.246
New	−1.1	−4.15	**<.001**	−1.6	−8.82	**<.001**	−1.4	−5.45	**<.001**
Repetitions	.48	1.88	.061	1	5.52	**<.001**	−.19	−.703	.483
Out	1.1	4.41	**<.001**	1.1	5.86	**<.001**	−.13	−.497	.62

Table 7 – Comparisons between focal patients groups and controls for each of the nine variables describing fluency performance. Besides the effect of age and education, the effect of the overall number of new words was also removed from all of the comparisons. For each comparison we report the effect size (ES, Hedges' g), the *t* value, and *p* value of the two-sample *t*-tests. Degrees of freedom for nfv-PPA, bv-FTD and sv-PPA were respectively 321, 338 and 320. *p*-values < .05 uncorrected for multiple comparisons are reported in bold, *p*-values < .05 corrected for multiple comparisons are underlined (Bonferroni critical threshold .007, *n* = 7).

	nfv-PPA			bv-FTD			sv-PPA		
	ES	*t value*	*p value*	*ES*	*t value*	*p value*	*ES*	*t value*	*p value*
Familiarity	.095	.373	.709	−.03	−.18	.859	.47	1.78	.077
N category	−.24	−.952	.342	−.32	−1.78	.076	−.13	−.496	.62
Proximity	−.48	−1.9	.059	−.005	−.03	.98	−.54	−2.07	**.04**
Switching	.39	1.51	.133	.25	1.38	.168	−.047	−.177	.86
Rel. switching	.66	2.58	**.01**	.25	1.4	.163	.4	1.52	.129
Order	−.82	−3.21	**.001**	−.35	−1.93	.054	−.27	−1.04	.298
Repetitions	.55	2.15	**.032**	1.11	6.07	<**.001**	−.09	−.34	.734

four variables had a performance that was not significantly higher than chance (Fig. 2).

3.2. *Alzheimer's Disease*

3.2.1. *Clinical severity of disease*

The ADL index was not different between the AD and the other focal patients [t(137) = .71, p = .48]. The ADL showed very low correlations with all fluency variables (mean of the absolute rho = .08; range: .003–.14). None of the correlations were significantly different from 0, also at $p < .05$ not corrected for multiple comparisons.

3.2.2. *Comparisons across groups*

We compared the average fluency performance of AD patients with that of healthy controls for each of the nine variables considered. AD patients differed from controls on all variables considered except the number of switches (Table 8). Next, we performed the same comparisons when the effect of the overall reduction in word production was taken into account. By contrast with the focal dementia patients, several effects remained significant even after correcting for the number of words (Table 8).

Table 8 – Comparisons between AD patients and controls. In the column "AD cov. words", besides removing the effect of age and education as in the "AD" column, we also removed the effect of the overall number of new words. For each comparison we report the effect size (ES, Hedges' g), the *t* value, and the uncorrected *p*-value of the two-sample *t*-tests. Degrees of freedom are 380, for both AD and AD cov. words. *p*-values < .05 uncorrected for multiple comparisons are reported in bold, *p*-values < .05 Bonferroni corrected for multiple comparisons are underlined (Bonferroni critical *p*-value threshold left column is .005; *n* = 9; right column .007, *n* = 7).

	AD			AD cov. words		
	ES	*t value*	*p value*	*ES*	*t value*	*p value*
Familiarity	.51	3.966	<**.001**	.15	1.167	.244
N category	−.62	−4.791	<**.001**	−.23	−1.779	.076
Proximity	−.45	−3.491	<**.001**	−.26	−2.036	**.042**
Switching	−.24	−1.848	.065	.35	2.71	.**007**
Rel. switching	.55	4.309	<**.001**	.37	2.853	**.005**
Order	−.59	−4.563	<**.001**	−.57	−4.444	<**.001**
New	−.72	−5.584	<**.001**	–	–	–
Repetitions	.71	5.529	<**.001**	.75	5.830	<**.001**
Out	.70	5.464	<**.001**	–	–	–

