Multilingual Prediction of Alzheimer's Disease Through Domain Adaptation and Concept-Based Language Modelling

Kathleen C Fraser, Nicklas Linz, Bai Li, Kristina Lundholm Fors, Frank Rudzicz, Alexandra König, Jan Alexandersson, Philippe Robert, Dimitrios Kokkinakis

Multilingual Prediction of Alzheimer's Disease Through Domain Adaptation and Concept-Based Language Modelling

Kathleen C. Fraser[1], **Nicklas Linz**[2], **Bai Li**[3], **Kristina Lundholm Fors**[4], **Frank Rudzicz**[3,5], **Alexandra König**[6], **Jan Alexandersson**[2], **Philippe Robert**[6], **Dimitrios Kokkinakis**[4]
[1]National Research Council Canada, Ottawa, Canada
[2]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[3]University of Toronto and Vector Institute, Toronto, Canada
[4]University of Gothenburg, Gothenburg, Sweden
[5]St. Michael's Hospital, Toronto, Canada
[6]University of Côte d'Azur, Nice University Hospital, and INRIA, Nice, France
kathleen.fraser@nrc-cnrc.gc.ca, nicklas.linz@dfki.de,
bai@cs.toronto.edu, kristina.lundholmfors@gu.se,
frank@cs.toronto.edu, alexandra.konig@inria.fr,
jan.alexandersson@dfki.de, probert@unice.fr,
dimitrios.kokkinakis@gu.se

Abstract

There is growing evidence that changes in speech and language may be early markers of dementia, but much of the previous NLP work in this area has been limited by the size of the available datasets. Here, we compare several methods of domain adaptation to augment a small French dataset of picture descriptions ($n = 57$) with a much larger English dataset ($n = 550$), for the task of automatically distinguishing participants with dementia from controls. The first challenge is to identify a set of features that transfer across languages; in addition to previously used features based on *information units*, we introduce a new set of features to model the order in which information units are produced by dementia patients and controls. These concept-based language model features improve classification performance in both English and French separately, and the best result (AUC = 0.89) is achieved using the multilingual training set with a combination of information and language model features.

1 Introduction

According to the World Health Organisation, the largest global challenge facing the world today is the rapid increase of the population aged over 65 years. It is projected to increase from 524 million in 2010 to 1.5 billion in 2050, with the largest increase in the developing world (Suzman and Beard, 2011). This demographic trend has profound societal implications; for example, the number of persons affected by dementia will increase worldwide from 46 million in 2015 to 131.5 million in 2050 (Prince et al., 2015). The most common underlying condition causing dementia is Alzheimer's disease (AD). Although no cure to this neurodegenerative disease has been found, experts agree that intervention in early stages is crucial to delay onset (Dubois et al., 2016).

AD is characterised by a global impairment of cognitive functioning, with specific deficits in episodic memory, executive functioning, perceptual speed and language (Bäckman et al., 2005; Weiner et al., 2008).

Machine learning experiments using speech and language for the detection of dementia or related disorders have been conducted in many languages, including English (Roark et al., 2011; Mirheidari et al., 2016; Fraser et al., 2016; Asgari et al., 2017), French (Tröger et al., 2017; König et al., 2018), German (Weiner et al., 2016), Hungarian (Szatloczki et al., 2015; Vincze et al., 2016), Spanish (Meilán et al., 2014), Greek (Satt et al., 2013), Swedish (Lundholm Fors et al., 2018; Fraser et al., 2018a), Japanese (Shibata et al., 2016), Portuguese (Aluísio et al., 2016), and Mandarin Chinese (Lai et al., 2009). Most studies acknowledge that small data sets are a limitation and describe the difficulties in gathering more data, including the challenges in patient recruitment, the expense of running clinically-based studies, and the manual effort required for transcription and annotation.

Here, we consider whether it could be possible to increase the amount of available data by augmenting a corpus in one language with data from another language, and thus improve predictive per-

Proceedings of NAACL-HLT 2019, pages 3659–3670
Minneapolis, Minnesota, June 2 - June 7, 2019.

formance without the need for new data collection. Specifically, we consider augmenting a relatively small French dataset with a much larger English one. The two aims of this study are: (1) to identify a set of features that are both useful for the detection of dementia and that we expect to transfer across different languages, and (2) to improve classification results on the French dataset by augmenting the training set with English data.

2 Background and Related Work

2.1 Narrative Analysis in AD

One way to assess language is through narrative speech, such as that elicited by the *Cookie Theft Picture* (CTP) task (Goodglass et al., 2000). In this task, participants are asked to describe the content of a line drawing of a kitchen scene, where a boy can be seen standing on a stool, trying to reach a cookie jar, while a woman is preoccupied washing dishes. In this study, we analyse CTP narratives, due to the widespread use of the task in multiple languages.

