Automated analysis of the semantics in narratives by persons with Primary Progressive Aphasia.

Anonymous ACL submission

Abstract

We present a method to detect differences in the semantics of the spontaneous language of persons with separate primary progressive aphasia syndromes (PPA) using automated Information Control Unit derivation. The resulting semantic clusters are evaluated for their use in a predictive model to identify speakers with PPA. A prototype description is automatically generated based on a picture description by control speakers. Clustering is used to identify topics. The semantic distance between the prototype and language from persons with PPA is used to quantify the degree to which the language of persons with PPA deviates from normal language. A classifier is used to classify individual fragments.

The vocabulary of speakers with PPA is found to be less diverse in speakers with PPA. Different clusters are identified automatically that correspond with categories of objects and actions. In several clusters, speakers with PPA show deviations from the prototype. Random Forest classification out-performs baseline in control vs PPA and control vs svPPA vs nfvPPA tasks. Whereas nfvPPA is usually associated with speech motor problems, our study also finds their language deviating on the level of semantics.

1 Introduction

One of the clinical manifestations of dementia is a decline of the ability to use language. Problems with language have been reported in individuals with dementia caused by Alzheimer's disease, Parkinson's disease or frontotemporal lobar degeneration. The term Primary Progressive Aphasia (PPA; Mesulam 2001) is used to describe a neurodegenerative condition in which the primary, dominant symptom is a progressive language disorder.

Individuals with PPA form a subclass of individuals with either Frontotemporal dementia (FTD) or Alzheimer pathology (Rohrer et al., 2012). There is commonly a threeway distinction of PPA types, each with different linguistic characteristics: a *semantic* variant (svPPA; characterized by fluent but increasingly empty speech with affected naming and word comprehension), a *nonfluent* variant (nfvPPA; characterized by agrammatism and/or hesitant or labored speech / apraxia of speech) and a *logopenic* variant (lpvPPA; characterized by aphasia with anomia and difficulties with repetition of sentences or phrases).

There is a large variation of language deficits and atrophy patterns, both within each of the PPA subgroups and between them (Louwersheimer et al., 2016; Patterson et al., 2006; Thompson et al., 1997, 2012; Wilson et al., 2010, 2018). Some patients present with language problems even if they don't yet meet the published guidelines for PPA; and some present with heterogeneous language problems and mixed phenotypical manifestations that do not clearly follow the threeway distinction.

One of the standard tasks in the clinical assessment of a person's language is an analysis of their spontaneous speech and language, through stimuli that elicit connected speech (Boschi et al., 2017). The usual stimulus is an image that provides a visual context for a narrative. In most cases (e.g., Goodglass, 2000; Swinburn et al., 2004), the image is associated with Information Control Units (ICUs; Yancheva and Rudzicz 2016), usually human-supplied (hsICUs), which represents the objects, actions and causality relations of the figures in the image. Previous studies have found that the scoring of ICUs and their comparison to predefined hsICUs can indicate differences between the narratives from healthy persons and those with aphasia (Hier et al., 1985; Croisile et al., 1996).

As one of the defining characteristics of svPPA is anomia, the typical scoring of ICUs for this group deviates, due to the difficulty with mapping an image's figures onto nouns and verbs (Bozeat et al., 2002; Garrard and Carroll, 2006; Hoffman et al., 2013). Persons with nfvPPA have poorer fluency and reduced syntax, however their ability to name things is relatively spared (Mancano and Papagno, 2023).

Analyses are usually based on multiple variables measured in a transcription, at different levels of detail. Some variables require more language data for reliable analysis than others, which impacts the required amount of effort (Ossewaarde et al., 2020). Transcribing what is said into individual tokens requires sufficient knowledge of the spoken language to identify the words used by the speaker. The annotation of word categories and their meaning requires knowledge of linguistic concepts (part-of-speech) and also consensus about the meaning expressed by the words in the language. ICU analysis, the measurement of the distance between the language in the transcription and the ICUs, requires interpretation of what is said.

Manual annotation is labor intensive, expensive, and error prone. Automatic annotation with software has been shown to be useful for speech assessments in the context of other forms of dementia (e.g. Robin et al., 2023). However, for PPA, it is still an open question how specifically the changes in the semantics of the language can be recognized with software such that human interpretation of the language is not necessary.

Therefore, this study investigates the degree to which the use of software can automate ICU analysis in such a way that machine learning models can detect whether a given speaker is from the PPA group or from the control group. To this end, we set out to automatically analyze fragments of semispontaneous, connected, spoken Dutch language. Any positive result on the classification task would provide suggestions for the way in which meaning expression can be quantified in a diagnostic setting.