3.2.3. *Classification analysis*

The best classification performance for AD patients versus healthy subjects was found when four variables were considered (Fig. 2). This combination of four variables reached an accuracy level of 72%, which is significantly higher than chance and significantly higher than the accuracy level reached by a classifier using fewer variables. The variables in the best performing classifier were number of categories, order index, new words, and repetitions (Table 12, Supplementary Online Material). In contrast with focal dementia patients, the classifier only relying on the number of new words variable showed an accuracy level much lower than the best four variables' model: only 61%. We then systematically explored which combinations of the nine variables considered would best discriminate AD patients from focal dementia patients. The best accuracy (72%) was found for a classifier using four variables. However, this performance was not significantly higher than that of a classifier only using the number of words. The latter classifier was able to reach an accuracy of 71%, which is significantly higher than chance alone (Table 12, Supplementary Online Material).

4. Discussion

Semantic fluency is a simple task that evaluates the ability to efficiently explore the semantic system. In this study, we had three main goals. First, we wanted to better understand the cognitive basis of the semantic fluency. Second, we wanted to explore the causes of the fluency impairment in focal dementias. Third, we wanted to evaluate whether any of the proposed fluency indices, either alone or in combination with others, would discriminate patients from healthy controls and assign each patient to his/her pathological group.

We found that to thoroughly describe semantic fluency performance at least four independent components are needed. We found theoretically relevant qualitative differences in fluency performance across pathological groups.

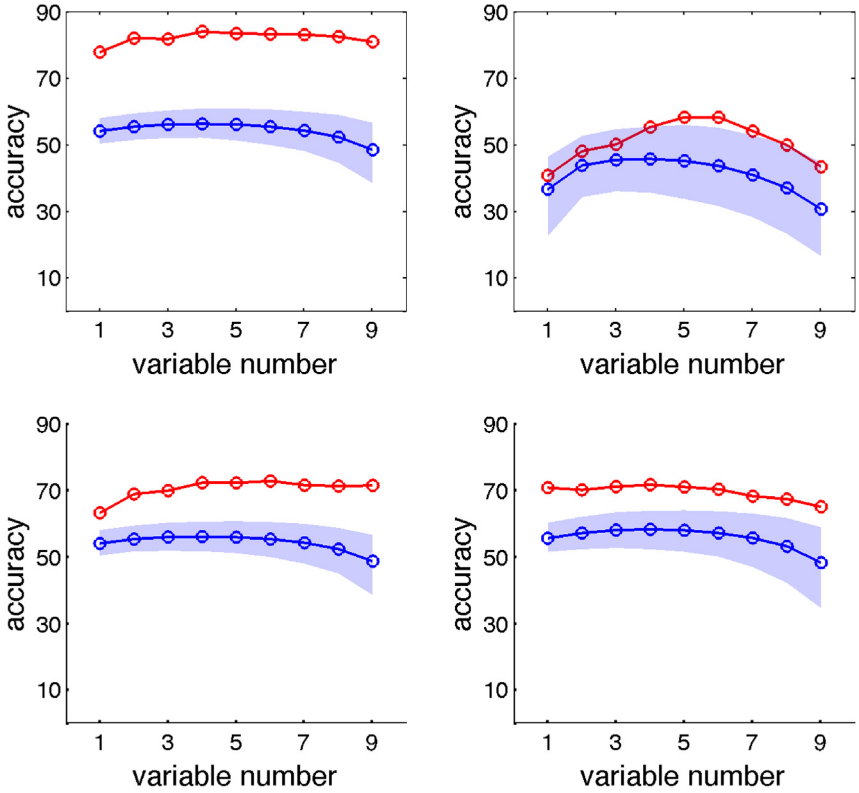

Fig. 2 – Classification analysis. On the top left, a classification of focal patients versus controls. On the top right, a classification of focal patients in each specific pathological group. On the bottom left, a classification of AD patients versus controls. On the bottom right, a classification of AD patients versus focal patients. Red: real data. Blue: permutation data (mean and 5th–95th percentiles).

Finally, we found that combinations of multiple variables can improve the classification accuracy of semantic fluency.

