Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 53 (2018) 121–139

www.elsevier.com/locate/csl

Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment☆

Kathleen C. Fraser*, Kristina Lundholm Fors, Dimitrios Kokkinakis

University of Gothenburg, Gothenburg, Sweden

Received 14 December 2017; received in revised form 27 April 2018; accepted 26 July 2018
Available online 2 August 2018

Abstract

We analyze the information content of narrative speech samples from individuals with mild cognitive impairment (MCI), in both English and Swedish, using a combination of supervised and unsupervised learning techniques. We extract information units using topic models trained on word embeddings in monolingual and multilingual spaces, and find that the multilingual approach leads to significantly better classification accuracies than training on the target language alone. In many cases, we find that augmenting the topic model training corpus with additional clinical data from a different language is more effective than training on additional monolingual data from healthy controls. Ultimately we are able to distinguish MCI speakers from healthy older adults with accuracies of up to 63% (English) and 72% (Swedish) on the basis of information content alone. We also compare our method against previous results measuring information content in Alzheimer's disease, and report an improvement over other topic-modeling approaches. Furthermore, our results support the hypothesis that subtle differences in language can be detected in narrative speech, even at the very early stages of cognitive decline, when scores on screening tools such as the Mini-Mental State Exam are still in the "normal" range.

Keywords: Machine learning; Topic modeling; Mild cognitive impairment; Dementia; Narrative analysis; Multilingual analysis

1. Introduction

Dementia is a progressive cognitive impairment due to neurodegenerative disease, affecting more people each year as the average lifespan increases (Prince et al., 2013). The most common cause of dementia is Alzheimer's disease (AD), although other types of dementia exist. In many cases, before the impairment is severe enough to be classified as dementia, an individual may experience a phase of subjective cognitive impairment (SCI; characterized by an individual's subjective experience of cognitive decline, but with no measurable deficit observed on standardized tests), or mild cognitive impairment (MCI; characterized by a mild but clinically observable deficit in at least one cognitive domain) (Gauthier et al., 2006; Reisberg and Gauthier, 2008). Detecting potentially incipient dementia in

☆ This paper has been recommended for acceptance by R. K. Moore.

* Corresponding author.
E-mail address: kathleen.fraser@nrc-cnrc.gc.ca (K.C. Fraser).

https://doi.org/10.1016/j.csl.2018.07.005

these prodromal or preclinical phases can help patients and their families prepare and allow for early intervention. Sensitive screening tools are also of crucial importance in selecting participants for clinical trials, as current research suggests that disease-modifying medications are most likely to be successful at the earliest stages of dementia (Posner et al., 2017).

Current methods of early detection, such as positron emission tomography (PET) and magnetic resonance imaging (MRI), are expensive and invasive (Nensa et al., 2014). However, recent work has suggested that analysis of speech and language may lead to the discovery of sensitive and non-invasive behavioural biomarkers of dementia and MCI (Szatloczki et al., 2015; Laske et al., 2015; Alberdi et al., 2016; Alm, 2016). Spontaneous speech production is a complex task involving multiple cognitive domains, such as memory, attention, and planning, in addition to language itself. As a result, subtle changes in language have been observed years or even decades before dementia is diagnosed (Snowdon et al., 1996; Garrard et al., 2004; Cuetos et al., 2007; Clark et al., 2009; Le et al., 2011; Ahmed et al., 2013a).

Numerous studies have made use of a machine learning approach to automatically classify text and/or speech samples from individuals with cognitive impairment (for example, Thomas et al., 2005; Roark et al., 2011; Jarrold et al., 2014; Rentoumi et al., 2014; Garrard et al., 2014; Orimaye et al., 2014; Prud'hommeaux and Roark, 2015; König et al., 2015; Fraser et al., 2016; Asgari et al., 2017; Masrani et al., 2017). However, a major challenge in this line of research has been the relative scarcity of high-quality, clinically-validated language data on which to train such machine learning models. In this paper, we consider two possible solutions to the data scarcity problem, as it applies to the topic-modeling[1] stage in our automated processing pipeline: (1) to augment the training set with additional monolingual normative data, or (2) to augment the training set with additional multilingual clinical data. In the latter case, we take advantage of recent advances in multilingual word embeddings to generate novel multilingual information units. We then evaluate the resulting topic models on the task of distinguishing MCI speakers from healthy controls in two different languages, on the basis of the information content of participants' narrative speech.

2. Related work

Many studies have examined the relationship between cognitive decline and various measures of speech and language. In the following, we review the findings regarding the use of a picture description task to elicit speech for the purpose of detecting dementia and MCI, focusing in particular on the so-called "Cookie Theft" picture. We then review manual and automated methods for measuring the information content of the elicited narratives.

2.1. The Cookie Theft task

The Cookie Theft picture is part of the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983). In the picture, a woman is seen drying some dishes while looking away absent-mindedly, not noticing that the sink is overflowing with water. Behind her back, a boy and a girl are stealing cookies from a cookie jar placed in a kitchen cupboard. The boy is standing on a stool and is about to fall down. The Cookie Theft picture is used for eliciting narrative speech, primarily when diagnosing speakers with different types of language and communication disorders. Over the years, the Cookie Theft picture has been used to elicit speech in many different languages (e.g. Japanese (Choi, 2009); Norwegian (Lind et al., 2009); Chinese (Lai et al., 2009); Hebrew (Kavé and Goral, 2016)), suggesting its potential for cross-linguistic comparison studies.

A number of studies have reported that performance on the Cookie Theft description task is affected in AD, and in particular marked by a reduction in the amount of information that is conveyed. Croisile et al. (1996) compared written and spoken Cookie Theft descriptions from 46 French participants; 22 with AD and 24 controls. They found that the written narratives were more diagnostically useful, but that in general the AD descriptions were always shorter and less informative than the control descriptions. They measured information content by scoring the narratives against a list of 23 expected information units. (Here, an *information unit* refers generally to a concept in the image, such as *woman, sink*, or *overflowing*.)

[1] Note that we use the term *topic model* in the generic sense of an unsupervised method for discovering topics in text data, rather than specifically methods based on latent Dirichlet allocation, for example.

Forbes et al. (2004) compared written descriptions from AD participants and controls on a simple picture (in half of cases, the Cookie Theft picture) and a complex picture. Participants with mild AD were distinguished from controls on the basis of reduced information content and the production of fewer pictorial themes, while participants with more severe AD also suffered difficulties in the production of writing (e.g. stroke and case errors). In subsequent work involving oral as well as written picture descriptions, Forbes-McKay and Venneri (2005) found that participants in the early stages of AD differed from controls on measures of information content, pictorial themes, word finding delays and the response given to word finding delays. For the Cookie Theft picture, they list 7 pictorial themes, each involving at minimum a subject and a verb (e.g. *boy stealing cookies, woman washing/drying dishes*).

Pekkala et al. (2013) compared written English Cookie Theft descriptions from 23 participants with AD and 24 healthy controls. They generated a list of "target words" by compiling the 22 most common words (and their morphological variants) from the control data, and an additional list of somewhat lower-frequency words from the control sample. They found that the AD group produced significantly fewer words from both lists. They also found that cognitive deficits were evident from the Cookie Theft analysis 7–9 years prior to death, while deficits on other language tasks (namely the Boston Naming Test and the letter verbal fluency task) were evident only 2–4 years prior to death.

Ahmed et al. (2013b) analyzed Cookie Theft narratives from 18 English-speaking participants in the early stages of AD and 18 matched controls. They extracted the same 23 information units as Croisile et al. (1996), as well as measures of idea density (i.e. the number of information units per word in the narrative) and efficiency (i.e. the number of information units per unit time). They found that the AD participants produced significantly fewer information units overall, and in particular fewer subjects and actions (in contrast to objects and places). Idea efficiency was also reduced.

