# Multilingual Learning for Mild Cognitive Impairment Screening from a Clinical Speech Task

# Multilingual Learning for Mild Cognitive Impairment Screening from a Clinical Speech Task

**Hali Lindsay**[1], **Philipp Müller**[1], **Insa Kröger**[1], **Johannes Tröger**[1 2], **Nicklas Linz**[2], **Alexandra König**[3 4], **Radia Zeghari**[3], **Frans RJ Verhey**[5], **Inez HGB Ramakers**[5]
[1]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[2]ki-elements, Saarbrücken, Germany, 66111
[3]CoBTeK (Cognition Behaviour Technology) Research Lab, University Côte d'Azur, France
[4] National Institute for Research in Digital Science and Technology (INRIA), Nice, France
[5]Maastricht University, Department of Psychiatry & Neuropsychology,
School for Mental Health and Neuroscience, Alzheimer Center Limburg, Maastricht, Netherlands, 6200
hali.lindsay@dfki.de, philipp.mueller@dfki.de, insak@coli.uni-saarland.de, johannes.troeger@dfki.de,
nicklaslinz@googlemail.com, akonig03@gmail.com, Radia.zeghari@univ-cotedazur.fr,
f.verhey@maastrichtuniversity.nl, i.ramakers@maastrichtuniversity.nl

## Abstract

The Semantic Verbal Fluency Task (SVF) is an efficient and minimally invasive speech-based screening tool for Mild Cognitive Impairment (MCI). In the SVF, testees have to produce as many words for a given semantic category as possible within 60 seconds. State-of-the-art approaches for automatic evaluation of the SVF employ word embeddings to analyze semantic similarities in these word sequences. While these approaches have proven promising in a variety of test languages, the small amount of data available for any given language limits the performance. In this paper, we for the first time investigate multilingual learning approaches for MCI classification from the SVF in order to combat data scarcity. To allow for cross-language generalisation, these approaches either rely on translation to a shared language, or make use of several distinct word embeddings. In evaluations on a multilingual corpus of older French, Dutch, and German participants (Controls=66, MCI=66), we show that our multilingual approaches clearly improve over single-language baselines.

## 1 Introduction

Mild Cognitive Impairment (MCI) is a medical condition (Petersen et al., 2014) that often precedes Alzheimer's Disease (AD). The development of cost-effective and scaleable screening approaches for MCI is crucial for the early treatment and management of AD (Dubois et al., 2016). The Semantic Verbal Fluency Task (SVF) is a promising screening approach as it combines a time-efficient testing procedure with the possibility of remote and automatic evaluation (Tröger et al., 2018). In this task, the testee is asked to name as many words as possible from a given semantic category (e.g. animals) in a given time (e.g. 60-seconds). Traditionally, the number of named within-category items is used to detect cognitive impairment. However, recent research has shown that in-depth analysis of the underlying cognitive strategies used for the SVF (e.g. semantic memory retrieval, executive control) enables a more fine-grained differential diagnosis (Tröger et al., 2019).

To harness the diagnostic power of the SVF, current automatic evaluation approaches identify semantic clusters in the participants' word sequences, based on semantic word embeddings (Woods et al., 2016; Linz et al., 2017b; Paula et al., 2018). As the word embeddings used in these approaches are language-specific, training diagnostic machine learning approaches for target languages with small available datasets of SVF tests is challenging. Despite the potential of improving MCI classification by training on larger, multilingual data, all existing approaches for automatic MCI diagnosis are trained and evaluated on data from a single language.



*Proceedings of Recent Advances in Natural Language Processing*, pages 830–838
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_095

In this paper, we for the first time investigate multilingual learning approaches for MCI screening from the SVF. To train a joint model that generalises across test languages we evaluate two approaches: (1) translation to a common language, and (2) the application of several distinct embedding resources to the same SVF productions. In line with the state of the art (Paula et al., 2018), we evaluate qualitative embedding-based approaches through an extrinsic quantitative downstream NLP application (Wang et al., 2018): classification between controls (HC) versus MCI from qualitative SVF features. In evaluations on French, Dutch, and German corpora we show clear improvements of the multilingual learning approaches over the single-language baselines. Our results show that the the performance of classical single-language, single-embedding approaches heavily depends on the combination of embedding and language, hindering generalizability. In contrast, by extracting features from several embeddings simultaneously and training over several languages, we achieve improved and more consistent classification performances across several test languages.