2 Methods

2.1 Participants

Language samples were collected from two different groups of Dutch speaking participants: one group that served as a control group ($n = 15$) and one group of participants with dementia related brain damage ($n = 16$), split evenly between nfvPPA and svPPA diagnosed.[1]

Participants in the PPA groups were under the care of neurologists at the Alzheimer Center of the Amsterdam University Medical Center and part of the Amsterdam Dementia Cohort (Van der Flier et al., 2014). They were asked to participate after their clinical consultation with a neurologist. Inclusion criteria were: able to understand and follow the task instructions, and able to generate speech (ie: not mutistic). The assessment of probable PPA was according to the diagnostic criteria of Gorno-Tempini et al. (2011). Their clinical workup followed a standardized healthcare pathway that includes a battery of diagnostic tests. In 12 cases amyloid biomarker assessment had taken place.

Participants in the control group were enrolled in a larger cohort of volunteer subjects in brain research studies (Dutch Brain Research Registry; Zwan et al., 2021). They were matched demographically by the selection algorithm of the registry. Control participants were included when they were native speakers of Dutch and had no history (self-reported) of neurological or psychiatric disorders. Demographic characteristics are reported in Table 1.

2.2 Elicitation

Spontaneous speech data was collected via the spontaneous speech task from the Comprehensive Aphasia Test (CAT-NL, Swinburn et al. 2004). The stimulus material consisted of an image portraying distinct elements, including a fish tank and an array of books, all of which are interlinked through a series of causal and consequential relationships. The participants were directed to describe the picture with the verbal prompts stipulated within the official guidelines of the CAT-NL. The assessment sessions with participants in the svPPA and nfvPPA groups

[1] Written informed consent was obtained from all participants. Ethical approval was determined exempt by the Medical Ethics Committee of the Amsterdam University Medical Center.

	control	nfvPPA	svPPA
Number of participants	15	8	8
Number of language samples	15	13	16
Persons:			
Women (%)	60.0	62.5	62.5
Samples:			
Women (%)	60	54	69
Months since symptom onset	n/a	35.9	34.6
Age at language recording	63.4 $\pm$8.4	66.7 $\pm$6.1	66.1 $\pm$3.0
MMSE	n/a	26.3 $\pm$1.4	27.5 $\pm$0.6

Data are shown as mean +/- standard deviation or frequency (%). Sample variables are computed at time of recording. Kruskal-Wallis test indicated no statistically significant different distributions with $alpha <= 0.05$. nfvPPA: nonfluent variant of PPA, svPPA: semantic variant of PPA, MMSE: Mini-mental State Examination score.

Table 1: Main clinical and demographic characteristics.

were conducted face-to-face within the clinical setting. Where possible, participants in these groups also contributed at follow-up visits. Sessions with control participants were held once per participant. Language in these groups was elicited via video conferencing (Google Meet) due to social distancing measures at the time of elicitation.

Some participants from the patient groups contributed samples at followup visits. Samples were assumed to be independent data points, given that the time between tests was sufficiently large to exclude memory effects ($>$ 90 days), and given the relatively heterogeneous character of the disease, which negatively affects the correlation that can be expected because samples are produced by the same individual.

2.3 Transcription and linguistic analysis

At transcription, the starts and ends of the recordings of participants were manually trimmed so that only the audio of the CAT-NL spontaneous speech task resulted. The start condition was the moment that the interviewer finished the instructions to the participant. The end condition was the signal from the participant that the storytelling was over.

The tokens in the spoken fragments were transcribed in a broad transcription, with special markings for filled pauses (*/eh/* and */ehm/*). There was no separate tier to transcribe temporal properties or (morpho-)syntax.

Part of speech tags were assigned automatically by RNNTagger (Schmid, 2019) trained on the Eindhoven Corpus[2].

2.3.1 Comparison of transcripts to prototype

A statistical way of capturing the meaning of a word is through meaurement of its similarity to other words in the same embedding context. (Distributional Hypothesis; Sahlgren, 2008). In practice, word embeddings are represented by a vector with enough dimensionality to be informative enough. Vectors are computed through large scale corpus analysis, resulting in either context independent word vectors (one vector for *left* in "the **left** side was **left** unpainted"; word2vec models; Mikolov et al., 2017) or context dependent word vectors (two vectors for *left* in the same example; BERT models; Devlin et al., 2019). Word embeddings have been shown to adequately capture the similarity of semantically similar words, thus acting as a proxy for the truth-conditional meaning of that sense of the word. The use of vectors allows a natural way to study the relatedness of the words that persons use in the retelling of a narrative.