Landfeldt and Söderbäck (2009) analyzed 141 Cookie Theft narratives written by Swedish-speaking persons with SCI, MCI, and dementia. In addition to syntactic variables, they calculated the number of propositions and the idea density (defined as number of propositions divided by total number of words), but did not include idea density in their statistical analysis. They did find a significant difference between the number of propositions produced by the participants with dementia, relative to both MCI and SCI participants.

However, the usefulness of the Cookie Theft picture in detecting dementia in the prodromal phase is unclear. Some studies have reported a reduction in information content in preclinical dementia. Cuetos et al. (2007) used the Cookie Theft task to elicit speech from 19 Spanish-speaking carriers of the E280A mutation (which inevitably leads to AD, so these participants were assumed to be in an asympotomatic preclinical stage) and 21 noncarrier family members. There was no significant difference on the number of sentences produced or mean sentence length, but the carriers of the mutation did produce significantly fewer "semantic units" and "objective situations", both measures of information content. Ahmed et al. (2013a) reported deficits in various aspects of connected speech in 15 English MCI participants who later went on to develop AD, although the deficits were heterogeneous, representing impairments ranging from speech production and fluency to syntactic complexity and semantic content. However, the most common profile was characterized by an increase in the production of pronouns, and a decrease in total information units and idea efficiency. This pattern of impairment continued to worsen as the disease progressed and a dementia diagnosis was made.

In contrast, Bschor et al. (2001) found that German participants with AD, MCI, and no cognitive impairment all produced the same number of words in their Cookie Theft descriptions, but that those with AD described significantly fewer people, objects, and places than healthy controls. However, they found that MCI participants performed similarly to controls in terms of the number of information units produced. Similarly, Tyche (2001) investigated subtle language impairments in Swedish-speaking persons with MCI, and found that the persons with MCI did not differ significantly from healthy controls on language tests or on semantic aspects of the oral Cookie Theft narratives, although some qualitative differences were seen with regards to the structure of the narrative, in that the persons with MCI for example tended to be more repetitive.

2.2. Automated analysis of information content

Given the time-intensive nature of manually annotating picture description narratives for information content, there has been some effort to automatically extract relevant features, using text analysis and natural language processing.

One approach to scoring information content was developed by Prud'hommeaux and Roark (2015), and involves selecting a picture description from the control group to act as the "source narrative", and then using a graph-based alignment algorithm to determine how well each other narrative in the corpus recalls the story elements from the source narrative. Using a subset of the DementiaBank[2] corpus (130 AD samples and 130 control samples), they achieve a best accuracy of 83%. However, this method may be better suited to situations where a single, gold-standard source narrative exists, such as in a story-recall task.

Pakhomov et al. (2010) analyzed Cookie Theft narratives from 38 English participants with frontotemporal lobar degeneration (FTLD). In addition to other speech and language variables, they computed a "Correct Information Unit count" for each narrative based on a manually compiled list of unigrams, bigrams, trigrams, and 4-grams. The complete list (available in the Appendix of their paper) contains 135 items, although some are simply morphological variants (*asking for cookie, asking for a cookie, asking for cookies, ask for cookie*, and so on). However, the number of correct information units was not found to differ significantly between subtypes of FTLD.

Fraser et al. (2016) employed a similar approach to extract information units from Cookie Theft narratives from participants with AD. For each of the 23 information units described in Croisile et al. (1996), they used WordNet to semi-automatically generate a set of possible synonyms. They then extracted 23 binary-valued information units by searching for those words in the texts. Additionally, they computed integer-valued frequency counts for each of the relevant words contributing to the information units, allowing them to capture potentially relevant lexical variation in how the information units are described. In combination with other linguistic and acoustic features, they reported a best accuracy of 81% on the task of distinguishing between 240 AD narratives and 233 control narratives from the DementiaBank corpus.

There are many limitations to simply searching for a pre-computed list of keywords or *n*-grams. First, it requires the set of expected information units to be defined — apparently not a trivial task, given the number of different possibilities described in the literature above. In particular, determining the level of specificity can be difficult: perhaps the woman counts as a single information unit, but what about her shoes, her dress, her apron? Are these separate pieces of information, or do they all refer essentially to the woman?

Second, the information units must be operationalized in some way, generally by compiling a list of likely synonyms (the *boy* could also be referred to as the *son, child*, etc.). In this process, there is always the possibility that a reasonable word choice will be omitted. Additionally, it is not always obvious how to differentiate the information units without context (e.g. the *dish* and *plate* information units from Croisile et al. (1996), which appear to be largely differentiated only by their location − on the counter versus in the woman's hand). Furthermore, it can be difficult to generate keywords that capture more complex information units, such as *the woman's indifference towards the children*.

Finally, it is obvious that any list of keywords will be both picture- and language-dependent. Having the ability to present different picture stimuli is critical for longitudinal monitoring, where familiarity with the image may induce a so-called "practice effect" (Forbes-McKay and Venneri, 2005; Goldberg et al., 2015). Being able to assess people in their dominant language is also essential to get an accurate evaluation of their language abilities, and the possibility to use different images makes it feasible to evaluate all languages in multilingual individuals, since it is not certain that all languages of an individual are affected equally in dementia (Stilwell et al., 2016).

Recent work has attempted to avoid some of these issues by using unsupervised learning techniques to automatically generate information units directly from the data. Yancheva and Rudzicz (2016) used *k*-means clustering to generate topic models from the AD and control narratives in DementiaBank. In this case, a "topic" refers loosely to the same concept as an "information unit". With $k = 10$, they were able to recall 97% of the human-annotated information units (with some of the human-annotated units being clustered into a single topic). Using a small set of features extracted from these topic models, they were able to classify AD versus control narratives with an F-score of 0.74, and by combining those features with lexical and syntactic features, the F-score improved to 0.80.

Sirts et al. (2017) applied a similar methodology and reproduced the results of Yancheva and Rudzicz (2016) on DementiaBank (with an F-score of 0.73 using a slightly different set of cluster features), and also applied it to a dataset of open-ended spontaneous speech data from participants with and without AD, where they report an F-score of 0.85.

[2] http://dementia.talkbank.org/.

Here, we propose to combine the idea of fully automated generation of information units with recent work on multilingual word embeddings to create *multilingual information units*. One major challenge in the area of clinical language analysis has been that data is very scarce, and predominantly English. The multilingual approach allows us to learn better information units by augmenting the topic model training data with datasets outside of the target language, and therefore leverage the few publicly available datasets (such as DementiaBank), even in languages where such resources are not available.

3. Methods

We first describe the different data sets and participant groups involved in the study, then provide details of the clustering, feature extraction, and classification procedures.

3.1. Participants

We make use of three datasets, all based on the Cookie Theft picture: the Gothenburg dataset, the Karolinska dataset, and the DementiaBank dataset. The Gothenburg dataset was collected within the present project; the others are existing datasets from external sources. The properties of the datasets are summarized in Table 1, with the group labels "MCI" indicating participants diagnosed with mild cognitive impairment, and "HC" indicating healthy controls. The Mini-Mental State Examination (MMSE) is a general test of cognitive status with a maximum score of 30 (Folstein et al., 1975), and a score above 24 is considered normal (Grut et al., 1993).