## 2 Related Work

Our work is related to clinical Evaluation, semantic word embeddings, as well as the automatic qualitative evaluation of verbal fluency tasks.

### 2.1 Clinical Evaluation

During an SVF trial, a person is asked to name as many words from a semantic category (e.g. animals) as they can in one minute. The person's response is then scored as the number of unique words named excluding any repetitions. Typically, this word count is then used to determine if the person shows signs of cognitive impairment.

In addition to the word count, qualitative measures to evaluate underlying strategy—clustering and switching—have been proposed (Troyer et al., 1997). For this evaluation, consecutive words that have a discernible semantic relationship are considered to be in a cluster. For instance, in the SVF response "cat, dog, whale, dolphin...". "Cat" and "dog" are common pets where as "whale" and "dolphin" are marine mammals. The process of going from one cluster to the next is called switching.

Computing these additional metrics by hand is time-consuming and subjective. This has lead to developing automated methods of clustering and switching based on distributional semantics, or semantic word embeddings (Linz et al., 2017a; Clark et al., 2016).

### 2.2 Semantic Word Embeddings

Semantic word embeddings map words to a vector space encoding their semantic meaning. Words with high semantic similarity are mapped to vectors close in this semantic space, semantically dissimilar words to distant vectors. These semantic vectors are learned through a variety of algorithms on any large corpora of text with two main varieties of embeddings: contextual and non-contextual (Miaschi and Dell'Orletta, 2020). In a non-contextual word embedding, the vector representation is static, whereas, in a contextual embedding, the surrounding words are considered. For example, if we had 'cutting paper' and 'cutting class', a non-contextual word embedding would assign the same vector to 'cutting' in both phrases whereas a contextual embedding would take into account the difference of meaning.

Given the nature of the verbal fluency task, a non-contextual list of animals, this paper focuses on using different types of non-contextual word embeddings to investigate how to model a persistent underlying cognitive structure while combining data from multiple languages. To keep results comparable and reproducible, pre-trained publicly available models that are available in a range of languages are investigated namely, FastText (Bojanowski et al., 2016a), Spacy (Honnibal et al., 2020), and Wikipedia2Vec (Yamada et al., 2020a).

As semantic vectors are learned from large amounts of text corpora (usually Wikipedia and OSCAR common crawl), embedding quality heavily depends on the quantity of the available training data. While French, German and Dutch are relatively well-supported Indo-European languages, they are at a large disadvantage in comparison to English model resources. For instance, Wikipedia offers 6,317,662 articles for English but much fewer for French(2,337,481) , German(2,586,965) or Dutch(2,058,488)[1].

This presents a trade-off for approaching multilingual learning with semantic embeddings for clinical applications between maintaining the nuance of verbal fluency response in its native language or translating the response to English to take advantage of larger resources. In this paper, we

[1]https://en.wikipedia.org/wiki/List_of_WikipediasDetails_table

| | French | | | German | | | Dutch | | |
|---|---|---|---|---|---|---|---|---|---|
| | **HC** | **MCI** | ***p*** | **HC** | **MCI** | ***p*** | **HC** | **MCI** | ***p*** |
| N | 27 | 27 | - | 23 | 23 | - | 16 | 16 | - |
| Age | 69.9(3.5) | 71.4(3.0) | 0.16 | 70.1(4.5) | 71.9(4.4) | 0.26 | 66.4(4.4) | 68.1(5.8) | 0.46 |
| Education | 12.0 (3.5) | 11.6 (3.4) | 0.28 | 14.0(3.0) | 13.6(3.2) | 0.51 | 13.9(2.7) | 13.7(2.2) | 0.71 |

Table 1: Demographic information for French, German and Dutch populations. Age and education in years. Statistically significant differences between the population reported as $p$. Statistical significance is set to $p \leq 0.05$. Healthy Controls (HC). Mild Cognitive Impairment (MCI). Number of Samples (N).

investigate both scenarios of multilingual machine learning for clinical models.

