Moving from Human Ratings to Word Vectors to Classify People with Focal Dementias: Are We There Yet?

Chiara Barattieri di San Pietro[1,2], **Marco Marelli**[1], **Carlo Reverberi**[1]
1. Università degli Studi di Milano-Bicocca, Milano, Italy
2. Università degli Studi di Verona, Verona, Italy
chiara.barattieridisanpietro@unimib.it,
carlo.reverberi@unimib.it, marco.marelli@unimib.it

Abstract

Fine-grained variables based on semantic proximity of words can provide helpful diagnostic information when applied to the analysis of Verbal Fluency tasks. However, before leaving human-based ratings in favour of measures derived from distributional approaches, it is essential to assess the performance of the latter against that of the former. In this work, we analysed a Verbal Fluency task using measures of semantic proximity derived from Distributional Semantic Models of language, and we show how Machine Learning models based on them are less accurate in classifying patients with focal dementias than the same models built on human-based ratings. We discuss the possible interpretation of these results and the implications for the application of distributional semantics in clinical settings.

1 Introduction

A Verbal Fluency (VF) task (Lezak et al., 2004) is a test routinely used in the neuropsychological practice that requires participants to produce as many words as possible belonging to a given semantic category (e.g., "colours, "animals", etc.) within a time limit (typically 60 sec). It is commonly used to study lexical retrieval, and the subject's performance is standardly rated by the number of correct words produced for a given cue. However, to overcome the opacity of the overall score and help distinguish the different cognitive functions underpinning VF performance, additional measures of VF performance have been proposed. Among these, the number of consecutive words produced that share similar properties such as being a citrus fruit (this is called "semantic cluster" and its size is a clinically useful variable), and the total number of transitions between clusters (called "number of switches" – Troyer et al., 1997). Indeed, by characterising a semantic VF task (category "fruits") using the number of semantic categories produced, the average semantic proximity between words, the number of new words and out-of-category words, it has been possible to classify people with and without focal dementias, as well as across three different subtypes of dementias (Fronto-Temporal Dementia *versus* Primary Progressive Aphasia *versus* Semantic Dementia) with good accuracy (78% accuracy for patients vs healthy control classification, and 58.3% accuracy for classification across three pathological subcategories – Reverberi et al., 2014). One shortcoming of this model, however, is that those VP indexes are built upon human-based ratings of semantic proximity between pairs of words collected from a sample of healthy controls, making it hard to extend the same approach to words for which human judgments were not previously collected, i.e., other semantic categories.

Recent advances in Natural Language Processing techniques could help overcome this limitation. Distributional Semantic Models (DSMs) of language start from lexical co-occurrences extracted from large text corpora (Turney & Pantel, 2010), and applying different computational techniques, end up representing word meanings as numerical vectors in a multidimensional space. Here, terms that are semantically related are located close to each other. Such models can be used to simulate the structure of conceptual knowledge implied in the performance of semantic tasks such as a VF task. Indeed, DSMs have been successfully applied to different tasks of semantic relationships (Mandera et al., 2017), including the analysis of VF tasks to classify patients with Alzheimer's disease (Linz et al., 2017) and reaching remarkable accuracy (F1 = 0.77). However, despite the success, questions have been posed concerning what exactly distributional models can learn (Erk, 2016) and if such models are sufficiently rich in terms of encoded features (Lucy and Gauthier, 2017) to be applied to all sorts of semantic tasks/problems.