4.1. *The cognitive basis of semantic fluency*

Semantic fluency is thought to involve at least two cognitive abilities: accessing a fully functional semantic store and engaging the executive functions that enable an efficient search into the semantic store. These abilities have been roughly mapped onto the left temporal and the left frontal lobes (Baldo, Schwartz, Wilkins, & Dronkers, 2006; Bousfield & Sedgewick, 1944; Costafreda et al., 2006; Gruenewald & Lockhead, 1980; Henry & Crawford, 2004; Hirshorn & Thompson-Schill, 2006; Katzev, Tüscher, Hennig, Weiller, & Kaller, 2013; Reverberi et al., 2006; Robinson et al., 2012; Troyer et al., 1998). While our findings are compatible with this general partition, they further suggest that the independent cognitive abilities should be expanded (Reverberi et al., 2006; Rosen & Engle, 1997; Unsworth et al., 2011). The number of independent components was assessed by two principal component analyses (PCA), one in healthy controls and one in patients. The PCA in the healthy group identified four components, while the PCA on patients identified five components that were characterised by loadings that were highly similar to those found in healthy subjects (Tables 4 and 2). These convergent findings clearly show that four, possibly five, components should be used to evaluate the performance on semantic fluency. Thus, the next step is to understand which cognitive abilities are measured by the identified components.

The first PCA component grouped all variables related to the main outcome variable, i.e., the number of new words produced. The new words count is a cognitively non-specific variable, measuring all functions contributing to the effectiveness of the search in the semantic store, from the integrity and size of the semantic store itself, to the basic processing speed, to “working memory” or “fluid intelligence” (Roca et al., 2010; Rosen & Engle, 1997; Unsworth et al., 2011). Given our results, this cognitive complexity should be extended to all other variables assigned to the same component, namely the number of categories and the number of switches. Consistent with this interpretation, the variables belonging to this component are impaired in all pathology groups.

The second component mainly involved variables that measured the generation and implementation of a search strategy that exploits the structure of the semantic store (Reverberi et al., 2006). Consistent with this interpretation, we found that the order index is specifically impaired in one of the frontal groups, i.e. the nfv-PPA patients, who are expected to show strategy deficits (see 4.2), but not in other patient groups. The same pattern was found for the relative switching index when the variance due to the number of words generated (first component) was removed. The average proximity shares a large part of its variance with the order index and the relative switching index, both in controls and patients. Nevertheless, proximity also shows a slightly different pattern than order index and relative switching in the group comparisons (Table 7). We speculate that, while semantic proximity is sensitive to strategy deficits, it is also sensitive to variations in the "semantic microstructure" of a sequence (see 4.2 for discussion).

The third component has a straightforward interpretation because it only included one variable: the number of repetitions. Thus, the ability to keep track of the uttered words to avoid repetitions depends on (partially) different cognitive functions than the other abilities involved in semantic fluency. Only patients with a degeneration involving the frontal lobe, particularly the bv-FTD patients, showed a significant increase in the number of repetitions, while the increase was completely absent when the degeneration only involved the temporal lobe (sv-PPA).

The fourth component only included the variable out-of-category words. Thus, the ability to represent and comply with the main general rule of the task (i.e., "fruits only") should be independent from other cognitive abilities. Nevertheless, unlike component 3, we are more careful in arguing the independency of this measure because this variable was extremely sparse, mostly in controls but also in focal patients. The observed independence was possibly due to the sparseness of the variable and not to the real independence of the measured cognitive abilities. More data are necessary to assess this possibility. Regardless of independence, this measure was specifically sensitive to degeneration involving the frontal lobes but not the temporal lobes.

The fifth component was only present in patients and included average familiarity, a variable that was assigned to the first component in controls. Thus, the patient PCA showed that familiarity is independent from the effectiveness variables (component 1). Familiarity is specifically sensitive to the ability to produce rare items. This ability is usually highly correlated with the number of uttered words: the more words you utter, the lower the average familiarity. However, if we hypothesise that damage to the semantic store would mostly affect rarer items, then the familiarity index could be disproportionately impaired. Thus, the correlation of the familiarity index with the number of uttered words would be weakened. This possibility is consistent with the features of the semantic store impairment in patients with sv-PPA (Hodges, Graham, & Patterson, 1995). In the present study, the sv-PPA group was the only group showing a trend (not corrected for multiple comparisons) in having an average familiarity higher than that expected on the basis of the number of words (Table 7). However, this speculation needs to be confirmed by further independent evidence.