### 2.3 Automatic Qualitative Evaluation of VF Tasks

Verma and Howard (2012) showed that pathological semantic organization of speech is an effective proxy for underlying cognitive impairment in early AD—MCI. As a result, MCI screening from the SVF has leveraged a variety of computational models of semantic coherence across many languages. Early approaches for automatic semantic modeling of the SVF relied on classic co-occurrence measures for capturing AD-related semantic SVF markers (Clark et al., 2016; Pakhomov et al., 2012), graph-based measures (Lerner et al., 2009), or employed latent semantic analysis (Pakhomov and Hemmy, 2014; Pakhomov et al., 2015).

Most recently, semantic word embeddings have been used for automatic evaluation of verbal fluency tasks, including the SVF (Linz et al., 2017b; Paula et al., 2018; Kim et al., 2019; Lindsay et al., 2021a). For MCI screening, encouraging results were obtained with a variety of semantic NLP resources including word2vec (Linz et al., 2017b; König et al., 2018), WordNet (Paula et al., 2018), and Wikipedia backlink vector space models (Kim et al., 2019); Paula et al. (2018) and Linz et al. (2017a) reported classification performances of *AUC* 0.71 with a random forest classifier and *F1* 0.77 with a support vector machine, respectively.

While the type of embedding was found to significantly influence classification performance (Linz et al., 2017b; Paula et al., 2018), an approach combining different embedding types was not presented. Similarly, studied languages include French (Linz et al., 2017a), Korean (Kim et al., 2019), English (Pakhomov and Hemmy, 2014) and Brazilian Portuguese (Paula et al., 2018), but to our knowledge, no multilingual classification was investigated.

We argue that by extracting qualitative SVF features with multiple language-specific resources, we can train machine learning models across languages. Overcoming the issue of small clinical data sets and possibly building more robust models that generalize cognitive impairment that is not language-specific.

## 3 Methodology

### 3.1 Data

This study included SVF data from clinical datasets in three languages; French collected at Nice Institut Claude Pompidou Memory Clinic in France; German collected at the University Medical Centre Freiburg, Germany; and Dutch from Maastricht University Clinic, Netherlands. All participants performed a 60-second SVF for the category "animals" in their native language—in addition to a battery of cognitive tests—administered by a clinician. The recordings were manually corrected according to the CHAT protocol (MacWhinney, 1991; Karakostas et al., 2017; Tröger et al., 2017).

For all corpora, participants were excluded if they presented with comorbidities (e.g. apathy or depression). To control for confounding cognitive factors, samples from healthy controls (HC) and those with mild cognitive impairment (MCI) were matched for age and education in each language using the MatchIt package in R (Ho et al., 2011). The resulting demographic information for each corpus is listed in Table 1. A wilcoxon non-parametric test is reported to check for differences in age and education between HC and MCI. All described studies were approved by national ethical committees and conform to the Declaration of Helsinki.

### 3.2 Embedding Resources

As the SVF does not evaluate language abilities but rather underlying processes of executive function and memory, we made use of non-contextual word embeddings. To keep results comparable and generalizable for future studies, we used pretrained mod-