Narrative speech can be analysed on a number of levels, including phonology, morphology, syntax, semantics, and pragmatics. Here, our goal is to extract features that both predict AD and are likely to transfer across different languages. Although other studies have used acoustic features for this task (Meilán et al., 2014; König et al., 2018), there are well-documented differences in the phonology and prosody of French and English (Bertrán, 1999; Vaissière, 2002). Syntax and morphology also differ across languages, and the degree to which they are impaired in mild to moderate AD is unclear (Taler and Phillips, 2008). Pragmatic ability in AD may be disrupted (Chapman et al., 1998; Boschi et al., 2017); however, the CTP is not ideally suited for assessing pragmatics.

Instead, we focus on the semantic level, with the assumption that while the specific vocabulary will be different across languages, the underlying meanings or semantic concepts expressed should be the same. Features relating to semantic content are also motivated by the AD literature. Cuetos et al. (2007) reported a significant reduction in semantic units produced by pre-clinical AD participants, relative to controls, on the CTP task. Croisile et al. (1996) studied CTP descriptions from French participants, and found that the AD descriptions were shorter and less informative than the control descriptions. They measured information content by scoring the narratives against a gold standard list of 23 expected "information units", which have been widely used in subsequent research.

2.2 NLP for AD Classification

Several recent studies have used NLP and machine learning to analyse speech samples from people with dementia and other cognitive disorders. Most relevant here, are those which focus on picture description tasks in English or French.

DementiaBank[1] is a large database of CTP narratives from AD patients and controls, containing primarily English data. A number of recent papers report classification results on this corpus (Prud'hommeaux and Roark, 2015; Fraser et al., 2016; Al-Hameed et al., 2016; Yancheva and Rudzicz, 2016; Sirts et al., 2017). Language analysis of English-language CTP data from other sources has also been used to differentiate between different underlying pathologies in AD (Rentoumi et al., 2014), and variants of frontotemporal lobar degeneration (Pakhomov et al., 2010).

In French, picture description was one of multiple tasks used to elicit speech for the classification of participants with mild cognitive impairment and AD reported by König et al. (2015) and König et al. (2018), although only acoustic processing was used.

2.3 Multi- and Cross-Lingual NLP

There has been very little prior work on multilingual or cross-lingual dementia classification. Rentoumi et al. (2018) presented preliminary results suggesting that some language features from CTP samples could transfer across Greek and English, but did not report classification results. Fraser et al. (2018b) studied a related task of detecting mild cognitive impairment (MCI), and found that classification results could be improved in both English and Swedish by incorporating multilingual topic modelling into the feature extraction pipeline; however, they did not consider multilingual classification directly.

More generally, multilingual NLP is an active and growing area of research. Some approaches to improving classifier performance on a resource-poor *target* language by leveraging a resource-rich *source* language include: translate the target language to the source language (or vice versa) and

[1] https://dementia.talkbank.org/

train a unilingual classifier (Wan, 2009); extract features from the two languages separately and then use domain adaptation techniques to train a classifier for the target language (Blitzer et al., 2006; Prettenhofer and Stein, 2010); or determine a common representation for both languages and then extract features from the combined corpus to train a multilingual classifier (Ammar et al., 2016). In the extreme case, one can also consider purely cross-lingual classification, in which the classifier is trained solely on the source language, but tested on the target language.

We use a supervised domain adaptation approach, similar to that of Daumé III (2007), by considering each language to be a different domain. In related (though not multilingual) work, Masrani et al. (2017) also used this approach to adapt a dataset of AD narratives to their MCI classification task.

2.4 Class-Based Language Modelling

In contrast to the previous work on AD classification, we measure not only which information units are mentioned, but also the order in which they are mentioned. Our approach has some similarity to class-based language models (Brown et al., 1992), in which words are first grouped into classes (or clusters), and then the language model is trained on the classes rather than the individual words. One benefit to this approach is improved generalisability (Hoidekr et al., 2006), and another is the ability of classes to span different languages (Täckström et al., 2012).

3 Methodology

3.1 Data

Data were taken from two corpora: a small French dataset ($n = 57$), collected at the Memory Clinic and Research Centre of the University Hospital Nice, and the Pitt subcorpus of DementiaBank, containing 550 English samples[2]. Detailed information about the protocols for each study can be found in Tröger et al. (2017) and Becker et al. (1994). In both cases, ethics approval for the data collection was obtained from the local governing bodies.

The demographics for the participants in each language are shown in Table 1. In both studies, the participants were asked to perform the CTP task in their respective languages. In English, the image was shown on paper and speech was digitally recorded, while in the French study, the image was displayed on a tablet and speech was recorded via the tablet microphone.

	English		French	
	HC	AD	HC	AD
N	241	309	25	33
Gender	154F/87M	189F/120M	19F/6M	22F/11M
Age	64.8 (7.7)	71.4 (8.4)	75.4 (7.0)	79.2 (6.6)
Education	14.2 (2.6)	12.8 (3.0)	14.0 (2.6)	11.3 (4.0)
MMSE (/30)	29.1 (1.1)	19.8 (5.7)	28.6 (1.4)	18.9 (3.9)

Table 1: Demographics of participants, where AD indicates Alzheimer's disease, and HC indicates healthy control. The Mini Mental State Examination (MMSE) is global measure of cognitive status.