3.1.1. Gothenburg

The Swedish participants recorded in Gothenburg are recruited from the ongoing Gothenburg MCI Study (Wallin et al., 2016). The Gothenburg MCI Study is a longitudinal in-depth phenotyping study of patients with different forms and degrees of cognitive impairment (i.e., from very mild to manifest dementia, but also including cognitively normal controls) using neuropsychological, neuroimaging, and neurochemical tools. The study is clinically based and aims at identifying neurodegenerative, vascular and stress related disorders prior to the development of dementia. All participants in the study undergo baseline investigations, such as neurological examination, psychiatric evaluation, cognitive screening (e.g., memory and visuospatial disturbance, poverty of language and apraxia), MRI imaging of the brain and cerebrospinal fluid collection. At biannual follow-ups, most of these investigations are repeated. The overall Gothenburg MCI Study is approved by the local ethical committee review board (reference number: L091-99, 1999; T479-11, 2011); while the currently described study is approved by the local ethical committee (decision 206-16, 2016).

A total of 31 MCI patients and 36 healthy controls were included in the present study, according to detailed inclusion and exclusion criteria (Kokkinakis et al., 2017). The participants in the current study all provided written informed consent. They were audio recorded while describing the Cookie Theft picture and performing some additional linguistic tasks not considered here. Participants were instructed to describe what they could see and what was happening in the picture. They were also told that they could talk for as long as they wanted and that they would not

Table 1
Demographic data for the different data sets included in the MCI analysis. *MMSE scores missing for 3 participants.

	In-domain				Out-domain	
Dataset	Gothenburg		DementiaBank		Karolinska	DementiaBank
Group label	MCI	HC	MCI	HC	HC	HC
N	31	36	19	19	96	78
Age (years)	70.1 (5.6)	67.9 (7.2)	66.7 (8.5)	66.4 (9.2)	57.2 (19.9)	63.9 (7.8)
Educ. (years)	14.1 (3.6)	13.1 (3.4)	14.9 (3.1)	14.2 (2.3)	13.0 (4.0)	13.9 (2.5)
Sex (M / F)	15 / 16	13 / 23	9 / 10	9 / 10	44 / 52	30 / 48
MMSE (/30)	28.2 (1.4)	29.6 (0.6)	27.4 (1.8)	29.1 (1.2)	–	29.1 (1.1)*
Task type	Spoken	Spoken	Spoken	Spoken	Written	Spoken
Language	Swedish	Swedish	English	English	Swedish	English

Table 2
Demographic data for the DementiaBank AD analysis. *MMSE scores missing for 2 participants. †MMSE scores missing for 3 participants.

Group	AD	HC
N	166	97
Age (years)	71.9 (8.3)	64.4 (8.1)
Education (years)	12.0 (2.7)	14.0 (2.4)
Sex (M / F)	55 / 111	39 / 58
MMSE (/30)	18.8 (5.1)*	29.1 (1.1) †
Task type	spoken	spoken
Language	English	English

be interrupted. The recorded narratives were subsequently manually transcribed by experienced transcribers according to guidelines provided by the authors.

3.1.2. Karolinska

The Karolinska corpus was collected by Cromnow and Landberg (2009) and contains only samples from healthy, Swedish controls. The 96 participants were divided into two groups depending on age (20–64 and 65–88 years old). The majority of the older participants in the dataset were members of Swedish retirement associations such as the "Pensioners' National Organization" (PRO) and "Active Seniors", whereas the younger participants were recruited through convenience sampling. The main criteria for participation in the Karolinska study were that subjects had Swedish as a first language, absence of clinical manifestations of linguistic impairment (such as dyslexia or aphasia) and absence of neurological disease.

The participants were instructed to produce a written description of what was happening in the Cookie Theft picture, while having the picture in front of them. They were given a time limit of 5 minutes, and wrote with pen on paper. The texts were manually transcribed into digital text files for the current study, using a set of guidelines to ensure consistency.

Because of the differences between this corpus and the Gothenburg corpus (most significantly, the lack of MCI participants, the wider range of ages, and the written modality), we consider this data to be "out-of-domain" with respect to the classification task, although the topic of the narratives is the same.

3.1.3. DementiaBank

Our English Cookie Theft data comes from DementiaBank, which is part of the TalkBank project (MacWhinney, 2007). These data were collected at the University of Pittsburgh as part of the Alzheimer Research Program. Detailed information about the original study is available from Becker et al. (1994). All participants received an extensive neurological, neuropsychological, psychiatric, and physical assessment.

Although the corpus primarily includes participants with AD (and healthy controls), we identified 19 participants who had been diagnosed with MCI. We then selected the 19 control participants who were a close match in terms of age, education, and sex. These comprise our English "in-domain" data. We then consider the remainder of the available control data (78 participants) to be our additional normative, or "out-of-domain" English data. Although many participants have contributed multiple samples to the DementiaBank database, we consider only the first available sample from each participant, so as to not bias the topic models or classifiers toward participants with multiple samples.

Additionally, to compare with the results previously reported by Yancheva and Rudzicz (2016) and Sirts et al. (2017), we consider the full DementiaBank corpus[3] comprising the 97 control participants and 166 participants with possible or probable AD. Participant demographics for this dataset are summarized in Table 2.

The DementiaBank data has been transcribed using the CHAT transcription protocol (MacWhinney, 2000).

[3] Version downloaded on November 22, 2013.

Table 3
Seed words used to initialize clusters.

English	boy, girl, woman, cookie, stool, sink, overflow, fall, window, curtain, plate, cloth, jar, water, cupboard, dish, kitchen, garden, take, wash, reach, attention, see
Swedish	pojke, flicka, kvinna, kaka, pall, diskho, rinna, ramla, fönster, gardin, tallrik, handduk, burk, vatten, skåp, fat, kök, trädgård, ta, diska, sträcka, märka, se

3.1.4. Group comparisons

Considering only the in-domain datasets, a multi-way ANOVA reveals no significant difference between the groups in terms of age or level of education. There is a significant main effect of MMSE ($p < 0.001$). Post-hoc tests reveal a significant difference between the Swedish MCI and HC groups ($p < 0.001$) and the English MCI and HC groups ($p < 0.001$). In each case, the HC group has the higher MMSE. However, we note that in general the average MMSE scores in the MCI groups are within what is considered the "normal" range (Grut et al., 1993), illustrating the very subtle impairment present at this stage. There is no significant difference in MMSE between the English and Swedish MCI groups, nor between the English and Swedish controls.

3.2. Clustering

Given the raw transcripts, we first pre-process the texts to remove filled pauses (e.g. *um* or *uh*), phonological fragments (e.g. *he's re- reaching*), and other non-lexical items (e.g. laughter). We part-of-speech (POS) tag and lemmatize each word, using the Stanford POS-tagger and NLTK WordNet Lemmatizer for English (Toutanova et al., 2003; Bird et al., 2009), and Sparv for Swedish (Borin et al., 2016). We then extract all nouns and verbs for the cluster analysis, following the assumption of the previous works that these word classes carry the most semantic information. Each extracted word is represented as a 300-dimensional vector using the pre-trained FastText word embeddings (Bojanowski et al., 2017). The FastText embeddings are based on the skip-gram model (Mikolov et al., 2013), but rather than learning word-level representations directly, character *n*-gram level representations are learned first, and then combined to generate word representations. The benefit of this approach is that it takes into account word morphology. This allows the sharing of subword information across related words, and means that representations can be obtained for words that did not occur in the training set. At the time of writing, pre-trained word vectors are available in 294 languages[4].

We chose this representation (in contrast to the GloVe embeddings used by Yancheva and Rudzicz (2016)) primarily due to the availability of transformation matrices for the alignment of FastText embedding spaces in any of 78 different languages, including English and Swedish[5]. Details of the transformation procedure are given by Smith et al. (2017). A dictionary of translations for 5000 frequently occurring words in the two languages is automatically generated, and the embedding spaces then rotated such that the mean cosine distance between translation pairs is minimized. This is accomplished by using a singular value decomposition to learn a linear transformation between the two embedding spaces. The assumption is that if the rotation aligns the 5000 known translation pairs, then the other vectors in the rotated space should lie close to their translations as well, based on the underlying structure of the embedding spaces. This transformation allows our English and Swedish word vectors to be represented in the same space, while preserving the relationships between words in the underlying monolingual models.