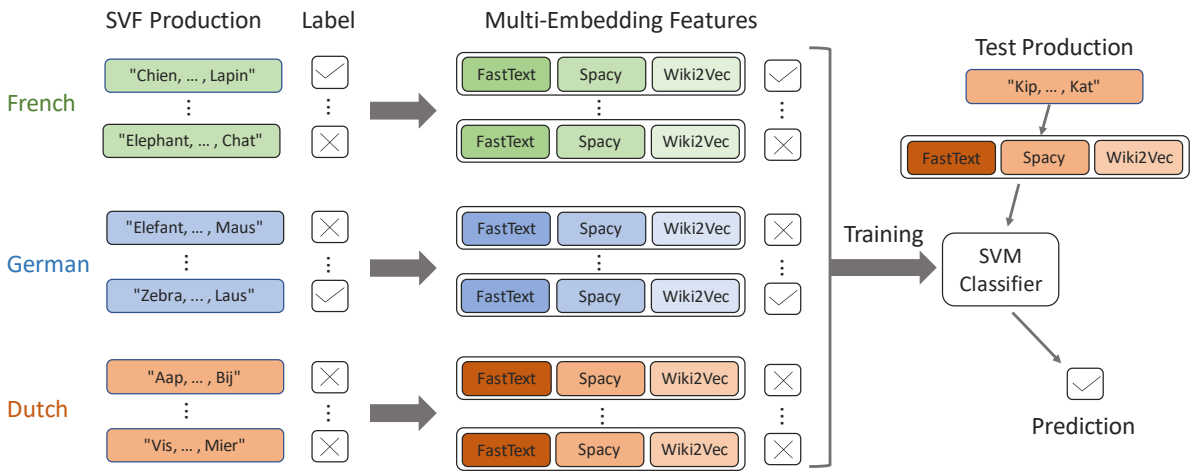


Figure 1: Overview of our Multi-Embedding Multilingual Learning framework. The training data consists of SVF productions with MCI labels in different languages (French, German, Dutch). For each training sample, features are extracted using multiple embedding resources (FastTest, Spacy, Wiki2Vec). With this training data, we learn a SVM classifier that is able to predict MCI versus HC at test time on SVF productions from any of the three languages.

els that did not require fine-tuning and were available in French, German and Dutch. Concretely, our approach integrated three different semantic word embeddings: FastText (Bojanowski et al., 2016b), Spacy (Honnibal et al., 2020), and Wiki2Vec (Yamada et al., 2020b). **FastText** models were trained using character n-gram models making them robust against out of vocabulary words. However, words shorter than the window of five characters could still go unrecognized. **Spacy** models used the same algorithm and training data as FastText models but contained much fewer key pairs (2,000,000 versus 500,000). **Wiki2Vec** combined three jointly optimized submodels; a word-based model and two models that represent semantic association using links between wikipedia pages which could inform the semantic relationships of the SVF task (Yamada et al., 2018). For semantic embedding parameters, please see the supplementary materials.

### 3.3 Clustering-Based Features

The implementation for determining clusters using semantic embeddings followed Linz et al. (2017a). Each participant's SVF production was transcribed and preprocessed into a sequence of only animal words represented by $a_1, ..., a_n$. A base threshold $T_p$ is determined for each participant $p$ by averaging the semantic similarity between all pairs of animal words in $p$'s production.

$$T_p = \frac{1}{n(n-1)} \sum_{i,j=1...n, i \neq j} sim(a_i, a_j)$$

Semantic similarity $sim$ was measured by the cosine distance between semantic embedding vectors $e_i$ extracted from words $a_i$, i.e. $sim(a_i, a_j) = \cos(e_i, e_j)$. Clusters were determined by comparing the semantic similarity of consecutive words $sim(a_i, a_{i+1})$ in the production to $T_p$. If the consecutive words were more similar than the base similarity threshold they were considered to belong to the same cluster. If the consecutive words were less similar than the base similarity threshold they introduced a cluster boundary, also referred to as a switch.

Based on the clusters obtained from a given participant with a given embedding, we computed the following features based on Linz et al. (2017a):

**Mean cluster size** computed as the average number of words in a cluster, **number of switches** calculated as the number of clusters minus 1, **mean cluster distance** computed as the average semantic distance between all words in a cluster, and **mean switch distance** as the average semantic distance between centroids of adjacent clusters.

### 3.4 Multilingual Approaches

To combine multilingual data, we investigate two approaches. Section 3.4.1 proposes a method using available language-specific resources for each language and the section 3.4.2 translates all of the data to a common language, English.