Overall, the PCAs and across group comparisons showed that semantic fluency involves multiple cognitive abilities that can be independently measured by several indices: the generation and application of a search strategy based on subcategories (component 2), the monitoring of the overall sequence to avoid repetitions (component 3), the monitoring of each word to avoid producing out-of-category items (component 4), and the full integrity of the semantic store (component 5). The integrated and effective work of all these cognitive abilities can be measured by the general effectiveness variables (component 1).

The identified cognitive components of semantic fluency are in good agreement with those proposed in preceding studies on semantic fluency (Rosen & Engle, 1997; Unsworth et al., 2011). In contrast, our findings prompt a revision on how to measure these components. It has been proposed (Troyer et al., 1998) that the switching index specifically measures the integrity of executive functions (i.e., strategic search processes, cognitive flexibility, and shifting), while the clustering index (i.e., relative switching in this study) specifically measures the integrity of the semantic store. Our findings suggest that the switching index should be better conceived as a more general index of effectiveness of performance (component 1), which is related to the integrity of multiple cognitive functions (Mayr, 2002; Reverberi et al., 2006). Consistently, the switching index was impaired in all patient groups, irrespective of whether executive functions were spared. Furthermore, the clustering index (i.e., relative switching) was not sensitive to semantic impairment. In contrast, it was sensitive to a strategy deficit, causing a production of disorganised sequences. Consistently, both relative switching and order index were impaired in one of the patient groups with degeneration affecting the frontal lobes (nfv-PPA), but not in the patient group with a damaged semantic store (see also Reverberi et al., 2006).

4.2. *Semantic fluency performance in focal dementias*

The pattern of failure on semantic fluency over the nine variables is different across the three types of focal dementia considered in this study, as the interaction group × variable shows. All three focal patient groups are impaired in the variables comprising the first effectiveness component. Importantly, however, when the variance associated with the number of new words (i.e., the effectiveness component) is removed (Table 7), the variables associated with the other components show a pattern of failure that is largely specific for each pathology group.

Sv-PPA causes a selective deficit in the semantic store (Hodges et al., 1992; Patterson et al., 2007; Warrington, 1975). In addition to showing a profound impairment in the number of new words generated and the other related variables (component 1), sv-PPA patients only showed a specific impairment on the semantic proximity index (Table 7, exploratory finding). We speculate that this impairment highlights a damage to the semantic "microstructure" of the category fruit. For a sv-PPA patient, a citrus fruit is still clearly different from a dry fruit, but more fine-grained differences within the subcategory may blur (e.g., the different semantic distance between orange-tangerine and orange-lime). However, a sv-PPA patient is perfectly able to use information at the subcategory level to cluster items from the same subcategory, as witnessed by the spared order index. This finding is consistent with preceding literature on sv-PPA showing that the natural history of the disease begins with an impairment at the lower levels of the semantic hierarchy, i.e., the level allowing distinctions between very similar items (Hodges et al., 1995; Rogers et al., 2004; Warrington, 1975). Finally, in striking contrast with the other two pathology groups, sv-PPA patients did not produce either more out-of-category words or more repetitions. Together, these observations confirm that all strategic and monitoring (i.e., "frontal") functions are spared in patients with sv-PPA, even when used to explore a damaged semantic system (Patterson et al., 2007).

The diagnosis of nfv-PPA is based on the selective impairment of the phonological and syntactical aspects of language. The associated degeneration mainly affects the left inferior frontal gyrus, i.e., Brodmann Areas (BA) 44/45/47 (Gorno-Tempini et al., 2004, 2011; Grossman, 2010; Neary et al., 1998). In this study, nfv-PPA patients showed a specific deficit on the strategic (e.g., order index) and monitoring (e.g., out-of-category words) aspects of semantic fluency task (Tables 6 and 7). Nfv-PPA patients did not cluster together fruits belonging to the same category; rather, they jumped between subcategories. This observation is consistent with the main locus of neural degeneration in patients with nfv-PPA, namely BA 45. Left lateral frontal lesions can produce an impairment to semantic fluency (Henry & Crawford, 2004; Stuss et al., 1998). More specifically, a previous study by our group found that patients with a focal lesion in the lateral prefrontal cortex produce disorganised sequences, thus suggesting a deficit in the strategic search within the semantic store (Reverberi et al., 2006; see also Stuss et al., 1994). Consistent with this hypothesis, recent neuroimaging studies have shown that BA 45 specifically activates during semantic but not phonemic fluency (Costafreda et al., 2006; Katzev et al., 2013). This finding can be linked to the proposed role of BA 45 in controlled semantic retrieval (Badre & Wagner, 2007; Katzev et al., 2013).