We use a monolingual Dutch transformer-based pre-trained language model (BERTje; de Vries et al., 2019) to map tokens to embeddings. The context dependent nature of BERTje means that semantic similarity comparisons of absolute values are less robust on the word level because contextual information influences the relatedness values.

[2]Eindhoven-corpus (Version 2.0.1) (2014) [Data set]. Available at the Dutch Language Institute: `http://hdl.handle.net/10032/tm-a2-n6`

BERT-like models boosts their performance for out-of-vocabulary words by computing embeddings on the sub-word level ('embedding' is represented as ['em' 'bed' 'ding']. Out-of-vocabulary words include neologisms and phonological paraphasias. In this study, we matched what is tokenized by BERTje with the tokens processed by the Stanford parser to ensure the correct alignment of sub-word tokens with words in in the transcript.

The embeddings are used to create a prototype of the picture description, based on the language of healthy speakers. The prototype is then used to investigate to what extent the divergence from the prototype is indicative of aphasia caused by PPA.

All content words (nouns, verbs) of the control speakers were clustered jointly based on their similarity scores. Similarity scores were derived from the hidden layers of the pretrained BERTje vectors through summation of the last four layers. The Elbow method (Satopaa et al., 2011) was used to determine the optimum number of clusters, optimizing for the lowest WCSS (distortion) score. The average optimal WCSS scores were found to be at $k = 12$ clusters.

Because high dimension data sets can encode its information in such a sparse way that subsequent clustering suffers in terms of performance, it is standard practice to apply a dimensionality reduction step as part of preprocessing before clustering. We applied the Uniform Manifold Approximation & Projection algorithm (UMAP; McInnes et al., 2018).

The data pipeline is illustrated in Figure 1.

The clustered space may be considered as a prototype: it is the summation of all content words used by control speakers clustered around centers that represent a dimension in the conceptual space. Each of the speaker's descriptions of the picture is a variant on the prototype. The distance of each of the content terms of each speaker to the cluster centers was computed to yield a fingerprint for the relation between prototype and variant as follows:

Each token that a participant uses is labeled as belonging to a cluster of the prototype. The *mean average distance* to each cluster is a measure for the semantic closeness to that cluster. If a cluster is about a specific concept that is part of the picture, such as the fish tank or the books in the picture of the CAT, then any speaker should be expected to also use words relating to those categories. If the speaker uses different but semantically similar terms, then the distance will be higher, but still closer than if a speaker uses semantically vague terms.

The labels of the words spoken by the participant form a bag (multiset). Comparison between the bag of unique labels spoken by the participant to the bag of labels in the prototype is quantified using the *Tversky index*, which is a widely used asymmetric similarity measure for comparison of a variant to a prototype.

$$S(A, B) = \frac{|A \cap B|}{|A \cap B| + \alpha|B - A| + \beta|A - B|} \tag{Tversky index}$$

Because the clusters of the variant are a priori derived from the prototype, the α parameter - a multiplier for the number of clusters in the variant that do not occur in the prototype - is necessarily zero. The β parameter was set to 1. The index that we use is insensitive to the number of times that a person mentions the same topic.

The number of words assigned to each cluster is an indication of the semantic fingerprint of the speaker's narrative. Between group comparisons are performed using one-way ANOVA tests.

2.4 Classification

A Random Forest Classifier (Breiman, 2001) was used to classify the participants. The independent variables were: the Tversky index, the frequency of each cluster label, and the average distance of the tokens to the cluster centers. The dependent variable was either the binary distinction *control* versus *patient* or the ternary distinction *control* versus *nfvPPA* versus *svPPA*. The cross-validation performance was used to tune the model. The number of trees was chosen as 100, with no constraints on the maximum depth of the tree. To evaluate the model, the out-of-sample performance was estimated using leave-one-out cross validation. Scoring of the classifier is reported using the balanced accuracy metric.

3 Results

The clusters and English translations of their tokens are reported in Appendix A. Their relative contributions are visualized in Figure 2.