#### 3.4.1 Untranslated Multilingual, Multi-Embedding Approach

The central idea underlying the untranslated multilingual, multi-embedding approach is to maximize the available clinical data by using generalizable semantic features that robustly model cognitive impairment. To mitigate the possibility of fluctuating performance between language and embedding type, we propose using multiple embeddings for untranslated, multilingual data.

#### 3.4.2 Translated to Common Language Approach

An alternative way to make use of training data from several source languages is to translate all SVF productions to a common language prior to feature extraction. We follow the methodology in (Paula et al., 2018) and translate all SVF productions to English before extracting word embeddings. For translation we first used Google translation API[2] and then manually checked and post-edited any words where the source word was identical to the target word. Due to privacy restrictions on medical data, a set of all mentioned animal names was extracted from the transcripts and a look-up dictionary was created mapping the animals of each language to its English equivalent.

### 3.5 Classification Experiments

From each of the French, Dutch and German productions as well as their English translations, the four described clustering-based features are extracted using each respective embedding resource.

#### 3.5.1 Multilingual, Multi-embedding (ML-ME)

Figure 1 gives an overview of the multilingual, multi-embedding framework. For both the untranslated and translated approaches, the four features of underlying cognition are extracted. Each of the features vectors are concatenated into a single feature vector. This is the new representation of the SVF production that is then used to train the model and predict a label of HC or MCI.

#### 3.5.2 Baseline Comparisons

**Single Language, Single Embedding (SL-SE)** To test how well each embedding resource models each language, we trained on each combination of language and embedding resource individually.

**Multilingual, Single Embedding (ML-SE)** To investigate how each embedding resource behaves in a multilingual training scenario, we trained a separate model for each embedding resource using all the language corpora.

#### 3.5.3 Out-of-Vocabulary (OOV) Rate

In addition to the classification experiments, the out of vocabulary rate for each language for each embedding is considered. This is used as a quality control test to ensure words are not being dropped from the transcript when the features are being computed. The OOV rate is calculated as the unique number of word that are not in the semantic model divided by total produced animal words.

### 3.6 Evaluation

In line with previous work (König et al., 2018), classification was performed by a Support Vector Machine (SVM) with Radial Basis Function kernel implemented in sci-kit learn[3] (Pedregosa et al., 2011), using default parameters for $\gamma$ and $C$. To maximize the amount of available data, testing for each model was done via leave-one-out cross-validation. Model performance was measured as *area under the receiver operator curve* (AUC). In the multilingual cases, a language-specific AUC was reported, where the multilingual model is evaluated separately on each target language. To compare the multilingual methods to the other approaches, AUC scores were averaged across the languages.

To nullify the effects of random initialization of the SVM optimization, we averaged AUC values obtained from 50 random initializations. To further test quality of the word embeddings, the rate of out of vocabulary words (OOV Rate) was reported as the percentage of words that did not have an emebedding vector in the specific model.

## 4 Results

All results described in Section 3.6 are in Table 2 and Table 3. Table 2 displays the OOV rate analysis and Table 3 contains classification results.

A baseline was created using single language, single embedding (SL-SE) classifications. In the untranslated approach, Dutch had the lowest average across embeddings with an average AUC of 0.29, then French with 0.58, and finally German with 0.55. Similarly, Dutch had the lowest value (average AUC=0.38) across resources in the

[2]https://cloud.google.com/translate

[3]sci-kit learn version 0.24.0 for Python 3.7

| | Embedding | French | German | Dutch | English | Lang AVG |
|---|---|---|---|---|---|---|
| OOV Rate | FastText | 0.0 | 0.04 | 0.0 | 0.0 | 0.01 |
| | Spacy | 0.0 | 0.07 | 0.0 | 0.0 | 0.018 |
| | Wiki2Vec | 0.0 | 0.001 | 0.0 | 0.0 | $\leq 0.001$ |