Bv-FTD patients are mainly characterised by behavioural problems, including apathy, disinhibition, and loss of social awareness (Neary et al., 2005). The degeneration mainly affects the frontal median regions (Salmon et al., 2003). In the present study, bv-FTD patients showed an extremely low production of new words in the context of a spared semantic structure of the sequences produced (Table 7). This pattern is compatible with the idea that the impaired fluency performance of bv-FTD patients is due to a deficit in initiation (Burgess & Shallice, 1996; Mayr, 2002; Reverberi et al., 2006). A pure initiation deficit should cause a reduced rate of word production. On the other hand, an initiation deficit should leave unaffected the semantic structure of the sequences produced, both at the macro and micro level. The presence of an initiation deficit in bv-FTD patients is also consistent with previous studies showing that patients with frontal medial lesions have a deficit in initiation both in semantic fluency (Reverberi et al., 2006) and in other types of fluency tasks (Robinson et al., 2012). In addition to the initiation deficit, bv-FTD patients shared an impaired ability to monitor the produced sequence for repetitions and out-of-category words with nfv-PPA patients. Interestingly, these two variables were specifically sensitive to damage in frontal lobe structures, but they did not clearly distinguish between pathology groups affecting different frontal regions. This finding suggests that these two variables measures multiple cognitive functions localised in different regions (e.g., inhibition of prepotent responses, representation of task rules and constraints) or, alternatively, that they measure the impairment of a general "multi-purpose" mechanism with a distributed neural basis (e.g., Duncan, 2010).

Overall, our study showed that sv-PPA, bv-FTD and nfv-PPA patients produce a pattern of failure on semantic fluency with both common and distinct aspects. The former are mainly related to the general effectiveness in coping with the task needs (measured by the number of words, the number of subcategories explored, and the number of switches), which is impaired in all groups. The distinct aspects are related to the use of subcategories for a strategic search, the ability to monitor for repetitions and out-of-category words, the sensitivity to nuances that differentiate items within a subcategory ("micro" semantic effects), and the production of rare items. All three pathology groups fail in semantic fluency but for different reasons that can emerge when measured with more specific behavioural indices.

4.3. *Semantic fluency performance in AD*

Alzheimer's patients showed impairments on all fluency indices but the switching index (Table 8). The switching index was most likely spared due to the balancing between a general decline in effectiveness (lowering the switching index) and an associated strategy deficit (increasing the switching index). A wide cognitive impairment was also confirmed when the general effectiveness component was removed (Table 8). These findings are consistent with the known deficit to multiple cognitive functions produced by AD (e.g., Snowden et al., 2011).

4.4. *Semantic fluency as a diagnostic test*

We showed that at least four independent components are necessary to fully describe semantic fluency performance. This finding suggests that the diagnostic performance of the fluency task can be improved by the concurrent consideration of several variables tapping different components. The classification analyses in this study formally show that this improvement is indeed attainable.

For focal dementia syndromes, semantic fluency showed a remarkable performance in distinguishing patients from healthy controls: the sensitivity was 86%, and the specificity was 82%. This performance is even more remarkable when one considers that the test is administered in approximately 1 min, and the patients are in the early stages of their pathology. Our study shows that the maximum classification accuracy can be attained by integrating the information of four variables. However, the accuracy of the classifier using four variables only granted a relatively minor and non-significant improvement (6%) over the simpler classifier using only the number of new words. Thus, a general measure of how effective the search of the semantic store was already captured most of the variance relevant for classifying a subject as healthy or cognitively impaired. In sharp contrast, our findings for the differential diagnosis in focal patients showed that the use of multiple variables is, in this case, mandatory. Indeed, at least five variables were necessary to obtain a performance that was higher than chance. With the best model available, the average performance was 58%, chance level being 33%. On one hand, these findings show for the first time that considering multiple features of the semantic fluency performance allows one to distinguish patients with an impairment due to different causes, while single variables cannot (as witnessed in Henry & Crawford, 2004). On the other hand, the findings place an upper bound on the performance attainable by using only the semantic fluency task, and this bound is clearly suboptimal for clinical purposes. In addition, the best model with five variables almost perfectly maps onto the PCA components on patients. This finding further supports the suggested identification of five independent components in semantic fluency.