One challenge in using pre-trained word vectors in Swedish is the presence of compound words in the narratives, which may in some cases be generated on-the-fly and thus not have a representation in the vector space. To manage this issue, for any word which does not have a vector representation, we again use Sparv to analyze its compound structure, and then check if any of its constituents have vector representations. In this way, for example, a person who describes the kitchen as a *femtiotalskök* (*1950s kitchen*) is still credited with the word *kök* (*kitchen*), which is represented in the set of pre-trained vectors.

[4] https://github.com/facebookresearch/fastText.
[5] https://github.com/Babylonpartners/fastText_multilingual.

Table 4
Data set combinations for training the cluster models for MCI classification.

	In-domain				Out-domain	
	Gothenburg		DementiaBank		Karolinska	DementiaBank
Swedish (in-domain)	✓	✓				
Swedish (all)	✓	✓			✓	
English (in-domain)			✓	✓		
English (all)			✓	✓		✓
Multilingual (in-domain)	✓	✓	✓	✓		
Multilingual (all)	✓	✓	✓	✓	✓	✓

For clustering, we use the Matlab implementation of k-means, with cosine distance as the distance metric. This choice of distance metric represents another departure from the methodology described in Yancheva and Rudzicz (2016) and reproduced in Sirts et al. (2017), but we consider this necessary since the procedure to generate the multilingual embedding space minimizes the cosine distance between translation pairs, rather than the Euclidean distance. Cosine distance also has some practical advantages over the d_{scaled} measure used in that study in that it is nonnegative, and well-defined for clusters containing a single word type. We consider three possible values for k: 10 (as suggested by the two previous studies), 23 (as Croisile et al. (1996) suggest there are 23 natural information units), and k_{sil}, where $k_{sil} \in \{2, 3, ..., 30\}$ and is chosen fully automatically using the silhouette method (Kaufman and Rousseeuw, 2009). We consider two initialization strategies: the first uses standard k-means++ initialization, which we then let run until convergence or for a maximum of 1000 iterations, and restart 100 times to avoid local minima. In the second strategy, we explore the effect of adding some expert knowledge by seeding the initial clusters with words representing the information units from Croisile et al. (1996), and again run for a maximum of 1000 iterations. The seed words were selected by the authors, in each case a native speaker of the given language, and are shown in Table 3. Given the variability that can occur in the output of cluster modeling, we generate 10 different cluster models using 10 different random seeds for each combination of parameters, and then incorporate model selection as part of the classifier training process (Section 3.4).

The cluster model can be trained on any combination of the available data; to examine more closely the effects of language and domain, we consider the six cluster model training configurations given in Table 4. When training on both languages together, the training set is balanced such that it contains the same number of word tokens from each language.

Additionally, to compare the multilingual approach to the previous work that has been done in English, we apply our methodology to the DementiaBank classification task, using cluster models trained on the full DementiaBank dataset (English), as well models trained after adding all available Swedish data (English + Swedish).

3.3. Feature extraction

Once we have the cluster model, we can extract features from the classification data set (either *English (in-domain)* or *Swedish (in-domain)*) using the cluster information. For each narrative, we first apply the same pre-processing steps as described in Section 3.2, represent the nouns and verbs as vectors, assign each vector to a cluster in the cluster model, and then extract the features listed in Table 5. The last two features listed in the table are baseline features, and are independent of the cluster model. We use these features to train our baseline classifiers only. Note that in the case where *no* words are discarded in the filtering step (i.e. all nouns and verbs are considered to be relevant to one of the topics), then N+V density is equivalent to information density, and similarly for N+V efficiency and information efficiency.

3.4. Classification

The main classification task is to distinguish between narratives from MCI speakers and controls. We use a linear SVM classifier (Pedregosa et al., 2011), and develop separate models for the English and Swedish MCI classification tasks. In both cases, we use a leave-one-out cross-validation framework, in which one narrative is set aside as the test

Table 5
Features extracted from the Cookie Theft narratives, based on the cluster model.

Cluster features	**C_i** For each cluster i, find the average cosine distance between the centroid and all words assigned to that cluster. This is equivalent to C_i in Yancheva and Rudzicz (2016), but using cosine distance rather than their d_{scaled} quantity.
	N_i For each cluster i, discard any word that lies more than 3 standard deviations away from the mean distance to the centroid in the training set, as proposed by Yancheva and Rudzicz (2016). (This filtering step is necessary because otherwise every word, no matter how irrelevant, will be assigned to some cluster.) Then count how many words are assigned to the cluster. The raw frequency count is N_i (indicating how many times a given topic i is mentioned).
	P_i For each cluster i, the frequency count N_i (above), divided by the total number of words in the narrative (indicating what proportion of the words produced belonged to this topic).
Summary features	**Idea density** The number of clusters that are mentioned, divided by the total number of words in the narrative.
	Idea efficiency The number of clusters that are mentioned, divided by the total time of the narrative in seconds.
	Information density The number of words which are assigned to clusters, divided by the total number of words in the narrative.
	Information efficiency The number of words which are assigned to clusters, divided by the total time of the narrative in seconds.
Baseline features	**N+V density** The total number of nouns and verbs, divided by the total number of words in the narrative.
	N+V efficiency The total number of nouns and verbs, divided by the total time of the narrative in seconds.

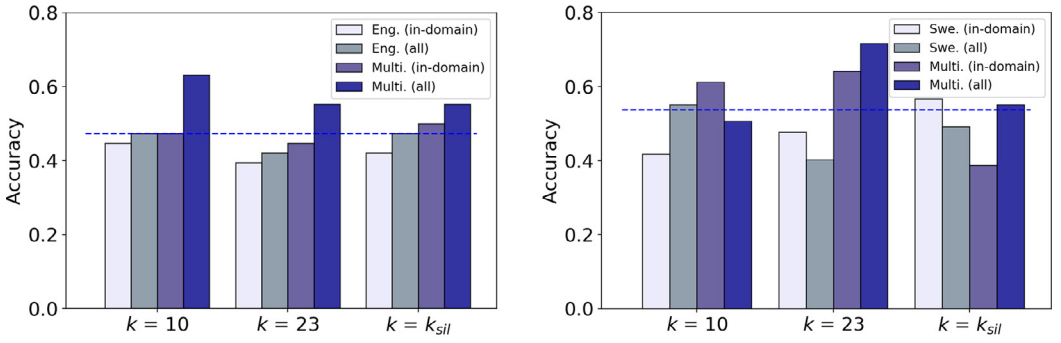
set, and model selection and training is performed on the remaining data. This allows us to maximize the size of our training set, given the relatively small number of samples available. On the training set, we run an inner loop of cross-validation to select the complexity parameter for the SVM classifier, and to choose the cluster model from the 10 generated for each configuration. We report the average accuracy across folds, as well as *sensitivity* (the true positive rate) and *specificity* (the true negative rate). These metrics are particularly relevant in a healthcare context, where it is desirable to have a test that is both highly sensitive (here, detects most cases of MCI) and specific (here, does not wrongly flag healthy individuals as having MCI).

Additionally, we consider the task of distinguishing between the AD and control narratives from DementiaBank, in order to compare our results against those previously reported by Yancheva and Rudzicz (2016) and Sirts et al. (2017). The methodology is the same as for the MCI classification task.