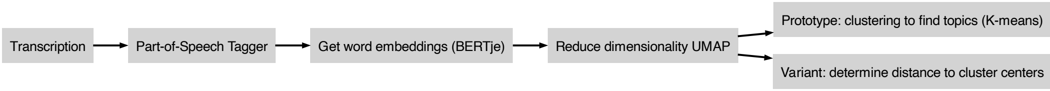

Figure 1: The pipeline from transcription to prediction

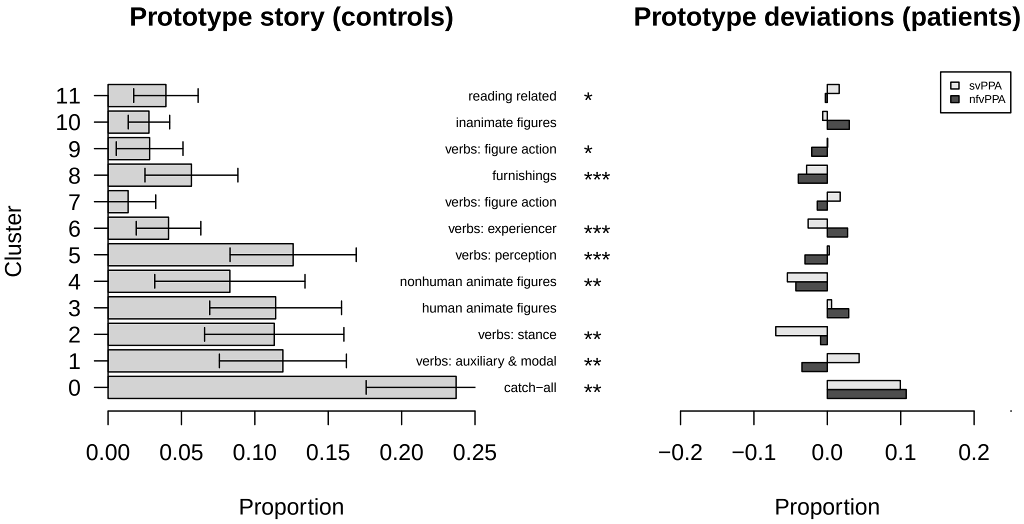

Figure 2: Relative contributions of clusters to the CAT-NL story.

The silhouette coefficient plot in Figure 3 visualizes the relationship between tokens and their assigned clusters.

The k-means algorithm, due to its deterministic character and its assumption of centers in spherical clusters, will always classify data, even words that are noise. Cluster 0 is the largest cluster in terms of number of different words, and the most heterogeneous. Both human inspection and the negative coefficient in the silhouette plot for this cluster suggest that this is the default cluster which is assigned to words that do not clearly belong to any of the other clusters.

Cluster 1 contains verbs that usually function as auxiliaries or modals. Clusters 2, 7 and 9 mostly contain the verbs that describe the action in the image, with one cluster (cluster 2) particularly associated with verbs of stance. Two clusters contain human and nonhuman animate figures respectively (clusters 3 and 4); one cluster (10) contains most of the inanimate figures that occur in the image.

Narratives from the nfvPPA group show significantly less words in almost all the clusters which is indicative of the generally shorter and more effortful speech in that group.

The comparison between participant groups for both the Tversky index variable and the absolute set size of each cluster is reported in Table 2. Except for clusters 3, 7 and 10, there are significant differences between each of the groups.

Cluster 0 is the largest cluster in terms of words, and the most heterogeneous. Both human inspection and the negative coefficient in the silhouette plot for this cluster suggest that this is the default cluster which is assigned to words that don't clearly belong to any of the other clusters. Cluster 1 contains verbs that usually function as auxiliaries or modals.

Both svPPA and nfvPPA group speakers produce relatively more words in cluster 0, the cluster with the most semantically distant (unrelated or vaguer) words. Both groups produce relatively fewer words in clusters 4 and 8, which contains the nonhuman animate figures and furnishings respectively.

Concerning verbs, nfvPPA speakers use fewer auxiliary and modal verbs (cluster 1) and

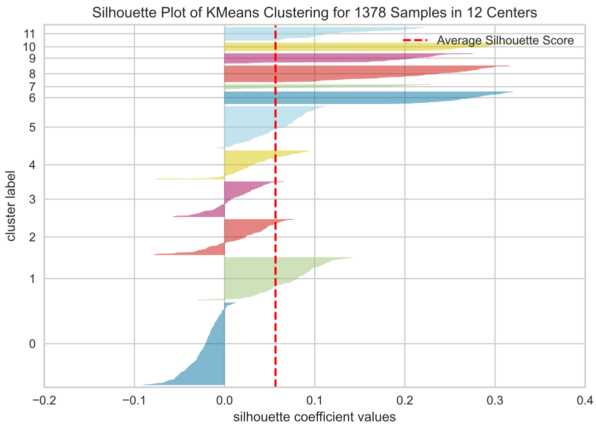

Figure 3: Silhouette plot of the clusters with their average silhouette score.

verbs that relate to what happens in the figure (clusters 5, 9). They use more verbs that relate to how they interpret the image (think, suspect; cluster 6). SvPPA speakers use more auxiliary verbs (cluster 1), fewer verbs that describe the stance of the figures in the image (cluster 2) and fewer verbs that relate to how they interpret the image (cluster 6).