Classification performance for patients with AD produced a pattern that was symmetrical to that of focal patients. Considering multiple fluency indices was unnecessary for distinguishing between AD patients and focal patients: a model only relying on the number of the produced words already reached the best classification accuracy attainable on the basis of the fluency test (70%, chance level being 50%). This finding is probably due to the fact that AD patients behaved differently from the focal groups on new word and related indices, while they behaved similarly on the other fluency indices. By contrast, for classifying AD compared with healthy controls, the use of multiple fluency indices granted a better accuracy. This finding is likely related to the fact that AD patients are severely impaired on multiple fluency indices, even after the removal of the effect due to the number of new words. The extra information available in other fluency components is exploited by the classifier to increase the overall performance. Finally, we should note that also for AD, the variables in the best model nicely mirror the PCA components found on healthy controls, further supporting the components' identification.

Overall our findings highlight that a more fine-grained evaluation of the pattern of performance of patients would grant a significant improvement of the diagnostic accuracy of semantic fluency. In agreement with others (e.g., Thompson, Stopford, Snowden, & Neary, 2005), our study advises the adoption in clinical practice of more sophisticated and multidimensional scoring procedures to maximise the information that can be extracted from clinical tests.

Acknowledgements

We thank Erminio Capitani and Marcella Laiacona for allowing us to use their data on healthy subjects and Anna Blumenthal for proofreading the first version of the manuscript. We would also like to thank Tim Shallice for his comments on an earlier draft of this manuscript and Marco Perugini and Alessio Toraldo for the interesting discussions on statistical methods.

Contributions: CR devised the experiment and the analyses; SB, SL, and CR collected data; CR performed the analyses; and CR, SL and PC wrote the paper.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cortex.2014.02.006.

REFERENCES

Abwender, D. A., Swan, J. G., Bowerman, J. T., & Connolly, S. W. (2001). Qualitative analysis of verbal fluency output: review and comparison of several scoring methods. *Assessment, 8*(3), 323–338.

Badre, D., & Wagner, A. D. (2007). Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia, 45*(13), 2883–2901.

Baldo, J. V., Schwartz, S., Wilkins, D., & Dronkers, N. F. (2006). Role of frontal versus temporal cortex in verbal fluency as revealed by voxel-based lesion symptom mapping. *Journal of the International Neuropsychological Society: JINS, 12*(6), 896–900.

Baldo, J. V., & Shimamura, A. P. (1998). Letter and category fluency in patients with frontal lobe lesions. *Neuropsychology, 12*(2), 259–267.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *BMJ, 310*(6973), 170.

Bousfield, W. A., & Sedgewick, H. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 52*, 83–95.

Burgess, P. W., & Shallice, T. (1996). Response suppression, initiation and strategy use following frontal lobe lesions. *Neuropsychologia, 34*(4), 263–272.

Capitani, E., Laiacona, M., & Barbarotto, R. (1999). Gender affects word retrieval of certain categories in semantic fluency tasks. *Cortex, 35*(2), 273–278.

Capitani, E., Rosci, C., Saetti, M. C., & Laiacona, M. (2009). Mirror asymmetry of category and letter fluency in traumatic brain injury and Alzheimer's patients. *Neuropsychologia, 47*(2), 423–429.

Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research, 12*(3), 289–325.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol., 2*(3), 1–27.

Costafreda, S. G., Fu, C. H. Y., Lee, L., Everitt, B., Brammer, M. J., & David, A. S. (2006). A systematic review and quantitative appraisal of fMRI studies of verbal fluency: role of the left inferior frontal gyrus. *Human Brain Mapping, 27*(10), 799–810.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences, 14*(4), 172–179.