3.2 Features

The English and French audio samples were manually transcribed using the CHAT protocol (MacWhinney, 2014). A set of pre-defined information units found in the CTP was determined as an extension to Croisile et al. (1996), and is given in Table 2a. Mentions of information units were determined using keyword-spotting (based on manually-constructed word lists specific to each language), and used to translate the full narratives to sequences of information units. As an example, the English *A boy is standing on a stool* and French *Le garçon est sur un tabouret* would both be mapped to the sequence BOY STOOL.

Features relating to the occurrence of each distinct information unit comprise the *info* feature set, described in Table 2b. Additionally, new features are derived from language models build on the sequence of information units. To this end, concept-based language models are trained for English and French in a leave-one-out fashion, using the `kenlm` framework (Heafield, 2011). Models up to 5-grams were constructed. For each participant, two language models are constructed for each n: one trained on the healthy control (HC) population and one trained on the AD population. The participant is left out of the model built on their associated diagnostic group. The trained language models are then applied to the held-out participant's sequence of information units and various language model (*LM*) features are extracted (Table 2c).

[2]In this analysis, we included all participants in the Dementia subfolder, regardless of specific diagnosis, to maximize the size of the source data.

Actions STEAL, FALL, WASH, OVERFLOW, GIRL'S ACTION, WOMAN'S INDIFFERENCE **Actors** BOY, GIRL, CHILD(REN), WOMAN **Places** KITCHEN, EXTERIOR **Objects** COOKIE, JAR, STOOL, SINK, DISHCLOTH, WATER, WINDOW, CUPBOARD, DISH, CURTAIN, COUNTER

(a) Information units.

has_*unit* Binary feature indicating presence or absence of each information unit (23 features) **ratio_*unit*** For each information unit, the number of times that unit was mentioned, divided by the total number of words in the original narrative (23 features) **unique_concept_density** Total number of information units which were mentioned at least once, divided by the total number of words in the original narrative (1 feature) **unique_concept_efficiency** Total number of information units which were mentioned at least once, divided by the duration of the sample in seconds (1 feature) **total_concept_density** Total number of words referring to information units, divided by the total number of words in the original narrative (1 feature) **total_concept_efficiency** Total number of words referring to information units, divided by the duration of the sample in seconds (1 feature)

(b) *info* features

perplexity_*class*_*n*-gram The perplexity assigned to the sample by each of the eight language models, where $n = 2,3,4,5$, and the models are trained on data from either the AD or HC class. (8 features) **score_*class*_*n*-gram** The log probability assigned to the sample by each of the eight language models. (8 features) **max_perplexity_*class*_*n*-gram** The maximum perplexity, computed over all n-grams in a sample, for each of the eight language models. (8 features) **min_score_*class*_*n*-gram** The minimum log probability, computed over all n-grams in a sample, for each of the eight language models. (8 features)

(c) *LM* features

Table 2: Top, the information units extracted from CTP narratives. Bottom, the *info* and *LM* features that are computed from the resulting sequence of information units.

3.3 Unilingual Classification

To evaluate the performance of the three proposed feature sets (*info*, *LM*, and *info+LM*), we first train classifiers to distinguish between HC and AD participants within a given language. To examine the importance of certain features, we restrict ourselves to more explainable linear models, namely logistic regression (LR) and linear support vector machines (SVM) (Pedregosa et al., 2011). In both cases, we use L_1 regularisation to promote sparsity in the feature weights.

Area under the Receiver-Operator curve (AUC) is reported as the evaluation parameter. Due to the small size of the French dataset, we use leave-pair-out cross validation (LPO-CV), which has been shown to produce an unbiased estimate for AUC on small datasets (Airola et al., 2009), and has also been used in related work (Roark et al., 2011). However, since LPO-CV is computationally very costly, we instead use 10-fold cross-validation (10-CV) for English, making sure that any samples for a given participant occur in either the training set or the test set, but not both. For LPO-CV we compute AUC and its standard deviation as described by Roark et al. (2011); for 10-CV we compute the AUC in each test fold and then report the average and standard deviation over folds.

Feature scaling and hyper-parameter optimisation is done on the training set in each fold. Features are scaled using Maximum-Absolute Scaling to preserve the binary nature of the *info* features. For both SVMs and LR, C was optimised between $C \in [10^{-4}, ..., 10^{4}]$ using a grid search.

3.4 Multilingual Classification

Our goal is to improve classification in French, by incorporating training data from English. To this end, we examine multiple ways to combine data from both English and French in the training set.

We first consider *domain adaptation*, where we treat French as the target domain and English as the source domain. We implement the AUGMENT method of Daumé III (2007), which involves augmenting the feature space with source-specific, target-specific, and combined versions of all the original features, allowing the classifier to assign a higher weight to the combined version when that feature transfers well across domains, while also retaining source- and target-specific information where appropriate.

We consider as well as the baseline methods described in Daumé III (2007): WEIGHT, in which the samples from the source domain are assigned reduced weights in the model; PRED, in which the prediction made by the source classifier is used as an additional feature in the target model; LININT, in which the predictions from the source and target models are linearly interpolated; and ALL, in which target and source data are simply combined in a single training set. Due to the limited size of our data, we do not optimise the weighting factors in WEIGHT and LININT, but rather assume the two languages should be given equal importance, and use a weighting factor of 0.1 in WEIGHT (since the English data is 10 times the size of the French data), and 0.5 in LININT.