The emerging semantic profile of the nfvPPA group is that of a narrative that has words for the humans in the figure and their stance actions in similar proportions as that of control speakers. However, the description of non-human animate figures (fish, cat, (teddy) bear, plant) and of furnishings are sparser, as well as their use of auxiliary verbs that are typically used in grammatically more complex expressions. Their language contains more words that are semantically remote from the prototype.

The svPPA group uses more auxiliary verbs and more words that are semantically remote. The typical difficulty with naming that develops over time in this group manifests through a smaller proportion of words assigned to the clusters with more extensional meaning (fewer words in clusters 2, 4, 6, 8, 11). The profile of this narrative fits the description of relatively intact syntax, but difficulties in recalling the specific words.

3.1 Results of the classification of individuals

The per-class results of the classification are summarized as confusion matrices in Tables 3a and 3b for the two and three class classifiers respectively. The observed performance is reported in Table 4. In both tasks, the classification performed significantly better than the baseline strategy of predicting the most frequent label.

4 Discussion

In this paper, we set out to quantify the degree to which the semantic content of a narrative by PPA participants differs from that by control speakers. In our methodology, we did not identify any topics a priori (hsICU), but rather used software to create a prototype from the narratives of controls, and then measured how the speech of PPA diverges.

One major finding is that the Tversky measure for both the svPPA and nfvPPA stories is significantly lower than that of control stories, with the nfvPPA group scoring lowest. The lower Tversky index for patient group speakers indicates that these speakers used relatively less distinct words to describe the story than speakers from the control group. This indicates that the vocabulary that is used by these speakers shows less variation, which relates to the general finding that vocabulary creativity decreases under the influence of PPA (Fraser et al., 2014). For nfvPPA participants, semantic effects are a surprising finding, given that nfvPPA is usually associated with effortful speech, in some cases caused by speech motor problems (Primary Progressive Apraxia of Speech, PPAOS; Duffy, 2006). In our grouping of participants, we did not subdivide the participants of the nfvPPA group, therefore categorizing nfvPPA

Group	**Tversky μ (σ)**	**0**	**1**	**2**	3	**4**	**5**	**6**	7	**8**	**9**	**10**	**11**
	**	**	**	**		**	***	***		***	*		*
Control	0.96 (0.03)	21.33	11.00	9.40	9.40	7.60	10.87	3.27	1.33	4.20	2.40	2.13	3.20
nfvPPA	0.77 (0.16)	11.00	3.00	3.75	4.67	1.83	3.58	1.92	0.00	0.83	0.33	1.92	1.33
svPPA	0.85 (0.17)	26.69	14.50	4.06	10.06	3.00	10.00	1.19	1.94	2.19	2.25	1.62	1.88

Table 2: The Tversky index and number of tokens assigned to each cluster. Alpha-values for significance: '*': 0.05, '**': 0.01, '***': 0.001.

(a) 2-class clustering

Prediction	**Actual value**	
	Control	**PPA**
Control	12	3
PPA	0	26

(b) 3-class clustering

Prediction	**Actual value**		
	Control	**nfvPPA**	**svPPA**
Control	12	0	3
nfvPPA	2	9	1
svPPA	0	1	15

Table 3: Confusion matrices for 2 and 3-class clustering.

Task	**Accuracy**	**Precision**	**F1**
2-class: control vs. PPA	0.77	0.81	0.80
3-class: control vs. nfvPPA vs. svPPA	0.70	0.71	0.71

Table 4: Observed performance of the Random Forest Classifier for the two classification tasks. The accuracy reported is the balanced accuracy, the average of recall obtained on each class. The scores for precision and F1 are micro averaged.

with PPAS in the same group as those without.

The clustering based on embeddings allows further introspection of the differences between the participant groups. Combining the Tversky findings and the cluster comparisons yields a quantification of the nfvPPA narratives as containing less content words in general (lowest Tversky index), and svPPA narratives as containing relatively more general nouns and verbs.

Our methodology shows how the contents of a story can be analyzed in an automatic way. We identified ICU's that should occur in a picture description through an automated analysis of descriptions by healthy speakers. This is an alternative to the approach in which humans predefine the elements, as in the hsICU approach that is often used in the field. One advantage of the use of software is that it scales well, even if narratives become longer or cover topics that are less predefined as those in a picture task.