Fagundo, A. B., López, S., Romero, M., Guarch, J., Marcos, T., & Salamero, M. (2008). Clustering and switching in semantic fluency: predictors of the development of Alzheimer's disease. *International Journal of Geriatric Psychiatry, 23*(10), 1007–1013.

Gorno-Tempini, M. L., Dronkers, N. F., Rankin, K. P., Ogar, J. M., Phengrasamy, L., Rosen, H. J., et al. (2004). Cognition and anatomy in three variants of primary progressive aphasia. *Annals of Neurology, 55*(3), 335–346.

Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, a, Mendez, M., Cappa, S. F., et al. (2011). Classification of primary progressive aphasia and its variants. *Neurology, 76*(11), 1006–1014.

Grossman, M. (2010). Primary progressive aphasia: clinicopathological correlations. *Nature Reviews. Neurology, 6*(2), 88–97.

Gruenewald, P. J., & Lockhead, G. R. (1980). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory, 6*(3), 225–240.

Henry, J. D., & Crawford, J. R. (2004). A meta-analytic review of verbal fluency performance following focal cortical lesions. *Neuropsychology, 18*(2), 284–295.

Hirshorn, E. a, & Thompson-Schill, S. L. (2006). Role of the left inferior frontal gyrus in covert word retrieval: neural correlates of switching during verbal fluency. *Neuropsychologia, 44*(12), 2547–2557.

Ho, A. K., Sahakian, B. J., Robbins, T. W., Barker, R. A., Rosser, A. E., & Hodges, J. R. (2002). Verbal fluency in Huntington's disease: a longitudinal analysis of phonemic and semantic clustering and switching. *Neuropsychologia, 40*(8), 1277–1284.

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: implications for the organisation of semantic memory. *Memory (Hove, England), 3*(3–4), 463–495.

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia. Progressive fluent aphasia with temporal lobe atrophy. *Brain, 115*, 1783–1806.

Katz, S., Downs, T. D., Cash, H. R., & Grotz, R. C. (1970). Progress in development of the index of ADL. *The Gerontologist, 10*(1), 20–30.

Katzev, M., Tüscher, O., Hennig, J., Weiller, C., & Kaller, C. P. (2013). Revisiting the functional specialization of left inferior frontal gyrus in phonological and semantic fluency: the crucial role of task demands and individual ability. *The Journal of Neuroscience, 33*(18), 7837–7845.

Klöppel, S., Stonnington, C. M., Barnes, J., Chen, F., Chu, C., Good, C. D., et al. (2008). Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. *Brain: A Journal of Neurology, 131*(Pt 11), 2969–2974.

Laisney, M., Matuszewski, V., Mézenge, F., Belliard, S., de la Sayette, V., Eustache, F., et al. (2009). The underlying mechanisms of verbal fluency deficit in frontotemporal dementia and semantic dementia. *Journal of Neurology, 256*(7), 1083–1094.

Mayr, U. (2002). On the dissociation between clustering and switching in verbal fluency: comment on Troyer, Moscovitch, Winocur, Alexander and Stuss. *Neuropsychologia, 40*(5), 562–566.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Jr., Kawas, C. H., et al. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 7*(3), 263–269.

Moscovitch, M. (1992). Memory and working-with-memory: a component process model based on modules and central systems. *Journal of Cognitive Neuroscience, 4*(3), 257–267.

Neary, D., Snowden, J. S., Gustafson, L., Passant, U., Stuss, D., Black, S., et al. (1998). Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. *Neurology, 51*(6), 1546–1554.

Neary, D., Snowden, J., & Mann, D. (2005). Frontotemporal dementia. *Lancet Neurology, 4*(11), 771–780.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Review Neuroscience, 8*(12), 976–987.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage, 45*(1 Suppl), S199–S209.

Price, S. E., Kinsella, G. J., Ong, B., Storey, E., Mullaly, E., Phillips, M., et al. (2012). Semantic verbal fluency strategies in amnestic mild cognitive impairment. *Neuropsychology, 26*(4), 490–497.

Rascovsky, K., Hodges, J. R., Knopman, D., Mendez, M. F., Kramer, J. H., Neuhaus, J., et al. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain, 134*(9), 2456–2477.