Table 2: Out of Vocabulary Rate across embeddings

| | | Untranslated | | | | Translated | | | |
|---|---|---|---|---|---|---|---|---|---|
| Approach | Embedding | French | German | Dutch | Lang AVG | French | German | Dutch | Lang AVG |
| SL-SE | FastText | 0.52 | 0.64 | 0.24 | 0.47 | 0.44 | 0.35 | 0.48 | 0.42 |
| | Spacy | 0.59 | 0.63 | 0.38 | 0.53 | 0.45 | 0.38 | 0.31 | 0.38 |
| | Wiki2Vec | 0.64 | 0.39 | 0.24 | 0.42 | 0.60 | 0.51 | 0.36 | 0.49 |
| ML-SE | FastText | 0.63 | 0.68 | 0.52 | 0.61 | 0.59 | 0.52 | 0.48 | 0.53 |
| | Spacy | 0.68 | 0.68 | 0.46 | 0.61 | 0.60 | 0.57 | 0.53 | 0.57 |
| | Wiki2Vec | 0.62 | 0.56 | 0.59 | 0.59 | 0.63 | 0.60 | 0.69 | 0.64 |
| ML-ME | All | 0.66 | 0.68 | 0.63 | 0.66 | 0.62 | 0.59 | 0.64 | 0.62 |

Table 3: Averaged AUC results for the Multilingual, Multi-embedding model and Baseline approaches. Single Embedding (SE), Multi-emebedding (ME), Single Language (SL), Multilingual (ML), Average (AVG), Out of Vocabulary (OOV). Cross-Language AVG is the average AUC performance for the values in the row.

translated approach, then German (0.41) and finally French (0.50). No single embedding type showed consistent best performances. In the untranslated approach, French and Dutch performed best with Spacy, and German performed best with FastText. In the translated approach, French and German achieved their best performance with English Wiki2Vec embeddings, whereas the Dutch data worked best with FastText embeddings. The overall finding from the SL-SE baseline showed that no single embedding type performed best over the setting.

In a next step, we combine the datasets to create a multilingual training scenario for each of the embedding types (ML-SE). In both approaches, every classification improves with the multilingual data with the exception of the French Wiki2Vec embeddings in the untranslated case. To make more meaningful comparisons to the SL-SE and ML-SE cases, we aggregate over the languages for each embedding type and report a cross-language average (shown in the table as Lang AVG). We then compare the cross language averages in the single language and multilingual scenarios. In both the untranslated and translated scenarios we see overall improvement. In the untranslated case, we see an average improvement of 12 AUC points, with the largest improvement coming from Wiki2Vec (16 AUC points). In the translated case, using the combined data, we see an average improvement of almost 15 AUC points.

In the case of the untranslated data, we see the largest overall improvement in the multilingual, multi-embedding scenario. Averaging over the cross language averages (Lang AVG) of the ML-SE scenario produces an AUC of 0.60, while the ML-ME average reaches an AUC of 0.66. For the ML-ME scenrios in the translated case, by averaging across the ML-SE models we achieve a comparable AUC of 0.58 which is outperformed by the ML-ME model (0.62). However, in the translated case, the best results are produced in the multilingual scenario using Wiki2Vec embeddings (0.64).

In addition to the classification analysis, we investigated the Out-of-Vocabulary (OOV) rate for each of the embeddings in each of the languages as a form of quality control. The results in Table 2 show that our embeddings were suited for the task. Overall, we had no OOV words for French, English and surprisingly Dutch. However, German animals seem to be lacking from models of equivalent size. This could also be due to language-specific differences in morphology.

## 5 Discussion

In this study we investigate two different methods of combining multilingual data to build clinical models to distinguish between healthy controls and early signs of Alzheiemer's Disease (MCI); untranslated and translated.

While the multi-embedding method is best for when data is kept multilingual, if the data is translated and no longer in need of multilingual re-

sources, a single embedding type did emerge with the best performance, Wiki2Vec. Given this, in the case where a common language (especially English) can be achieved, according to our data, it may be best to find and use one embedding type.

In addition, we found that embedding type does make a difference in classification performance. Therefore, caution should be used when deciding on semantic resources. For instance, in the untranslated case, if we were to build a multilingual model and only use Spacy emebeddings, we would have relatively good performing classifier in French and German but the Dutch classification would not exceed chance performance. While combining embeddings may not yield the best results for each individual language, it results in the most uniform improvement in a multilingual—versus translated—setting.