In our approach, we used verbs and nouns, as these add most of the truth-conditional semantical content. The implication of our results is that persons with nfvPPA have less problems finding content words but produce less language overall, and that persons with svPPA will use content words that are emptier in meaning. The classification results indicate that verbs and nouns alone are informative enough to result in a classification between the groups. However, one aspect of the task is that of the causality between different elements in the picture. Although classification without the causality element is already promising, future research may target a way to also include words that describe the causality relation, such as subordinating conjunctions or prepositional phrases (*because, then*).

BERT models encode features such as familiarity, age-of-acquisition, frequency, and concreteness internally into their vector representations. These features have been shown individually to be predictive for the selective loss of concepts in persons with PPA. The black box nature of Transformer models does not allow direct introspection of the importance of such features for the classification prediction. Future research may focus on post hoc analyses of such features in the clusters that influence the classification.

BERTje embeddings are context dependent. For some tasks, context independent embeddings, generated by more traditional dictionary approaches (such as Word2Vec and GloVE; (Pennington et al., 2014)), perform as well as context dependent ones (e.g., Arora et al. 2020). The prediction is that context dependent models fare better when the language has more complex structure, more ambiguity in its word usage and contains more Out of Vocabulary words. This is relevant for the application to persons with language problems because their language is often marred by syntactic problems or word finding problems.

The hyperparameters in the dimensionality reduction dictate the performance of the algorithm, especially the selection of the number of clusters. Although the parameter setting was governed by best practices, a different choice for the number of clusters may result in a clustering of the labels that is more aligned with human intuitions. The silhouette visualization (cf. Figure 3) indicates a good convergence of all clusters except for cluster 0, which is the catch-all cluster for words. Participants in the svPPA group use a significantly higher number of words related to this cluster, which indicates a strategy of replacing target words with more general counterparts.

Because no topics have been identified a priori, our methodology can be seen as agnostic about the stimulus that is used to elicit the narratives. It scales to other narratives and to other languages, under the condition that pretrained embeddings are available.

5 Limitations

Words that have no truth-conditional semantics (such as pronouns) are not included in the clustering. It is expected that nfvPPA participants use more frequently constructions that are more referential (Çokal et al., 2018); our methodology yields no further insight in the usage of such words, but see [citation: name deleted to maintain review integrity] for an analysis of word usage differences between the groups of this study.

Our choice of dimensionality reduction algorithm and subsequent k-means classification is partly inspired by the white box properties of these algorithms (Leijnen et al., 2020). It is possible that other AI methods (such as artificial neural networks) show better performance, even though our training set is relatively modest.

The expression of meaning through embeddings carries the same bias as the training data used to generate the embeddings. Our expectation is that the choice of embedding model is relatively insignificant, given the nature of the analyzed texts: for the specific image in this task, descriptions are expected to contain mostly high frequency vocabulary items that name everyday things. One form of bias is significant for applications in healthcare: the training data for BERT data is derived from large scale corpora, with demographics of the speakers regressing towards the mean of the general population. Persons with PPA are generally older. Language production declines with age, even in healthy speakers (Kemper, Thompson, and Marquis, 2001). In our study, we use a task and a stimulus specifically designed for this target population; however, when extending our methodology towards stimuli with more freedom, care must be taken that the bias of the trained embeddings does not translate into bias effects in the analysis.

The pipeline includes Dutch specific elements: the parser and the embedding model. Because the quality of the subsequent analysis depends on the quality of the software elements for that language, scaling to other languages is not a given, unless similar resources are available. Some approaches to developing BERT models actively include multiple languages in the same model (e.g. multilingual BERT; Wu and Dredze 2020). The assumption is that some linguistic constructs are shared between languages, and so that the training effort of multiple languages combined is less than a per language training approach. The high interest in embeddings for different languages bodes well for the ability to scale our approach to other languages.

6 Conclusion

The use of parsing software combined with pretrained embeddings can aid in the analysis of spontaneous speech. In this study, we classified participants between control and PPA, and between the control and two of the three dominant subtypes of PPA, with a high degree of confidence. The classification is based on a comparison to the language of healthy persons, which makes the method cost effective and agnostic of predefined

References

Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual embeddings: When are they worth it?

Veronica Boschi, Eleonora Catricalà, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology*, 8:269.