Reverberi, C., Capitani, E., & Laiacona, M. (2004). Variabili semantico-lessicali relative a tutti gli elementi di una categoria semantica: indagine su soggetti normali italiani per la categoria "frutta." *Giornale Italiano Di Psicologia*, (3/2004), 497–522. http://dx.doi.org/10.1421/16262.

Reverberi, C., Görgen, K., & Haynes, J.-D. (2012a). Distributed representations of rule identity and rule order in human frontal cortex and striatum. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 32*(48), 17420–17430.

Reverberi, C., Görgen, K., & Haynes, J.-D. (2012b). Compositionality of rule representations in human prefrontal cortex. *Cerebral Cortex (New York, N.Y.: 1991), 22*(6), 1237–1246.

Reverberi, C., Laiacona, M., & Capitani, E. (2006). Qualitative features of semantic fluency performance in mesial and lateral frontal patients. *Neuropsychologia, 44*(3), 469–478.

Robinson, G., Shallice, T., Bozzali, M., & Cipolotti, L. (2012). The differing roles of the frontal cortex in fluency tests. *Brain: A Journal of Neurology, 135*(Pt 7), 2202–2214.

Roca, M., Parr, A., Thompson, R., Woolgar, A., Torralva, T., Antoun, N., et al. (2010). Executive function and fluid intelligence after frontal lobe lesions. *Brain: A Journal of Neurology, 133*(Pt 1), 234–247.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., et al. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological Review, 111*(1), 205–235.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General, 126*(3), 211–227.

Salmon, E., Garraux, G., Delbeuck, X., Collette, F., Kalbe, E., Zuendorf, G., et al. (2003). Predominant ventromedial frontopolar metabolic impairment in frontotemporal dementia. *NeuroImage, 20*(1), 435–440.

Snowden, J. S., Goulding, P. J., & Neary, D. (1989). Semantic dementia: a form of circumscribed cerebral atrophy. *Behavioural Neurology, 2*(3), 167–182.

Snowden, J. S., Neary, D., & Mann, D. M. A. (1996). *Fronto-temporal lobar degeneration: Fronto-temporal dementia, progressive aphasia, semantic dementia*. New York: Churchill Livingstone.

Snowden, J. S., Thompson, J. C., Stopford, C. L., Richardson, A. M. T., Gerhard, A., Neary, D., et al. (2011). The clinical diagnosis of early-onset dementias: diagnostic accuracy and clinicopathological relationships. *Brain: A Journal of Neurology, 134*(Pt 9), 2478–2492.

Stuss, D. T., Alexander, M. P., Hamer, L., Palumbo, C., Dempster, R., Binns, M., et al. (1998). The effects of focal anterior and posterior brain lesions on verbal fluency. *Journal of the International Neuropsychological Society, 4*(3), 265–278.

Stuss, D. T., Alexander, M. P., Palumbo, C. L., Buckle, L., Sayer, L., & Pogue, J. (1994). Organizational strategies with unilateral or bilateral frontal lobe injury in word learning tasks. *Neuropsychology, 8*(3), 355–373.

Thompson, J. C., Stopford, C. L., Snowden, J. S., & Neary, D. (2005). Qualitative neuropsychological performance characteristics in frontotemporal dementia and Alzheimer's disease. *Journal of Neurology, Neurosurgery and Psychiatry, 76*(7), 920–927.

Troster, A. I., Fields, J. A., Testa, J. A., Paul, R. H., Blanco, C. R., Hames, K. A., et al. (1998). Cortical and subcortical influences on clustering and switching in the performance of verbal fluency tasks. *Neuropsychologia, 36*(4), 295–304.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology, 11*(1), 138–146.

Troyer, A. K., Moscovitch, M., Winocur, G., Alexander, M. P., & Stuss, D. (1998). Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia, 36*(6), 499–504.

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2011). Variation in verbal fluency: a latent variable analysis of clustering, switching, and overall performance. *The Quarterly Journal of Experimental Psychology, 64*(3), 447–466.

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology, 27*(4), 635–657.

Wixted, J., & Rohrer, D. (1994). Analyzing the dynamics of free recall: an integrative review of the empirical literature. *Psychonomic Bulletin & Review, 1*(1), 89–106.