The present study aims to test if the analysis of a VF task based on DSM-derived measures would reproduce the results of an analysis based on human-derived measures. In particular, we decided to re-analyse the original data of a semantic VF task (category "fruit") that Reverberi et al. collected on a cohort of participants with focal dementias and healthy controls (CTR). Focal dementias are neurodegenerative diseases that cause deterioration of cognitive function, including language. The original cohort included people with Fronto-Temporal Dementia (FTD), Primary Progressive Aphasia (PPA), and Semantic Dementia (SD). Each diagnostic group presents peculiar linguistic symptomatology, making these syndromes ideal candidates for a differential approach. The human-based indexes of VF (see Section 2 for details) were adapted to be computed on different DSMs (Landauer & Dumais, 1997; Mikolov et al., 2013). Specifically, we adopted two predict and one count model. All three semantic spaces were based on the itWac web-crawled corpus (Baroni et al., 2009). The two predict models (Word-Embeddings Italian Semantic Space 1 and 2 - "WEISS1" and "WEISS2") were obtained from Marelli (2017) and were chosen for both their practical accessibility (http://meshugga.ugent.be/snaut-italian) and their proven good performance in previous studies (Mancuso et al., 2020; Nadalini et al., 2018). WEISS1 is based on a CBOW model with 400 dimensions and a 9-word window; WEISS2 is based on a CBOW model with 200 dimensions and a 5-word window. Both models consider words with a minimum frequency of 100 in the original corpus. The count-model based on Latent Semantic Analysis ("LSA") was created ad-hoc for this study following Günther and colleagues' (2015) procedure. Many psycholinguistic studies applying LSA in the English language used the TASA corpus (http://lsa.colorado.edu, including 12,190,931 tokens), which is a far smaller corpus than ItWac (about 1.9 billion tokens). To ensure comparability with this previous literature, we extracted a subset of the itWac corpus to match the TASA size. We selected an untagged set of 91,058 documents randomly extracted from itWAC, comprising the same set of words (N = 180,080) of the WEISS semantic spaces. The creation of a matrix of co-occurrences was carried out using the DISSECT toolkit (Dinu et al., 2013), and applying a Positive Pointwise Mutual Information weighting scheme (Niwa & Nitta, 1995), followed by dimensionality reduction by Singular Value Decomposition. We set the number of dimensions at 300 following the study of Landauer and Dumais (1997), which indicates good performance for dimensionalities ranging from 300 to 1,000.

2 Materials and Methods

The verbal production to a sematic VF (category "fruits") from the original cohort of 371 subjects (Table 1) was analysed. Overall datapoints were N = 3,642 words, with 133 unique words.

	PPA	**FTD**	**SD**	**CTR**
Number	16	33	15	307
Age	73.6±3.4	67.0±6.1	67.9±6.5	54.9±17
Education	7±4.6	8.6±4.4	9.3±4.9	9.6±5

Table 1: Demographic information for all the subject groups.

Data were entered in an R pipeline, leveraging on two word2vec (Mikolov et al., 2013) semantic spaces ("WEISS1" and "WEISS2"), and an LSA space with identical vocabulary size ("LSA"). For each participant, the pipeline outputs three sets of semantic indexes computed according to five different thresholds (set to identify the occurrence of a semantic switch), corresponding to the 10th, 30th, 50th, 70th, and 90th quantiles of the distribution of semantic relatedness values (Table 2), computed considering the cosine proximity of all adjacent words produced by the whole study cohort.

	10th	**30th**	**50th**	**70th**	**90th**
WEISS1	.185	.226	.247	.268	.287
WEISS2	.303	.371	.405	.434	.463
LSA	.336	.431	.479	.519	.582

Table 2: Cosine values adopted as thresholds for the three semantic spaces.

For each participant, we computed the following 9 indexes of VF:

1) *Total number of valid words*, produced in 1 minute, excluding repetitions. Differently from the original work, words not

included in the vocabulary of the semantic space were obligatory excluded, but words not belonging to the category "fruit" were kept. Due to limitations of the semantic space's vocabulary, 53 words and compound expressions (8 from the patient group and 45 from the control group) out of the 3,642 (1.5%) were removed from the data;

2) *Repetitions* ("rep"): the total number of repeated words*;*

3) *Total number of switches* ("switch"): computational equivalent of the "number of switches between subcategories" in the original work. Semantic switches were identified based on measures of semantic relatedness obtained from three semantic spaces and according to five different thresholds (Table 2);

4) *Total number of semantic clusters* ("NC"): computational equivalent of the "number of subcategories" in the original work. Clusters were identified based on the occurrence of a semantic switch, i.e., when the mean value of cosine similarity of words within a cluster drops below the identified threshold (Table 2);

5) *Mean size of clusters* ("SC"): mean number of words within a semantic cluster; computational equivalent of the "relative switching" index in the original work;

6) *Average semantic proximity* (*"*prox"), the semantic distance between adjacent words. Unlike the original index, based on human-derived estimated of semantic proximity (Reverberi et al., 2006), we derived this index from the mean cosine between the vectorial representation of adjacent words in the participants' production.