4. Results

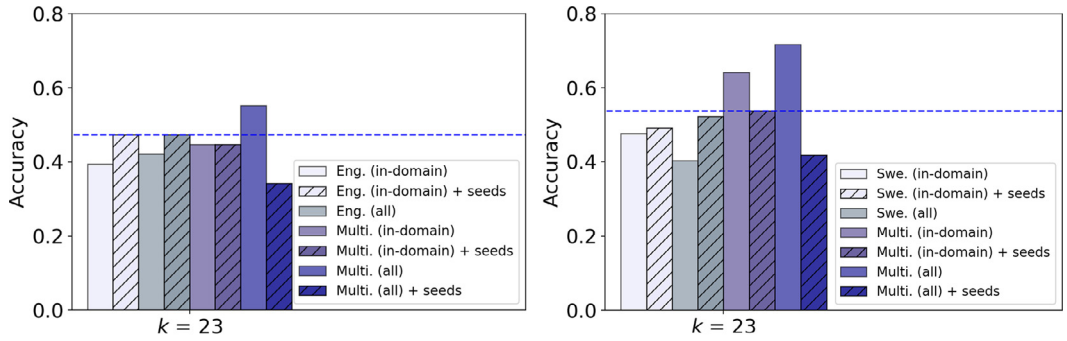
4.1. Monolingual and multilingual data augmentation

The classification accuracies for the two languages under different experimental conditions can be seen in Fig. 1. In the English classification task (Fig. 1a), the highest accuracy is always achieved using the topic model trained on the *multilingual (all)* data. This pattern does not hold in the Swedish classification task (Fig. 1b), where the topic model trained on *multilingual (in-domain)* performs better in the case of $k = 10$, and the *Swedish (in-domain)* model



(a) English MCI vs HC. (b) Swedish MCI vs HC.

Fig. 1. Classification accuracies using features extracted from cluster models trained on different data sets (English, Swedish, and multilingual). The dashed line indicates the accuracy achieved using the N+V baseline features described in Section 3.3.



(a) English MCI vs HC. (b) Swedish MCI vs HC.

Fig. 2. Classification accuracies when the clusters are initialized with *k*-means++ versus with centroids based on the information units given by Croisile et al. (1996).

for the automatically chosen k; however, the overall best accuracy on this task is again achieved using the *multilingual (all)* data, with a value of 0.72. This is higher than the best accuracy of 0.63 in the English case.

An ANOVA reveals a significant effect of topic model type (multilingual versus monolingual) on accuracy ($F(1, 18) = 7.893$, $p = 0.01$), with the multilingual models leading to higher accuracies. There are no significant effects of domain (in-domain versus all), test language (Swedish versus English), or number of clusters k.

Table 6
Summary of all classification accuracies, as well as sensitivity and specificity scores for each configuration.

				Test set: English			Test set: Swedish		
Training set	k	# features	seeds?	Acc	Sens.	Spec.	Acc	Sens.	Spec.
English	10	34	False	0.45	0.47	0.42	–	–	–
(in-domain)	23	73	False	0.39	0.37	0.42	–	–	–
	23	73	True	0.47	0.42	0.53	–	–	–
	2–30	10–94	False	0.42	0.32	0.53	–	–	–
Swedish	10	34	False	–	–	–	0.42	0.48	0.36
(in-domain)	23	73	False	–	–	–	0.48	0.48	0.47
	23	73	True	–	–	–	0.49	0.52	0.47
	2–30	10–94	False	–	–	–	0.57	0.52	0.61
Swedish+English	10	34	False	0.47	0.37	0.58	0.61	0.61	0.61
(in-domain)	23	73	False	0.45	0.37	0.53	0.64	0.58	**0.69**
	23	73	True	0.45	0.42	0.47	0.54	0.55	0.53
	2–30	10–94	False	0.50	**0.53**	0.47	0.39	0.35	0.42
English	10	34	False	0.47	0.37	0.58	–	–	–
(all)	23	73	False	0.42	0.37	0.47	–	–	–
	23	73	True	0.47	0.37	0.58	–	–	–
	2–30	10–94	false	0.47	0.47	0.47	–	–	–
Swedish	10	34	False	–	–	–	0.55	0.55	0.56
(all)	23	73	False	–	–	–	0.40	0.39	0.42
	23	73	True	–	–	–	0.52	0.58	0.47
	2–30	10–94	false	–	–	–	0.49	0.45	0.53
Swedish+English	10	34	false	**0.63**	**0.53**	**0.74**	0.51	0.48	0.53
(all)	23	73	false	0.55	**0.53**	0.58	**0.72**	**0.77**	0.67
	23	73	True	0.34	0.32	0.37	0.42	0.48	0.36
	2–30	10–94	False	0.55	**0.53**	0.58	0.55	0.58	0.53
Baseline	–	2	–	0.47	**0.53**	0.53	0.54	0.55	0.53

4.2. Effect of seeding the topics

We now consider the effect of adding some expert knowledge to the cluster model, by restricting $k = 23$ and initializing the k-means algorithm with words representing Croisile's 23 information units. The classification results for $k = 23$ with and without the seed words are given in Fig. 2.

We observe that while in some cases, initializing the clusters with seed words leads to small increases in performance, using the seed word initialization never leads to classification accuracies that exceed the baseline. This suggests that the fully automated cluster models are better able to capture patterns in the data which distinguish the two groups. We will present some examples of this in the discussion, Section 5.2.

4.3. Summary of MCI results

A complete overview of the results presented in the previous sections is given in Table 6, including the sensitivity and specificity for each configuration. Looking first at the English MCI classification task, we see that even the best performing model has low sensitivity (not better than the baseline). This is undesirable from the perspective of a dementia screening tool, as it means that individuals with MCI may be missed. The specificity in the best case is 0.74.

However, in the Swedish MCI classification task, we observe the opposite pattern: in the best result, the sensitivity (0.77) is higher than the specificity (0.67). An imbalance of this type is typically less problematic, as any individuals flagged as potentially having MCI would be referred to a specialist, who would conduct a more detailed examination and rule out any false positives.

Considering only the experiments with k-means++ initialization (corresponding to the results reported in Section 4.1), we find a significant effect of topic model type (multilingual versus monolingual) on both sensitivity ($F(1, 18) = 4.845$, $p = 0.04$) and specificity ($F(1, 18) = 6.418$, $p = 0.02$), with higher values achieved using the multilingual models. There is also an effect of classification task on sensitivity ($F(1, 18) = 5.043$, $p = 0.04$), with higher sensitivities obtained in the Swedish MCI classification.

4.4. DementiaBank classification

Finally, although the focus of this paper is on MCI rather than AD, we use the same methodology to classify DementiaBank narratives as belonging to the AD or control groups, and compare against the results reported in the literature. We compare directly against the results of Yancheva and Rudzicz (2016) and Sirts et al. (2017), since they use a similar methodology of clustering to identify information units. However, we note that other previous work has taken alternative approaches to DementiaBank classification, including Prud'hommeaux and Roark (2015) (83% accuracy) and Fraser et al. (2016) (81% accuracy), described in Section 2.2, as well as Orimaye et al. (2014) (F-score of 0.74, using lexical and syntactic features), and Orimaye et al. (2017) (AUC of 0.94, using 1000 n-gram features).

The comparison is not exact, as the previous works used different classifiers and train-test configurations. However, Table 7 shows that by using our methodology and increasing the number of clusters from 10 to 23, we can

Table 7
F-scores reported in the previous literature and obtained using this method for the classification task of distinguishing AD and control narratives from DementiaBank.