Sasha Bozeat, Matthew A. Lambon Ralph, Karalyn Patterson, and John R. Hodges. 2002. When objects lose their meaning: What happens to their use? *Cognitive, Affective, & Behavioral Neuroscience*, 2(3):236–251.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Derya Çokal, Gabriel Sevilla, William Stephen Jones, Vitor Zimmerer, Felicity Deamer, Maggie Douglas, Helen Spencer, Douglas Turkington, Nicol Ferrier, Rosemary Varley, Stuart Watson, and Wolfram Hinzen. 2018. The language profile of formal thought disorder. *NPJ schizophrenia*, 4(1):18.

B. Croisile, B. Ska, M. J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet. 1996. Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language*, 53(1):1–19.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph R Duffy. 2006. Apraxia of speech in degenerative neurologic disease. *Aphasiology*, 20(6):511–527.

Kathleen Fraser, Graeme Hirst, and NL Graham. 2014. Comparison of different feature sets for identification of variants in progressive aphasia. *ACL 2014*, pages 17–26.

Peter Garrard and Erin Carroll. 2006. Lost in semantic space: A multi-modal, non-verbal assessment of feature knowledge in semantic dementia. *Brain*, 129(5):1152–1163.

Harold Goodglass. 2000. *Boston Diagnostic Aphasia Examination: Short Form Record Booklet.* Lippincott Williams & Wilkins.

Maria Luisa Gorno-Tempini, A E Hillis, S Weintraub, A Kertesz, M Mendez, S F Cappa, J M Ogar, J D Rohrer, S Black, B F Boeve, F Manes, N F Dronkers, R Vandenberghe, K Rascovsky, K Patterson, B L Miller, D S Knopman, J R Hodges, M M Mesulam, and M Grossman. 2011. Classification of primary progressive aphasia and its variants. *Neurology*, 76:1006–1014.

D. B. Hier, K. Hagenlocker, and A. G. Shindler. 1985. Language disintegration in dementia: Effects of etiology and severity. *Brain and Language*, 25(1):117–133.

Paul Hoffman, Roy W. Jones, and Matthew A. Lambon Ralph. 2013. Be concrete to be comprehended: Consistent imageability effects in semantic dementia for nouns, verbs, synonyms and associates. *Cortex*, 49(5):1206–1218.

Susan Kemper, Marilyn Thompson, and Janet Marquis. 2001. Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and aging*, 16(4):600.

Stefan Leijnen, Huib Aldewereld, Rudy van Belkom, Roland Bijvank, and Roelant Ossewaarde. 2020. An agile framework for trustworthy AI. In *NeHuAI@ ECAI*, pages 75–78.

Eva Louwersheimer, M. Antoinette Keulen, Martijn D. Steenwijk, Mike P. Wattjes, Lize C. Jiskoot, Hugo Vrenken, Charlotte E. Teunissen, Bart N. M. van Berckel, Wiesje M. van der Flier, Philip Scheltens, John C. van Swieten, and Yolande A. L. Pijnenburg. 2016. Heterogeneous Language Profiles in Patients with Primary Progressive Aphasia due to Alzheimer's Disease. *Journal of Alzheimer's disease : JAD*, 51(2):581–590.

Martina Mancano and Costanza Papagno. 2023. Concrete and Abstract Concepts in Primary Progressive Aphasia and Alzheimer's Disease: A Scoping Review. *Brain Sciences*, 13(5):765.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

M M Mesulam. 2001. Primary progressive aphasia. *Annals of neurology*, 49(4):425–32.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Roelant Ossewaarde, Roel Jonkers, Fedor Jalvingh, and Roelien Bastiaanse. 2020. Quantifying the Uncertainty of Parameters Measured in Spontaneous Speech of Speakers With Dementia. *Journal of Speech, Language, and Hearing Research*, 63(7):2255–2270.

Karalyn Patterson, Naida Graham, Matthew A Lambon Ralph, and John Hodges. 2006. Progressive non-fluent aphasia is not a progressive form of non-fluent (post-stroke) aphasia. *Aphasiology*, 20(9):1018–1034.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Jessica Robin, Mengdan Xu, Aparna Balagopalan, Jekaterina Novikova, Laura Kahn, Abdi Oday, Mohsen Hejrati, Somaye Hashemifar, Mohammadreza Negahdar, William Simpson, and Edmond Teng. 2023. Automated detection of progressive speech changes in early Alzheimer's disease. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, 15(2):e12445.

Jonathan D Rohrer, Martin N Rossor, and Jason D Warren. 2012. Alzheimer's pathology in primary progressive aphasia. *Neurobiology of aging*, 33(4):744–752.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171.

Helmut Schmid. 2019. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, pages 133–137, New York, NY, USA. Association for Computing Machinery.

Kate Swinburn, Gillian Porter, and David Howard. 2004. *CAT: Comprehensive Aphasia Test*. Psychology Press.