In addition, to ascertain the replicability of original results with computational methodologies, the following indexes were adapted from the original work:

7) *Mean familiarity* ("fam"). As a computational equivalent of the original index, calculated according to familiarity scores collected from a sample of healthy controls (Reverberi et al., 2004), we computed the raw word frequency as derived from the corpus of reference (itWac), converted to lower case and excluding metadata;

8) *Out-of-category words* ("OOC"): number of words not pertaining to the 15 subcategories of "fruit" as identified in previous works by the same Authors (Reverberi et al., 2004; 2006). Given that the vectorial representation of words differs according to inflectional morphology, data were not normalised (singular to plural) but kept as originally produced;

9) *Order Index* ("OI"): computed following the formula proposed in Reverberi et al., 2006. In its simplified notation, the Order Index is equivalent to the difference between the theoretical maximum number of switches (total number of words minus 1) and the actual observed switches, divided by the range of theoretically possible switches (total number of words minus 1, minus total number of clusters minus 1). To avoid non-linearity problems, the participant production is represented in a three-dimensional space having number of words, number of switches, and number of subcategories as axes: the order index is then transformed using the arctangents of the resulting segments.

2.1 Statistical Analyses

All variables of interest were pre-processed to remove variance due to differences in age, level of education, and the total number of words. We ran a linear regression analysis with the relevant variable as the dependent factor and with age, education, and the total number of words as regressors (only considering healthy subjects to avoid any potential bias in the estimates due to brain damage). We then used the regression coefficients to compute the residuals for each variable and all subjects. Residuals were then used as predicting variables for the classification analysis. The average for each variable and each patient group was compared with the respective average in the control group through a two-sample t-test, Bonferroni corrected.

2.2 Classification Analysis

The R packages `caret` and `e1071` (interfaces to the LIBSVM by Chang & Lin 2011) were used. The aim of the classification analysis was to determine: i) which variables, alone or in combination, would be able to classify a subject as being either a patient or control, and; ii) which variables, alone or in combination, would best classify a patient as being member of one of the three frontal dementia group (FTD, PPA, SD).

After removing variance due to differences in age and education, we performed a Leave-One-Out Cross-Validation (LOOCV) analysis. The model kernels were set as linear, and relative weights were added to counterbalance the difference in group numerosity. In LOOCV, a data instance is left out, and a model is constructed on all other data instances in the training set. The model is tested against the data point left out, and the associated error is recorded. The process is then repeated for all data points, and the overall prediction error is calculated by taking the average of the recorded test error estimates. The LOOCV analysis was repeated for each combination of the 9 variables of interest, for each of the 3 semantic spaces, and each of the 5 thresholds, resulting in 7,665 models.

3 Results

We compared the performance of each group to that of healthy controls for each of the nine variables considered. All pathological groups significantly differed from the controls on at least one variable (Table 3). In the classification analysis, we investigated which variables (alone, or in all the possible combinations with other variables, i.e., 511 combinations) would best predict the membership of participants. We carried out two sets of analysis: i) healthy controls versus participants with focal dementias (PPA, FTD, and SD); and ii) participants with PPA versus participants with FTD versus participants with SD. The analysis was performed for each semantic space and for each preidentified threshold for a total number of 7,665 models.

	FTD	PPA	SD
Proximity	+		
Familiarity			
New words	+		+
Out-Of-Category			
N Switches	+		
N Cluster	+		
Size Cluster	+	+	+
Order Index	+	+	
Repetitions			

Table 3: Variables that are significantly different between a given pathological group vis-à-vis healthy controls. Results Bonferroni-corrected for multiple comparison are reported.

The best classification performances for patients versus healthy controls was found when we considered the variables "total number of new words" and "Order Index" at any threshold and with all semantic spaces. In these cases, the overall accuracy of the models was 61.2%, with sensitivity of 57.4% and specificity of 79.7% (Table 4).

SS	Thres.	Vars	Acc.	Sens.	Spec.
Human-Based		**NC + prox + new + OOC**	**84**	**86**	**82**
all	all	New + OI	61.2	57.4	79.7
-	-	New	61.0	57.0	79.7
all	all	OI	61.0	57.0	79.7
all	all	Rep + new + OI	60.7	55.7	84.4
-	-	OOC	60.4	56.4	79.7

Table 4. Top 5 performing classification models (patients vs controls).