	Cluster features	Cluster features + Summary features
Yancheva and Rudzicz (2016)	0.68	0.74
Sirts et al. (2017)	0.66	0.75
English ($k = 10$)	0.81	0.83
English + Swedish ($k = 10$)	0.80	0.80
English ($k = 23$)	0.80	0.79
English + Swedish ($k = 23$)	**0.85**	0.81
English + seeds ($k = 23$)	0.79	0.78
English + Swedish + seeds ($k = 23$)	0.77	0.78

improve the performance on this task to a best F-score of 0.85. We also consider the effect of including Swedish data in the topic modeling step, even though we do not have any Swedish AD data data. Using the multilingual data decreases performance in the $k = 10$ configuration but increases performance in the $k = 23$ case, leading to the best results of $F = 0.85$ using the cluster features alone. For the sake of comparison, this corresponds to an accuracy of 82%, a sensitivity of 0.82, a specificity of 0.81, and an AUC of 0.89. We also consider the effect of initializing with seed words, as before, but find that it confers no benefit over the k-means++ initialization. Finally, we note that the cluster features (first column) appear to be generally more discriminative here than in previous work, suggesting that either the FastText word embeddings or our modified feature definitions offer some benefit over the previously reported methodology.

5. Discussion

In the following section we discuss in more detail the implications of the results, including possible factors contributing to the strength of the multilingual approach, examples of the kinds of topics that are learned in the different configurations, and a comparison of the summary measures of density and efficiency across the MCI and control groups.

5.1. Why is the multilingual approach effective?

The classification results show that when training the cluster model, it is often more effective to add data from a different language, rather than more data from the same language. Why could this be? One possibility is that adding more data from the same language does not actually add any new information. That is, because the picture is relatively simple, after some number of picture descriptions are seen, all the relevant words have been mentioned and adding additional training data does not improve the topic modeling.

Observing many repeated instances of the same word does not tend to lead to richer topic models, but simply concentrates the centroids around those highly frequent words. In the extreme case, where a cluster contains only multiple instances of a single word type, the standard deviation of the distance to the centroid will be zero, and it will be impossible for any other word type to be assigned to that topic during evaluation; that is, the "topic modeling" approach will reduce to simple keyword-spotting.

However, artificially increasing the type-token ratio does not necessarily lead to better classification accuracy either. To test this scenario in the extreme, we ran a set of experiments in which we filtered the cluster model training data to exclude all repetitions of a word; that is, we trained on the set of *tokens* rather than the set of *types*. This maximized the type-token ratio of the training set to be 1.0. However, the classification accuracies were poorer in all cases. In this scenario, all words in the training data appear with uniform probability, while in reality we want the cluster centroids to be closer to *stool* than to *chair*, closer to *cookie* than *cake*, and so on. Clearly, the frequency information that comes from including the natural distribution of word tokens is also important.

Thus, one possible explanation for the effectiveness of the multilingual approach here is that it helps to balance the trade-off between too-tight clusters (occurring due to many repeated instances of highly frequent words in the monolingual dataset) and too-broad clusters (occurring when each word type appears with uniform density). Another way of thinking about this is that the multilingual data adds more "synonyms" to the dataset, enriching the vocabulary in the relevant areas of the semantic space (assuming, of course, a good alignment between the English and Swedish spaces). This helps to better define where the clusters should be located in the space. A concrete example of how this can lead to better topics will be seen in the following section.

5.2. What topics are being learned?

To better understand how the different topic models differ in practice, we consider some examples of the topics that are learned. For each configuration (combination of training set, k, and classification language), we consider examples from the model that leads to the overall best accuracy on the given classification task. For the sake of space, we do not attempt an exhaustive analysis, but simply offer some illustrative examples.

Table 8 shows two examples of clusters learned using *English (all)* data with $k = 23$, initializing with either k-means++ or the seed method. For each cluster, we generally list the top five words in the cluster, ranked by their

Table 8
Comparison of the topic models resulting from the supervised and automatic initialization methods, for the dataset *English (all)*, with $k = 23$. The *k*-means++ clusters have been manually aligned to the seeded clusters to aid in comparison.

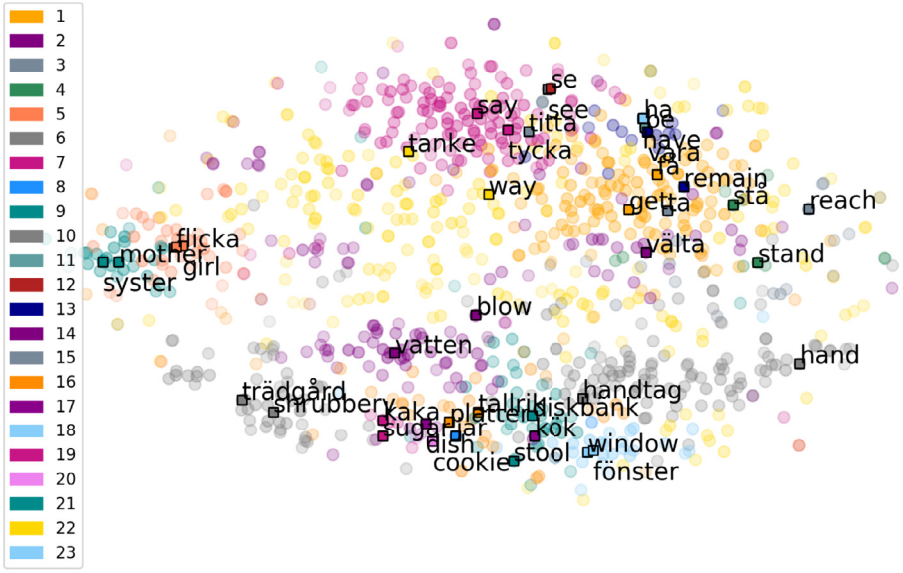
#	Seed	Initialization: seed words	Initialization: k-means++
1	boy	boy kid youngster lad johnny ...	boy girl kid man youngster ...
2	girl	girl woman man mama mommy ...	
3	woman	mother daughter sister child son ...	mother daughter sister child son ...
4	cookie	cookie jelly	cookie
5	stool	stool stepstool	stool stepstool
6	water	water dry rain basin wind ...	water sink dry wash faucet drip spilling overflow flow ...
7	sink	sink	
8	overflow	overflow flow fill torrent cascade ...	
9	wash	wash spilling mouth towel faucet ...	
10	fall	fall spring collapse winter summer ...	fall spring collapse winter summer
11	window	window door windowsill glass side ...	window door curtain windowsill roof ...
12	curtain	curtain wall lip valance ladder ...	shrubbery drape skirt ruffle cloth ...
13	plate	plate cup bowl finish pan ...	cup time finish end back ...
14	cloth	hand finger wrist foot toe ...	hand finger wrist toe nose...
15	jar	jar lid pot	jar lid pot
16	dish	dish pudding platter egg fry	dish platter pudding plate egg ...
17	kitchen	floor room kitchen roof house ...	floor kitchen room cupboard house closet countertop driveway garage ...
18	cupboard	cupboard cabinet closet countertop dishwasher ...	
19	garden	shrubbery tree grass shrub garden ...	
20	take	want know get say think ...	get go want take ask ...
21	reach	reach climb grow extend	reach climb grow extend path ...
22	attention	action counter perspective time step ...	leave cause hear attention disturb ...
23	see	be have remain include	see look glance appear detail
24	–		say think know something suppose thing anything ...
25	–		mhm board chair shh sirt ...
26	–		run stand running walk standing ...
27	–		action counter actio motion movement
28	–		do em xxx uff
29	–		be have

closeness to the cluster centroid (in some cases, the cluster contains only five or fewer words, in which case there is no ellipsis; in other cases we list more than five words to aid in the example). We have manually aligned similar clusters for the purpose of comparison.