Another option is to combine the French

and English datasets before extracting features. Specifically, we first replace the word-level transcripts with the sequence of information units, and then combine the two datasets and train the language models over the multilingual corpus, thus generating *multilingual language models*.

3.5 Cross-Lingual Classification

To understand how well a trained classification model in one language could be applied to another, we also perform cross-lingual experiments. For this, we train language and classification models in one language and test it on the other.

4 Results

The results of the classification experiments are presented in Figure 1.

4.1 Unilingual Classification

In French, for both LR and SVM, using *LM* features leads to higher AUC than the *info* features, and the combination of features is more effective than either feature set alone. In the English case, the *LM* and *info* features lead to equivalent performance individually, but the AUC is again marginally improved when the feature sets are combined, suggesting that they are capturing at least somewhat complementary information.

4.2 Domain Adaptation Results

For French, the *LM* features generally do not benefit from domain adaptation, with equivalent or poorer AUC relative to the unilingual case. The best result with the *LM* features is achieved in the AUGMENT scenario, where the classifier can select the French *LM* features only (although this result holds only for the SVM classifier). In contrast, the *info* features do benefit from the additional data available through domain adaptation, and lead to better results than the unilingual baseline. The best overall result of AUC = 0.89 is achieved by combining the feature types in the ALL configuration.

For English, we do not expect to see much benefit from including the (much smaller) French dataset. The WEIGHT adaptation technique is not feasible when the source data is smaller than the target data, and the LININT technique performs poorly, as it assigns too much importance the smaller and out-of-domain dataset. However, we do see marginal improvements using ALL and AUGMENT, reflecting the value of increasing the training set size by roughly 10%. The best result of AUC = 0.84 is achieved in the ALL condition, using the combined feature set.

4.3 Multilingual LM Results

Using the multilingual LM does not affect the *info* features, and therefore Figure 1 shows only the *LM* and *info+LM* results. Clearly, the multilingual LM approach does not work well here. Unlike in domain adaptation, combining the datasets using this method assumes that information units will be produced in the same order in the two languages. While French and English are similar in this respect, there are many possible counter-examples, such as *cookie jar* (COOKIE JAR) versus *boîte à biscuits* (JAR COOKIE).

4.4 Cross-Lingual Classification

When training entirely on English data and testing on French, the results using *info* and *info+LM* features are significantly improved over the unilingual baseline, while the *LM* results are reduced, once again indicating that the *info* features transfer better across languages. The results are very similar to those using the ALL technique for domain adaptation, suggesting that in that case, model training is dominated by the English data.

To further explore the similarity in performance in the ALL and cross-lingual cases, we examine the effect of incrementally increasing the amount of English data in the training set, when testing on French data. Figure 2 displays the classification performance of SVM and LR classifiers trained either using the ALL method of domain adaptation or cross-lingually with increasing amounts (10% at a time) of the English data. Considering first the ALL method (red and blue), at $x = 0$ there is no English data, and so we recover the French unilingual baseline. As we increase the amount of English data in the training set, performance slowly increases, eventually reaching the values reported in Figure 1. Considering next the cross-lingual case (yellow and green), we see that training on only 10% of the English data (55 samples) results in much poorer AUC values. However, each further 10% increases the classification performance. At 80% of English data (440 samples) the multi- and cross-lingual cases converge in performance. Thus, it would appear that domain adaptation is more data-efficient, as we achieve close to optimal results with a smaller proportion of English data,