However, translating the data to English, drastically improved Dutch performance, specifically with the Wiki2Vec models. Speculatively, the improved overall performance of the English translations with Wiki2Vec could be due to the backlink model where relationships are modeled through linked wikipedia page, situating Wiki2Vec to be very useful in modelling these semantic relationship from a cognitive standpoint in verbal Fluency tasks. However, based on these results, it seems that these relationships are mainly found in the English Wiki2Vec model, most likely due to the large discrepancy in the amount of available training data between the languages.

Beyond just the brand of embedding, there are pros and cons that come with each the untranslated and translated approach. By translating the data to English, we introduce possible errors based on how the data is translated. In this study, we chose to combine an automatic approach with a manual post-editing step, making the translated approach not fully automatic. From a clinical perspective, we do not know if the previous work on cognition applies to data that is translated to another language and then assessed. However, from a computational standpoint, if a reliable translation service for the source language to English exists, using the monolith of English resources presents as a reliable and effective alternative.

There are many challenges that arise when trying to concatenate data from multiple sources, thus specific caution should be taken on how to model data that has health implications. Our investigation of the two approaches (untranslated and translated) shows that SVF speech data can be combined to achieve results comparable to previous models. As no unified benchmark exists for HC vs. MCI detection from the SVF, our results can only cautiously be compared with previous work. However, we noticed that our best results from our best models for French(0.66), German(0.68) and Dutch (0.69) are in line with reported AUC values for French (0.76 (König et al., 2018)) and Spanish (0.75 (Paula et al., 2018)). It is worth noting, that these results are achieved without using the overall word count, which typically the strongest indicator for MCI detection from the SVF task.

## 6 Conclusion

Using multilingual cognitive data in both a untranslated multilingual, multi-embedding approach and translated to a common language approach improved classification over single language baselines.

This is promising not only for the feasibility of increasing the size of small clinical datasets in quick and cost-effective way, but it also opens the door for methodology on how we can use multilingual data to build more robust understanding of underlying cognitive conditions (Lindsay et al., 2021b; Fraser et al., 2019).

Future work should look at exploratory analysis for the compatibility of features computed from translated transcripts in the current clinical understanding. This could present translation as viable option for low-resource languages, or taking advantage of larger resources, while still presenting explainable clinical solutions and improved classification performance.While the languages in this study are all in the same language family, this methodology should be tested with data from different language families to test for robustness of the solution.

As such, our proposed methodology provides insight into the effect of NLP resources for classifications on cognition as well as a tentative solution to the problem of combining multiple clinical datasets. This addresses the issue of small clinical data sets as well as opens the door for building robust models of cognition for clinically actionable solutions leveraging multilingual data, paving the way towards reaching the societal goal of cost-effective early AD detection.

## References


Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016a. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016b. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

D. G. Clark, P. M. McLaughlin, E. Woo, K. Hwang, S. Hurtz, L. Ramirez, J. Eastman, R. M. Dukes, P. Kapur, T. P. DeRamus, and L. G. Apostolova. 2016. Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimers Dement (Amst)*, 2:113–122.

B. Dubois, H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, H. Bakardjian, H. Benali, L. Bertram, K. Blennow, K. Broich, E. Cavedo, S. Crutch, J. F. Dartigues, C. Duyckaerts, S. Epelbaum, G. B. Frisoni, S. Gauthier, R. Genthon, A. A. Gouw, M. O. Habert, D. M. Holtzman, M. Kivipelto, S. Lista, J. L. Molinuevo, S. E. O'Bryant, G. D. Rabinovici, C. Rowe, S. Salloway, L. S. Schneider, R. Sperling, M. Teichmann, M. C. Carrillo, J. Cummings, and C. R. Jack. 2016. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimers Dement*, 12(3):292–323.