In many cases, the two methods of initialization result in very similar clusters (e.g. clusters 14, 15, 16). In other cases, the *k*-means++ initialization merges topics that are distinct in the seeded initialization. For example, in the seeded case, clusters 1, 2, and 3 are intended to refer to the boy, the girl, and the woman respectively (although it is clear from the top-ranked words that there is much semantic overlap between the clusters 2 and 3). In the automatically initialized case, we end up with only two clusters relating to these topics: cluster 1 generally contains words that objectively refer to the people in the image (*boy, girl,* etc.), while cluster 3 contains words that describe an inferred relationship between the participants (*mother, daughter, son*, etc.). While these clusters do not accurately distinguish between the three actors in the scene, previous work has sometimes drawn a similar distinction between *interpretive* versus *literal* information units in the Cookie Theft picture (Hillis Trupe and Hillis, 1985), and the ability to infer or interpret non-literal pictorial elements may be impaired in dementia and its pre-clinical stages (Stevens, 1985; Cuetos et al., 2007).

We also observe that the objects *water* and *sink*, and the actions *wash* and *overflow*, have all been assigned to a single topic in the automated case, and similarly words relating to the *kitchen, cupboard*, and *garden* have been collapsed into a single topic.

In contrast, there are some topics generated in the automated case that do not exist in the seeded topic models. One interesting example is clusters 28 and 29, which contains verbs that can be used as auxiliary verbs (*be, do*, and *have*). These clusters are likely not very informative on their own, as most narratives will include these highly frequent verbs. However, allocating these verbs to a separate cluster allows the other clusters to be more specific with respect to the verbs that qualify as a reference to a given information unit. Cluster 24 is another interesting example, where the words again do not appear to reference the actual information content of the image, but rather a speaker's



1	få försöka get komma kunna hålla låta gå göra sätta hjälpa vänta ...
2	vatten water varmvatten vattenpöl vattensamling torka dry diskvatten ...
3	see look titta glance
4	stå stand ligga falla fall run standing sit heta running hylla lay
5	flicka pojke girl boy mamma flickebarn tjej kille kid tioårsåldern ...
6	trädgård shrubbery garden träd hus buskage plantering gräsmatta ...
7	say think know tycka tänka want something anything thing presume ...
8	jar lid
9	mother syster daughter barn dotter mor syskon child fru sister son ...
10	hand handtag tygstycke handduk skjorta klnning wrist snöre finger ...
11	be
12	se bild pl
13	vara finnas bli kännas förefalla innehålla verka bestå böra remain ...
14	välta slå förstöra skjuta blow action blåsa stänga skratta wipe vrida ...
15	ta reach nå ramla kliva klättra växa climb slingra grow tippa walk ...
16	tallrik pall stol kopp gardin skopa balja platter barstol plate ...
17	dish kök kitchen pudding coffee egg fry
18	ha have
19	kaka pepparkaka smaka sugar nalla kanke puff
20	cookie jelly
21	diskbänk köksgolvet köksskåp stool diskho skåp cupboard diskmaskin ...
22	tanke gång sätt way ögonblicksbild början tillfälle säga svänga ...
23	fönster window golv dörr roof door floor spröjs tak vägg windowsill ...

Fig. 3. Two-dimensional representation of a cluster model using *Multilingual (all)* and $k = 23$. Color indicates cluster membership. For each cluster, the English and Swedish words closest to the centroid are annotated, and the top closest words listed below.

uncertainty about the task, often reflected in statements like *I don't know, I can't say anything else*, and vague terms like *thing* and *something*.

These differences notwithstanding, in almost half of the automatically initialized clusters, the word closest to the centroid actually belongs to the list of seed words. This suggests that if the human-generated list of information units is essentially a list of frequently-mentioned elements in the image, then given enough (normative) data, we can learn those topics from the frequency information available in the data. However, the classification results would seem to

Table 9
Example: the polysemous English word *fall*. In the multilingual case, it belongs to a cluster with a centroid that is closer to words relating to the action of falling, rather than the autumn season.

	Monolingual	Multilingual
Words in the cluster:	*fall* spring collapse winter summer	stå stand ligga falla *fall* run standing sit heta running hylla lay
Distance of centroid to related words:	**autumn 0.50** **harvest 0.70** tumble 0.61 drop 0.56	autumn 0.65 harvest 0.72 **tumble 0.58** **drop 0.51**

suggest that these frequently-mentioned topics are not the most diagnostically useful features, at least in the case of very mild cognitive impairment.

We turn now to the multilingual topic models. Here we consider the *multilingual (all)* dataset with $k = 23$ and k-means++ initialization, which led to the best classification result in Swedish, and the second-best in English. To visualize the clusters in two dimensions, we use t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008). In Fig. 3, cluster membership is indicated through different colors, with the intensity of the color scaled by the closeness of the word vector to the cluster centroid. For each cluster, we annotate the word closest to the centroid from each language (unless only one language is represented in that cluster). We can see that these words are often, but not always, direct translations of each other (e.g. *girl* and *flicka, window* and *fönster*). However, there are also examples where the multilingual approach failed. For example, *cookie* and *kaka* belong to separate clusters (20 and 19), as do *stool* and *pall* (21 and 16). If we compare the cosine distances of these five example pairs in the multilingual space, the first three are indeed closer than the latter two, perhaps due to polysemy (e.g. *kaka* could also be translated as *cake*, and in fact *kaka* is closer to *cake* than *cookie* in the vector space).

However, in some cases the multilingual approach actually appears to resolve sense ambiguity. In Table 8, cluster 10 shows that in the monolingual English case, both the seed and k-means++ initialization lead to the word *fall* being clustered with the words *spring, winter*, and *summer* (which are mentioned in reference to the exterior seen through the window, albeit infrequently). Clearly, this is not the sense of *fall* that is most relevant to the Cookie Theft picture. In the multilingual case, *fall* appears in a cluster with the Swedish word *falla*, which does not have the same ambiguity with the season of autumn (*höst*). Table 9 shows a comparison of the two clusters: in the monolingual case, the cluster centroid is closer to words like *autumn* and *harvest*, while in the multilingual case the centroid is closer to words like *tumble* and *drop*, suggesting that this cluster is more appropriate for the sense of *fall* that we expect to see most often in this corpus. This is an example of how the multilingual approach can lead to better topic models, even if an individual speaker will only use words from one language or the other.

Table 10
Average values for the four summary features, computed across every configuration. Boldface indicates a significant difference between the groups (within a language). $*p = 0.003$, $\dagger p = 0.0002$.

	Gothenburg (Swedish)		DementiaBank (English)	
	HC	MCI	HC	MCI
Idea density	0.09	0.09	0.12	0.14
Idea efficiency	0.19	0.18	0.29	0.28
Information density	0.36	0.35	0.39	0.38
Information efficiency	**0.77**	**0.72***	**0.93**	**0.81**†

5.3. Is there a change in information content in MCI?

Our assumption throughout has been that the production of information will be somehow disrupted or reduced in the MCI participants, but do our data support this assumption? In Table 10 we show the averages for the four summary features, computed over all combinations of k and training set (selecting at random one of the 10 models generated for each configuration). In both Swedish and English, we see no significant difference between the groups in idea density, idea efficiency, or information density. However, in both languages, there is a significant reduction in information efficiency (that is, the number of words referring to relevant information units, divided by the total time). Note that although the MCI participants do tend to speak somewhat slower than the control participants, there is no significant difference in speech rate (as measured in words per minute) between the MCI and HC groups in either Swedish or English. Thus, the difference really does appear to be one of efficiency: MCI participants are producing relevant, information-bearing words at a significantly slower rate than controls, despite not speaking at a significantly slower rate in general.

It is worth noting that information efficiency (and to some extent, all four measures) also tended to be higher in English than in Swedish, pointing to the difficulty of directly comparing such numerical results across languages without taking into account structural and morphological differences between the languages.