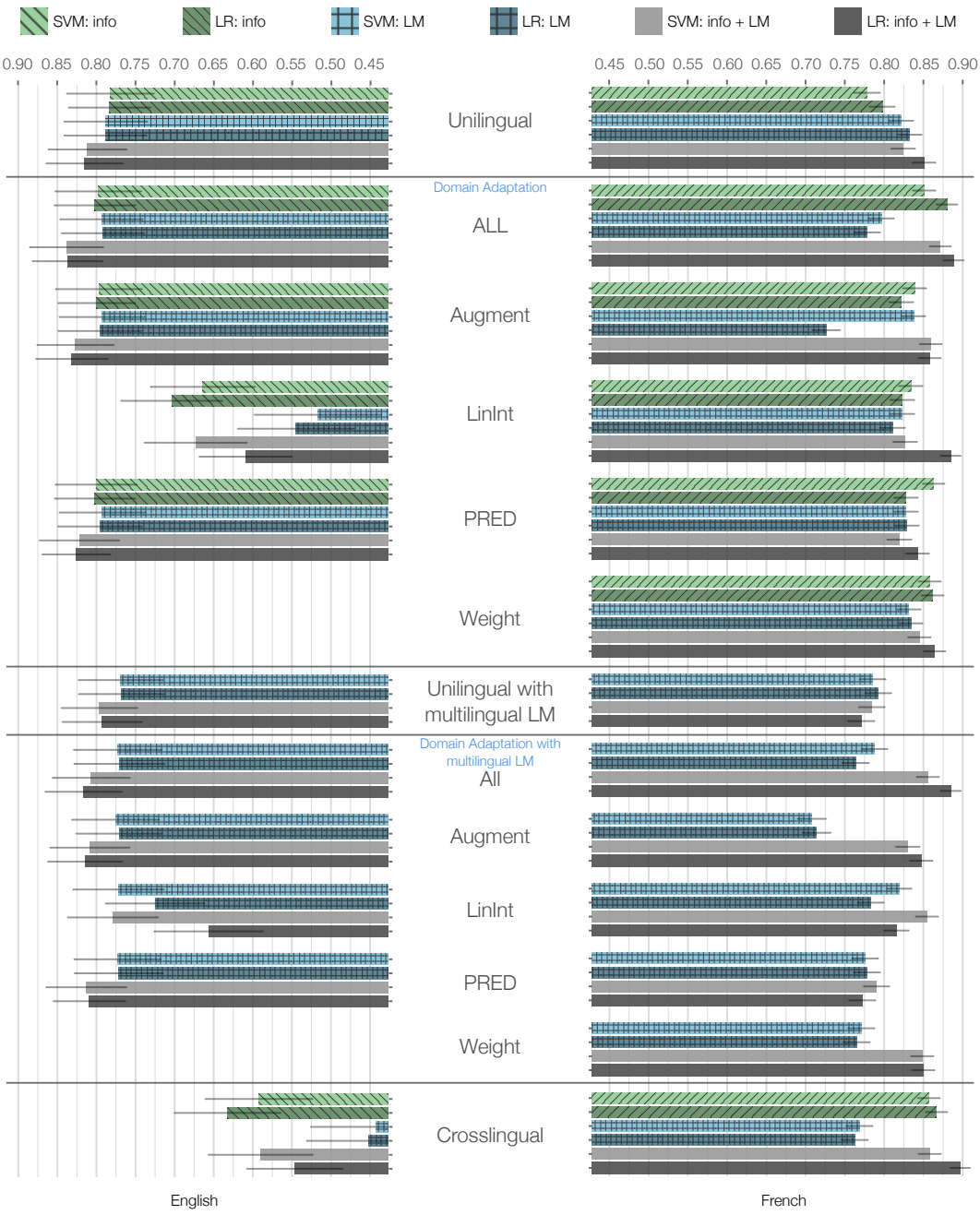

Figure 1: Results of uni-, multi- and cross-lingual classification experiments. Left panel displays results for English, right panel for French. Labels in the middle indicate the classification scenario and method of domain adaptation. Colours indicate the feature set and classifier. Bars indicate the AUC; error bars represent standard deviation.

but that the cross-lingual approach can be equally effective, given a large enough corpus.

4.5 Feature Analysis

Finally, we examine the features to determine which features are most useful to the task of dementia detection, and to compare the selected features in the unilingual and multilingual cases. Figure 3 shows the median absolute value of the weights assigned to each feature, for English and French, in the unilingual and multilingual ALL condition. The L_1 regularisation serves to set many feature weights to zero.

As a high-level observation, in both the uni- and multilingual cases, relatively more *info* features are selected, and relatively fewer *LM* features. Of the *LM* features that are selected, those which relate to the maximum perplexity or minimum probability appear to be more useful. These features

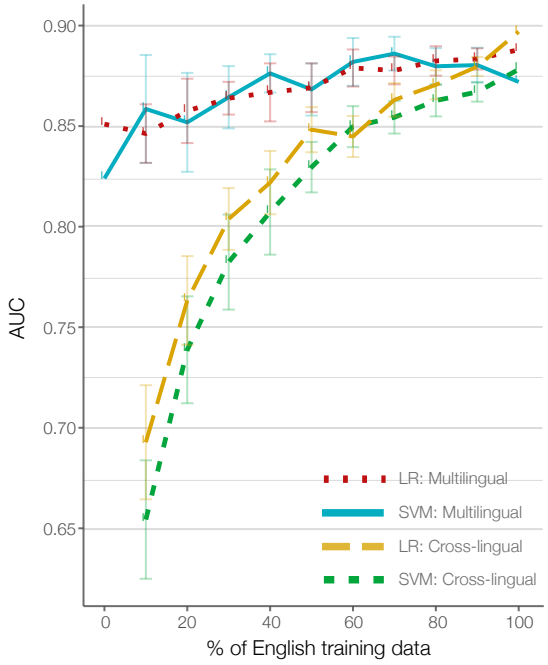

Figure 2: *AUC* as a function of the amount of English data used in the training set, for both multi- and cross-lingual cases. Error bars indicate 95% confidence intervals.

capture locally anomalous speech patterns, relative to either the AD or control language models.

In the unilingual case, the French models show a preference for the binary "has" features (indicating whether or not an information unit has been mentioned). Only 4 of the "ratio" features and none of the density or efficiency features have a median value greater than zero. However, these features *are* relevant to the task, and potentially more generalisable (e.g., total concept efficiency differs between the French AD and HC groups with $p < 0.001$ on a t-test, and represents an aggregate score rather than depending on the presence or absence of a single information unit). Such features are selected more often in the multilingual case, and lead to improved performance. One explanation for this could be that in the small French training set, spurious correlations due to noise can overpower the real signal, and lead to less relevant features being assigned high weights, while correlated (but perhaps actually more relevant) features are suppressed. By increasing the size of the training set with English data, the signal-to-noise ratio is improved, and a better set of features is selected.