The best classification performances for patients in their specific pathology group was found when we considered the variables "out of category words", "average semantic proximity", and "size of clusters" computed at the 3rd threshold (50th) of the WEISS2 space (Table 5). In this case, the overall max accuracy was 43.8%. Sensitivity and specificity for each pathology group were: PPA = 87.5% and 62.5%; FTD = 36.4% and 71%; SD = 13.33% and 81.6%, respectively.

SS	Thres.	Vars	Acc.	PPA	FTD	SD
Human-Based		**Fam + NS + OI + new + rep**	**58**	**NA**	**NA**	**NA**
W2	50	OOC + prox + SC	43.8	87.5/62.5	36.4/71	13.3/81.6
W1	10	OOC + SC	42.2	87.5/56.3	39.4/74.2	0/83.7
W1	30	NS + NC	40.6	93.8/50	33.3/77.4	0/85.7
W1	70	OOC + SC	40.6	87.5/62.5	36.4/64.5	0/81.6
W2	90	SC	39.1	68.8/60.4	42.4/64.5	0/81.6

Table 5. Top 5 performing classification models (patients in each specific pathology group).

4 Discussion

In this work, we replaced human-based measures of semantic proximity with DSM-derived measures of semantic proximity to compute a set of indexes of VF that was found to be able to classify with good accuracy people with and without focal dementias based on their verbal production to a semantic VF task (category "fruits", which was originally adopted to limit the set of possible items as compared to broader categories such as “animals”). The objective of the study was to assess the accuracy of Machine Learning (ML) models based on DSM measures of semantic information, in view of their possible extension to words and semantic categories for whom the measure of semantic proximity is not available. Despite being above chance in both cases, ML models based on DSM-derived measures of semantic proximity showed lower accuracy compared to models built on human-based ratings. This was true both for the classification of patients versus controls (61.2% and 84%, respectively), as well as for the subclassification of diagnosis (43.8% and 58%, respectively).

The observed differences might be due to the functional adaptations needed to transpose the original VF indexes to DSM-derived measures. For example, the computational equivalent of the "familiarity" index, calculated according to familiarity scores collected from the sample of healthy controls, was approximated via the raw word frequency as derived from the corpus of reference. Moreover, given that the vectorial representation of words differs according to inflectional morphology, data were not normalised (singular to plural) but kept as originally produced, unlike the original work. Hence, it might be possible that these operations introduced some distortions that could explain the differences observed compared to the original study.

In terms of parameter setting, it is worth noting that our choices might have affect the overall performance of the adopted models, possibly reducing their ability to avoid noise and biases. For example, according to Tripodi (2017), hyperparameter setting for Italian has specific requirements in terms of vector size, negative sampling, vocabulary threshold cutting, to maximize performance in an analogy task (although to what extent such recommendation can be extended to VF is an empirical question that remains to be addressed). Also, the choice of a CBOW model, instead of “more predictive” algorithms such as Skipgram and Mask might have reduced the ability of the model to mimic the human ratings of word associations.

However, a different explanation might be related to the type of information encoded into the human proximity ratings. Given its evolutionary relevance, the neural substrate underpinning the notion of "fruits" might encode a rich multidimensional semantic characterisation (including sensory information such as taste, smell, sight, touch). As such, the representation of this semantic category might not be simply derivable by the lexical distribution of its items in a corpus. Differently, other semantic categories might leverage on less perceptual and more encyclopaedic semantic knowledge, such as, for example, the category "animals", another semantic cue widely used for the assessment of VF. Indeed, while people do generally have first-hand, real-life experience of "fruits", knowledge about "animals" may be more commonly derived from indirect exposure to encyclopaedic information (i.e., the media). In other words, when we think about a cherry, we may not only recall the meaning of the lemma as compared to, for example, an apple, but at the same time, we might also recall the sensory information attached to the drupe (round, red, juicy, etc.). Conversely, apart from common pets, it is unlikely that participants have first-hand experience about most of the items commonly included "animals" category (e.g., "lion", “whale”, etc.).

This means that distributional models might be not the best-suited tool to resolve semantic problems when the semantic task under investigation makes use of a subset of words pertaining to a semantic category perceptually rich (such as that of “fruits”).