Kathleen C. Fraser, Nicklas Linz, Bai Li, Kristina Lundholm Fors, Frank Rudzicz, Alexandra König, Jan Alexandersson, Philippe Robert, and Dimitrios Kokkinakis. 2019. Multilingual prediction of Alzheimer's disease through domain adaptation and concept-based language modelling. 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 3659–3670, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Anastasios Karakostas, Alexia Briassouli, Konstantinos Avgerinakis, Ioannis Kompatsiaris, and Magda Tsolaki. 2017. The dem@care experiments and datasets: a technical report. *CoRR*, abs/1701.01142.

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

A. König, N. Linz, J. Töger, M. Wolters, J. Alexandersson, and P. Robert. 2018. Fully automatic analysis of semantic verbal fluency performance for the assessment of cognitive decline. *Dementia and Geriatric Cognitive Disorders*. Accepted.

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

Hali Lindsay, Philipp Mueller, Nicklas Linz, Radia Zeghari, Mario Magued Mina, Alexandra König, and Johannes Tröger. 2021a. Dissociating semantic and phonemic search strategies in the phonemic verbal fluency task in early dementia. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 32–44.

Hali Lindsay, Johannes Tröger, and Alexandra König. 2021b. Language impairment in alzheimer's disease—robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. *Frontiers in aging neuroscience*, 13:228.

Nicklas Linz, Johannes Tröger, Jan Alexandersson, and Alexandra König. 2017a. Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

Nicklas Linz, Johannes Tröger, Jan Alexandersson, Maria Wolters, Alexandra König, and Philippe Robert. 2017b. Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 719–728.

Brian MacWhinney. 1991. *The CHILDES project: Tools for analyzing talk.* Lawrence Erlbaum Associates, Inc.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

Serguei VS Pakhomov and Laura S Hemmy. 2014. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex*, 55:97–106.

Serguei VS Pakhomov, Laura S Hemmy, and Kelvin O Lim. 2012. Automated semantic indices related to cognitive function and rate of cognitive decline. *Neuropsychologia*, 50(9):2165–2175.

Serguei V.S. Pakhomov, Susan E. Marino, Sarah Banks, and Charles Bernick. 2015. Using Automatic Speech Recognition to Assess Spoken Responses to

Cognitive Tests of Semantic Verbal Fluency. *Speech Communication*, 75:14–26.

Felipe Paula, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. Similarity measures for the detection of clinical conditions with verbal fluency tasks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 231–235.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ronald C Petersen, Barbara Caracciolo, Carol Brayne, Serge Gauthier, Vesna Jelic, and Laura Fratiglioni. 2014. Mild cognitive impairment: a concept in evolution. *Journal of internal medicine*, 275(3):214–228.

Johannes Tröger, Nicklas Linz, Jan Alexandersson, Alexandra König, and Philippe Robert. 2017. Automated Speech-based Screening for Alzheimer's Disease in a Care Service Scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*.

Johannes Tröger, Nicklas Linz, Alexandra König, Philippe Robert, and Jan Alexandersson. 2018. Telephone-based dementia screening i: automated semantic verbal fluency assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 59–66.

Angela K Troyer, Morris Moscovitch, and Gordon Winocur. 1997. Clustering and Switching as Two Components of Verbal Fluency: Evidence From Younger and Older Healthy Adults. *Neuropsychology*, 11(1):138–146.

Johannes Tröger, Nicklas Linz, Alexandra König, P. Robert, Jan Alexandersson, Jessica Peter, and Jutta Kray. 2019. Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in alzheimer's disease. *Neuropsychologia*, 131.

Malvika Verma and Robert J Howard. 2012. Semantic memory and language dysfunction in early alzheimer's disease: a review. *International journal of geriatric psychiatry*, 27(12):1209–1217.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

David L. Woods, John M. Wyma, Timothy J. Herron, and E. William Yund. 2016. Computerized Analysis of Verbal Fluency: Normative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury. *PLOS ONE*, 11(12):1–37.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020a. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020b. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2018. Wikipedia2vec: An optimized tool for learning embeddings of words and entities from wikipedia. *CoRR*, abs/1812.06280.