Generally, the feature values (not shown) support the intuition that controls mention more of the information units in the image (higher "has" feature values), convey information more efficiently, with fewer off-topic words (higher density and efficiency scores), and organize the narrative in a more predictable way (narratives have lower perplexity and higher probability) than the AD participants. Again, these trends are more apparent in the English data than the French data, likely due to the relatively larger number of samples.

5 Discussion

One perhaps surprising result of this study was that naively combining features in the ALL condition led to better results than the AUGMENT algorithm. However, this is in line with the original findings of Daumé III (2007), where he identified a set of tasks where AUGMENT performed sub-optimally: specifically, those cases where training on source-only data was better than training on target-only data. This is precisely the case we have here, as training cross-lingually (on English source data) leads to better results than training unilingually (on French target data). The explanation offered by Daumé III is, "If the domains are so similar that a large amount of source data outperforms a small amount of target data, then it is unlikely that blowing up the feature space will help." In some sense, then, these results are confirmation that we have indeed identified a set of features over which the two languages (i.e. domains) are very similar.

The fact that the ALL configuration is optimal in both French and English has an added practical benefit: since there is no distinction between source and target features, the resulting classifier is language-agnostic. This means that test data could come from either language, in a hypothesized future screening application.

While our goal in this paper was not to push the state-of-the-art on the DementiaBank dataset, we do find that our best English result (AUC=0.84, which corresponds to an accuracy of 75% and F_1 score of 0.77) is comparable to the other published results on this dataset (Prud'hommeaux and Roark, 2015; Yancheva and Rudzicz, 2016; Sirts et al., 2017; Fraser et al., 2016; Hernández-Domínguez et al., 2018). There are no previously published results on the French dataset.

6 Conclusion and Future Work

In this work, we have shown that there are features which can both distinguish AD patients from healthy controls with a high degree of accuracy,