5 Conclusions and Future Works

The past decades have witnessed an increasing interest towards the application of NLP techniques to answer, or support the resolution of, different clinical problems, from patients’ classifications to disease monitoring, and from differential diagnosis to prediction of treatment response (see de Boer et al., 2018 for a comprehensive review). All these applications implicitly rely on the assumption that these techniques are agnostic/transparent to the semantic task under investigation and, given the good results obtained, that they are equipped with sufficiently rich semantic information to solve any kind of task based on linguistic data. Our findings challenge this idea and align with previous works pointing to a lack of basic features of perceptual meaning in DSM (Lucy and Gauthier, 2017).

Implications for the application of DSM-derived measures to clinical work and research indicate that the choice of the verbal task and the associated DSM can affect the results. For this reason, we plan to assess the classification accuracy of ML models built both on human ratings and DSM-derived measures of semantic proximity for other categorical VF tasks, as well as adopting word vectors derived from lemmatised corpora.

Before moving to more recent language models such as the last generation of deep neural language models like BERT (Devlin et al., 2019), consideration should be given to the trade-off between computational and data resources needed to train them (Bender et al., 2021) on one hand, and what kind of added value they can give compared to traditional "static" embeddings (Lenci et al., 2021) on the other. Further research might address the limits of current DSM models by enriching the information encoded, integrating experiential and distributional data to induce reliable semantic representations (Andrews et al., 2009). Additional sources of multimodal information (e.g., Lynnott et al., 2020) including visual and audio information, might help overcome these current limitations (Chen et al., 2021).

References

Baroni Marco, Bernardini Silvia, Ferraresi Adriano and Zanchetta Eros. 2009. The waCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3): 209–226.

Bender Emily M., Gebru Timnit, McMillan-Major Angelina & Shmitchell Shmargaret. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 610-623.

Chang Chih-Chung and Lin Chih-Jen. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology* (TIST), 2(3): 1-27.

Chen Wei, Wang Weiping, Liu Li and Lew Micheal S. 2021. New ideas and trends in deep multimodal content understanding: A review. *Neurocomputing, 426*:195-215.

De Boer Jann N., Voppel Alban E., Begemann Marieke J.H., Schnack Hugo G., Wijnen Frank and Sommer Iris E.C. 2018. Clinical use of semantic space models in psychiatry and neurology: a systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews, 93:* 85-92.

Devlin Jacob, Chang MW, Lee K, Toutanova K. 2019 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACLHLT 2019*, 4171–4186

Dinu Georgiana and Baroni Marco. 2013. Dissect-distributional semantics composition toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 31-36.

Günther Fritz, Dudschig Caroline and Kaup Barbara. 2015. Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology, 69*(4):626–653.

Landauer Thomas and Dumais Susan. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.

Lenci Alessandro, Sahlgren Magnus, Jeuniaux Patrick, Gyllensten Amaru Cuba and Miliani Martina 2021. A comprehensive comparative evaluation and analysis of Distributional Semantic Models. arXiv preprint arXiv:2105.09825.

Lezak Muriel, Howieson Diane, Loring David, Hannay Julia and Fischer Jill. 2004. *Neuropsychological assessment*. New York: OUP, USA.

Lucy Li and Gauthier Jon. 2017. Are Distributional Representations Ready for the Real World? Evaluating Word Vectors for Grounded Perceptual Meaning. *Proceedings of the First Workshop on Language Grounding for Robotics.*

Lynott Dermot, Connell Louise, Brysbaert Marc, Brand James and Carney James. 2020. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3), 1271-1291.

Mandera Paul, Keuleers Emmanuel and Brysbaert Marc. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57-78.

Marelli Marco. 2017. Word-embeddings Italian Semantic spaces: A semantic model for psycholinguistic research. *Psihologija*, 50(4): 503–520.

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg and Dean Jeffrey. 2013. Distributed Representations of Words and Phrases and their Compositionality. Retrieved from http://arxiv.org/abs/1310.4546

Niwa Yoshiki and Nitta Yoshihiko. 1995. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. arXiv preprint cmp-lg/9503025

R CoreTeam. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from https://www.r-project.org.

Reverberi Carlo, Cherubini Paolo, Baldinelli Sara and Luzzi Simona. 2014. Semantic fluency: Cognitive basis and diagnostic performance in focal dementias and Alzheimer's disease. *Cortex*, 54, 150-164.

Tripodi Rocco and Pira Stefano Li. 2017. Analysis of Italian word embeddings. arXiv preprint arXiv:1707.08783.

Turney Peter D. and Pantel Patrick. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141-188