Cynthia Thompson, Kathleen Ballard, Mary Tait, Sandra Weintraub, and Marsel Mesulam. 1997. Patterns of language decline in non-fluent primary progressive aphasia. *Aphasiology*, 11(4-5):297–321.

Cynthia Thompson, Soojin Cho, Chien-Ju Hsu, Christina Wieneke, Alfred Rademaker, Bing Bing Weitner, M Marsel Mesulam, and Sandra Weintraub. 2012. Dissociations between fluency and agrammatism in primary progressive aphasia. *Aphasiology*, 26(1):20–43.

Wiesje M. Van der Flier, Yolande A. L. Pijnenburg, Niels Prins, Afina W. Lemstra, Femke H. Bouwman, Charlotte E. Teunissen, Bart N. M. van Berckel, Cornelis J. Stam, Frederik Barkhof, Pieter Jelle Visser, Evan van Egmond, and Philip Scheltens. 2014. Optimizing patient care and research: The Amsterdam Dementia Cohort. *Journal of Alzheimer's disease : JAD*, 41(1):313–327.

Stephen Wilson, Maya Henry, Max Besbris, Jennifer Ogar, Nina Dronkers, William Jarrold, Bruce L Miller, and Maria Luisa Gorno-Tempini. 2010. Connected speech production in three variants of primary progressive aphasia. *Brain*, 133(7):2069–2088.

Stephen M Wilson, Dana K Eriksson, Sarah M Schneck, and Jillian M Lucanie. 2018. A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PloS one*, 13(2):e0192773.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Maria Yancheva and Frank Rudzicz. 2016. Vector-space topic models for detecting Alzheimer's disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2337–2346, Berlin, Germany. Association for Computational Linguistics.

Marissa D. Zwan, Wiesje M. van der Flier, Solange Cleutjens, Tamara C. Schouten, Lisa Vermunt, Roos J. Jutten, Ingrid S. van Maurik, Sietske A. M. Sikkes, Derek Flenniken, Taylor Howell, Michael W. Weiner, Philip Scheltens, and Niels D. Prins. 2021. Dutch Brain Research Registry for study participant recruitment: Design and first results. *Alzheimer's & dementia (New York, N. Y.)*, 7(1):e12132.

A Clusters and the words they contain.

Clusters and English translations of the words they contain. Stars indicate significant deviations in absolute word counts with arrows indicating the deviation direction. The cluster labels were assigned post hoc by the authors.

Cluster		**Tokens**		**nfvPPA** vs control	**svPPA**
0	**	attention, number, picture, alcohol, busy, fold, holes, effect, speak, . . .	*catch-all category*	↑	↑
1	**	will, must, have, was, wants, succeeds, happens, would, gets, have, has, knows, is, seems, lets, been, am, causes, us, may, goes, finds, can, will, tried, give, comes, can	*auxiliary & modal verbs*	↓	↑
2	**	lay, falls, sit, stand, hang, lays, hangs, placed, stands, happens, fall, sits	*figure stance verbs*		↓
3		family, little, granddaughter, girl, dad, young, mister, small, child-, girl, on, daughter, children, daddy, child, father, man	*animate human figures*	↑	↑
4	**	cat, teddy bear, fish (pl), fish (sg), gold fish, plant, cat	*animate non-human figures*	↓	↓
5	***	see, watch, look, find	*perception verbs*	↓	
6	***	think, suspect	*experiencer verb*	↑	↓
7		awake, become, tell, receive, fishing, hit, do, say, pull, make, take, comes, getting, point, catch, fall, interfere, fell, placed, care, fetch, wake, holes, throw, hear, want, warn, can	*figure action verbs*	↓	↑
8	***	curtain, living room, table, armchair, cabinets, wall, ground, small table, chair, window, floor, upper, cabinet, shelf, coffee table, dresser, window sill, paper, walls, couch, living room, stack	*furnishings*	↓	↓
9	*	awake, plays, lays, occupied, warns, want, happens, sleep, tries, sleeping, sit, about to, points, put, holds, asks, says, plays, seems, does, comfortable, try, stay, hunts, sits, sleeps, baby sits, peaceful, light	*figure action verbs*	↓	
10		cd, living room, alcohol, audio, table, booze, plays, video, takes record, plant, empty, toy, window, speaker, vase, wine, doll, door, nice, salon, bottle, stereo, glass, pane, panes, little jar, little glass, flower, music, box, drank, cognac, house, drill, sound-, liquor, windows, stack, radio	*inanimate figures*	↑	
11	*	read, books	*about reading*		↑