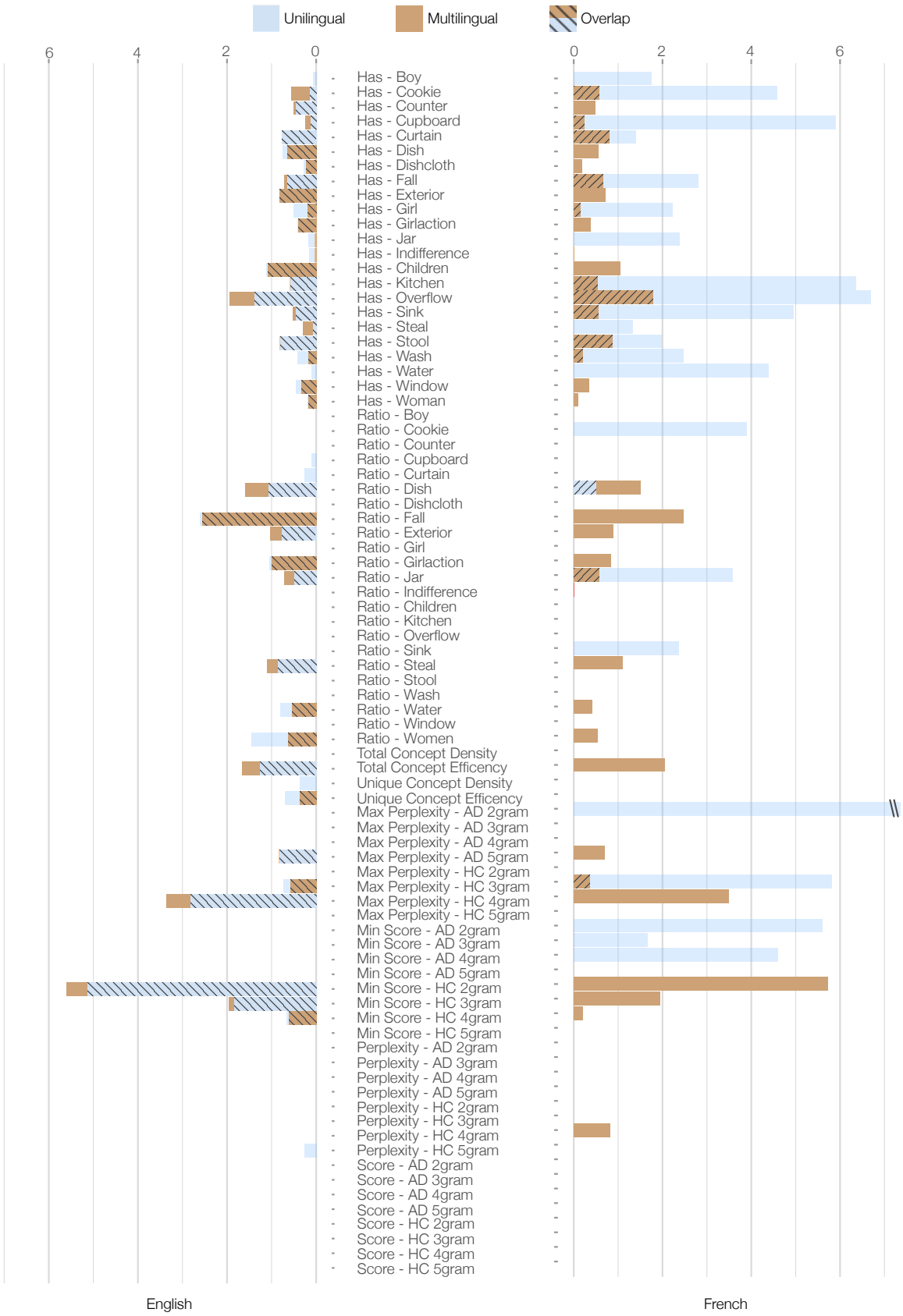

Figure 3: Visualisation of feature weights for uni- and multilingual experiments. Median feature importances over LPO- and 10-CV are displayed. The left panel displays the English and the right panel the French data sets. Unilingual experiments are given in blue and multilingual in yellow.

and also generalize across languages. By incorporating a large English dataset, we were able to improve the AUC on the French dataset from 0.85 to 0.89. We also developed a new set of features for this task, using concept-based language modelling, which improved AUC from 0.80 to 0.85 in the unilingual case, and 0.88 to 0.89 in the multilingual case.

Future work will involve extending the set of features involved, incorporating data from other languages, and testing whether similar techniques can be effective for detecting earlier stages of cognitive decline, such as MCI. Other work from our group has also begun to explore the use of unsupervised methods and out-of-domain data sources (Li et al., 2019).

Technical challenges aside, collaborations of this nature can be difficult due to the sensitive nature of the data, and the need to respect ethical guidelines and participant consent when sharing and storing data. With this in mind, we recommend to other researchers working in similar domains to consider from the outset whether their data could eventually be shared, and to make suitable provisions in their ethics protocols and participant consent forms. We look to DementiaBank as a model for this kind of data-sharing and openness, and hope that researchers can continue to find ways to share resources of this nature.

Acknowledgements

This research was partially funded by the Riksbankens Jubileumsfond – The Swedish Foundation for Humanities and Social Sciences, grant no: NHS 14-1761:1, and the EIT Digital Wellbeing Activity 17074, ELEMENT. The French data was collected during the ELEMENT project and the FP7 Dem@Care project (grant number 288199). The original acquisition of the DementiaBank data was supported by NIH grants AG005133 and AG003705 to the University of Pittsburgh, and the data archive is supported by NIHNIDCD grant R01-DC008524 to Carnegie Mellon University.

References

Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski. 2009. A comparison of AUC estimators in small-sample studies. In *Machine Learning in Systems Biology*, pages 3–13.

Sabah Al-Hameed, Mohammed Benaissa, and Heidi Christensen. 2016. Simple and robust audio-based detection of biomarkers for Alzheimer's disease. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 32–36.

Sandra Aluísio, Andre Cunha, and Carolina Scarton. 2016. Evaluating progression of Alzheimer's disease by regression and classification methods in a narrative language test in Portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 109–114. Springer.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. 2017. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228.

Lars Bäckman, Sari Jones, Anna-Karin Berger, Erika Jonsson Laukka, and Brent J Small. 2005. Cognitive impairment in preclinical Alzheimer's disease: A meta-analysis. *Neuropsychology*, 19(4):520–531.

James T. Becker, François Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Antonio Pamie Bertrán. 1999. Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages. *Language Design: Journal of Theoretical and Experimental Linguistics*, 2:103–130.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics.

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

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Sandra Bond Chapman, Amy Peterson Highley, and Jennifer L Thompson. 1998. Discourse in fluent aphasia and Alzheimer's disease: Linguistic and pragmatic considerations. *Journal of Neurolinguistics*, 11(1-2):55–78.

Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. 1996. Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language*, 53(1):1–19.

Fernando Cuetos, Juan Carlos Arango-Lasprilla, Claramónika Uribe, Claudia Valencia, and Francisco Lopera. 2007. Linguistic changes in verbal expression: A preclinical marker of Alzheimer's disease. *Journal of the International Neuropsychological Society*, 13(3):433–439.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 256–263.

Bruno Dubois, Harald Hampel, Howard H Feldman, Philip Scheltens, Paul Aisen, Sandrine Andrieu, Hovagim Bakardjian, Habib Benali, Lars Bertram, Kaj Blennow, et al. 2016. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 12(3):292–323.

Kathleen C. Fraser, Kristina Lundholm Fors, Marie Eckström, Charalampos Themistokleous, and Dimitrios Kokkinakis. 2018a. Improving the sensitivity and specificity of MCI screening with linguistic information. In *Proceedings of the 2nd Workshop on Resources and processing of linguistic, paralinguistic and extra-linguistic data from people with various forms of cognitive/psychiatric impairments (RaPID)*, pages 19–26.

Kathleen C Fraser, Kristina Lundholm Fors, and Dimitrios Kokkinakis. 2018b. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, 53:121–139.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Harold Goodglass, Edith Kaplan, and Barbara Barresi. 2000. *Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT)*, pages 187–197.

Laura Hernández-Domínguez, Sylvie Ratté, Gerardo Sierra-Martínez, and Andrés Roche-Bergua. 2018. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:260–268.

Jan Hoidekr, Josef V Psutka, Ales Prazák, and Josef Psutka. 2006. Benefit of a class-based language model for real-time closed-captioning of TV ice-hockey commentaries. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 2064–2067.

Alexandra König, Aharon Satt, Alex Sorin, Ran Hoory, Alexandre Derreumaux, Renaud David, and Phillippe H Robert. 2018. Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, 15(2):120–129.

Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H Robert, and Renaud David. 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124.

Yi-hsiu Lai, Hsiu-hua Pai, et al. 2009. To be semantically-impaired or to be syntactically-impaired: Linguistic patterns in Chinese-speaking persons with or without dementia. *Journal of Neurolinguistics*, 22(5):465–475.

Bai Li, Yi-Te Hsu, and Frank Rudzicz. 2019. Detecting dementia in Mandarin Chinese using transfer learning from a parallel corpus. In *The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Kristina Lundholm Fors, Kathleen C. Fraser, and Dimitrios Kokkinakis. 2018. Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment. In *Proceedings of the Medical Informatics Europe (MIE) Conference*, pages 705–709.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Vaden Masrani, Gabriel Murray, Thalia Shoshana Field, and Giuseppe Carenini. 2017. Domain adaptation for detecting mild cognitive impairment. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 248–259.

Juan José G Meilán, Francisco Martínez-Sánchez, Juan Carro, Dolores E López, Lymarie Millian-Morell, and José M Arana. 2014. Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.

Bahman Mirheidari, Daniel Blackburn, Markus Reuber, Traci Walker, and Heidi Christensen. 2016. Diagnosing people with dementia using automatic conversation analysis. In *Proceedings of Interspeech*, pages 1220–1224. ISCA.

Serguei VS Pakhomov, Glenn E Smith, Dustin Chacon, Yara Feliciano, Neill Graff-Radford, Richard Caselli, and David S Knopman. 2010. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23(3):165.

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.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL)*, pages 1118–1127.

Martin Prince, Anders Wimo, Maelenn Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, and Matthew Prina. 2015. World Alzheimer report 2015: The global impact of dementia. *Alzheimer's Disease International (ADI)*.

Emily Prud'hommeaux and Brian Roark. 2015. Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549–578.

Vassiliki Rentoumi, George Paliouras, Dimitra Arfani, Katerina Fragkopoulou, Spyridoula Varlokosta, and Peter Garrard. 2018. Automatic detection of linguistic indicators as a means of early prediction of Alzheimer's and of related dementias: A cross-linguistics analysis. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 14:1302.

Vassiliki Rentoumi, Ladan Raoufian, Samrah Ahmed, Celeste A. de Jager, and Peter Garrard. 2014. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease*, 42(S3):S3–S17.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffery Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090.

Aharon Satt, Alexander Sorin, Orith Toledo-Ronen, Oren Barkan, Ioannis Kompatsiaris, Athina Kokonozi, and Magda Tsolaki. 2013. Evaluation of speech-based protocol for detection of early-stage dementia. In *Proceedings of Interspeech*, pages 1692–1696.

Daisaku Shibata, Shoko Wakamiya, Ayae Kinoshita, and Eiji Aramaki. 2016. Detecting Japanese patients with Alzheimer's disease based on word category frequencies. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 78–85.

Kairit Sirts, Olivier Piguet, and Mark Johnson. 2017. Idea density for predicting Alzheimer's disease from transcribed speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 322–332.

Richard Suzman and John Beard. 2011. Global health and aging. *National Institute on Aging and National Institutes of Health, WHO*, 11-7737.

Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, Janos Kalman, and Magdolna Pakaski. 2015. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 7:1–7.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 477–487.

Vanessa Taler and Natalie A Phillips. 2008. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556.

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*, pages 292–297.

Jacqueline Vaissière. 2002. Cross-linguistic prosodic transcription: French versus English. In N.B. Volskaya, N.D. Svetozarova, and P.A. Skrelin, editors, *Problems and Methods of Experimental Phonetics. In honour of the 70th anniversary of Pr. L.V. Bondarko*, pages 147–164. St Petersburg State University Press.

Veronika Vincze, Gábor Gosztolya, László Tóth, Ildikó Hoffmann, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. 2016. Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 181–187.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 235–243.

Jochen Weiner, Christian Herff, and Tanja Schultz. 2016. Speech-based detection of Alzheimer's disease in conversational German. In *Proceedings of Interspeech*, pages 1938–1942.

Myron F Weiner, Katherine E Neubecker, Mary E Bret, and Linda S Hynan. 2008. Language in Alzheimer's disease. *The Journal of Clinical Psychiatry*, 69(8):1223.

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 (ACL)*, pages 2337–2346.