5.4. Limitations

As we observed, one limitation of the word embedding approach is that the word vectors are not disambiguated for word sense, or even part-of-speech. This problem is potentially compounded in the multilingual case, as different word senses may be better aligned to different word vectors in the multilingual space. We discussed the issue of *fall*; another example that we noticed in this corpus is the Swedish word *skola*, which occurred most often as the lemma of the helper verb *ska* expressing something that will happen (similar to *will* in English). However, *skola* also translates as the English noun *school*, and it seems that this was the dominant sense captured by the word vector, as it was often assigned to clusters referring to the two children. One solution to the problem of multiple senses could involve fuzzy clustering approaches, where words can be assigned to more than one cluster. Another solution could lie in embedding approaches that distinguish between word senses and parts-of-speech (Chen et al., 2014; Trask et al., 2015).

In keeping with the previous work, we also limited our analysis to nouns and verbs. However, there could be important information in the adverbs and adjectives as well. In particular, two of the information units suggested by Croisile et al. (1996) involve the mother appearing unconcerned, indifferent, or distracted — all adjectives. Future work should consider how to best incorporate this information in the models, as well as better handling of multiword expressions and particle verbs.

Of course, quantifying information content is just one aspect of a complete linguistic analysis. We have started our multilingual analysis here, since it seems reasonable that the information conveyed should be comparable across languages, on an abstract semantic level. Nonetheless, we did find differences in the numerical values of the scores for density and efficiency across languages. It is not at all obvious that the approach of supplementing training data with samples from other languages will be appropriate when examining other linguistic levels, such as syntax, acoustics, and so on.

Finally, our methodology does not allow us to definitively state whether the lower information efficiency observed in the MCI participants truly represents a *change* due to their declining cognitive status, since we do not have a pre-morbid baseline against which to compare. However, longitudinal data will be collected from the Swedish participants, and follow-up research will document whether information density and efficiency decline on an individual level over the course of the progression from MCI to dementia.

6. Conclusion

In many studies involving the analysis of clinical language samples, the size of the data is necessarily quite limited, as data collection is expensive, time-consuming, and limited by various factors (e.g. difficulty recruiting participants into a study, limited access to the population of interest, privacy issues, etc.). In this paper, we have considered

how we can use external data to boost the performance of automatic and machine learning methods when faced with the challenge of small data.

Although we have focused on the topic modeling and classification tasks, when we look at the big picture, there are actually three broad levels of data involved in the end-to-end analysis:

1. *Very large, completely out-of-domain data.* These data were used to train the part-of-speech taggers, lemmatizers, and word embedding models that we used as off-the-shelf tools. They include things like Wall Street Journal texts and Wikipedia articles.
2. *Small, slightly-out-of-domain unlabeled data.* These are the additional DementiaBank and Karolinska data we used to help train the unsupervised topic models. They *are* Cookie Theft descriptions, but they do not have the same distribution as the classification data, in terms of patient group membership or demographic variables.
3. *Small, in-domain, labeled data.* These are the clinical datasets for the supervised classification tasks (and which we also used to help train the topic models, without considering the labels).

Figuring out the best way to learn from *all* available data will be an important step forward in successfully applying machine learning in the clinical domain. Here, we find that using data from a different language to help train the topic model can lead to better classification results than simply augmenting the training data with additional normative data from the same language. This result held in both English and Swedish, although future work will involve additional languages. Another avenue of future research will be investigating the optimal proportion of each language; here, we only considered an equal mix of the two available languages, but it could be that biasing the topic model towards the target language, while still enriching the model with some information from another language (or languages), could be even more effective.

Regarding initialization strategies, our results do not support using expert knowledge to seed the clusters, at least for the downstream application of automated classification. In any case, the completely data-driven approach is preferable in terms of generalization to different languages and different stimulus images.

Finally, we report a significant reduction in information efficiency in the MCI group, in both English and Swedish. Ongoing work will assess whether this finding generalizes across more varied languages, and seek to identify other general linguistic markers of MCI. As dementia is a global problem, the comparison of results between and across languages will be necessary to build robust and accurate automated tools for screening and monitoring the early indicators of cognitive decline.

Acknowledgments

This work has received support from *Riksbankens Jubileumsfond – The Swedish Foundation for Humanities and Social Sciences*, through the grant agreement no: NHS 14–1761:1. The original acquisition of the DementiaBank data was supported by NIH grants AG005133 and AG003705 to the University of Pittsburgh, and maintenance of the data archive is supported by NIH-NIDCD grant R01-DC008524 to Carnegie Mellon University. We are also thankful to Ing-Mari Tallberg for providing us with the Karolinska data, and to the anonymous reviewers for their helpful feedback.

References

Ahmed, S., Haigh, A.-M.F., de Jager, C.A., Garrard, P., 2013. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain 136 (12), 3727–3737.

Ahmed, S., de Jager, C.A., Haigh, A.-M., Garrard, P., 2013. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. Neuropsychology 27 (1), 79–85.

Alberdi, A., Aztiria, A., Basarab, A., 2016. On the early diagnosis of Alzheimer's disease from multimodal signals: A survey. Artif. Intel. Med. 71, 1–29.

Alm, C.O., 2016. Language as sensor in human-centered computing: clinical contexts as use cases. Lang. Linguist. Comp. 10 (3), 105–119.

Asgari, M., Kaye, J., Dodge, H., 2017. Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimer's Demen. 3 (2), 219–228.

Becker, J.T., Boiler, F., Lopez, O.L., Saxton, J., McGonigle, K.L., 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Arch. Neurol. 51 (6), 585–594.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python. O'Reilly Media, Inc.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. 5, 135–146.
Borin, L., Forsberg, M., Hammarstedt, M., Rosen, D., Schäfer, R., Schumacher, A., 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. The Sixth Swedish Language Technology Conference (SLTC), Umeå University, 17–18 November.
Bschor, T., Kühl, K.-P., Reischies, F.M., 2001. Spontaneous speech of patients with dementia of the Alzheimer type and mild cognitive impairment. Int. Psychogeriat. 13 (3), 289–298.
Chen, X., Liu, Z., Sun, M., 2014. A unified model for word sense representation and disambiguation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1025–1035.
Choi, H., 2009. Performances in a picture description task in Japanese patients with Alzheimer's disease and with mild cognitive impairment. Kor. J. Commun. Dis. 14 (3), 326–337.
Clark, L.J., Gatz, M., Zheng, L., Chen, Y.-L., McCleary, C., Mack, W.J., 2009. Longitudinal verbal fluency in normal aging, preclinical, and prevalent Alzheimers disease. Am. J. Alzheimer's Dis. Demen. 24 (6), 461–468.
Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., Trillet, M., 1996. Comparative study of oral and written picture description in patients with Alzheimer's disease. Brain Lang. 53 (1), 1–19.
Cromnow, K., Landberg, T., 2009. Skriftliga beskrivningar av bilden Kakstölden. Insamling av referensvärden från friska försökspersoner.
Cuetos, F., Arango-Lasprilla, J.C., Uribe, C., Valencia, C., Lopera, F., 2007. Linguistic changes in verbal expression: A preclinical marker of Alzheimer's disease. J. Int. Neuropsychol. Soc. 13 (3), 433–439.
Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J. Psychiat. Res. 12 (3), 189–198.
Forbes, K.E., Shanks, M.F., Venneri, A., 2004. The evolution of dysgraphia in Alzheimers disease. Brain Res. Bull. 63 (1), 19–24.
Forbes-McKay, K.E., Venneri, A., 2005. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. Neurol. Sci. 26, 243–254.
Fraser, K.C., Meltzer, J.A., Rudzicz, F., 2016. Linguistic features identify Alzheimer's disease in narrative speech. J. Alzheimer's Dis. 49 (2), 407–422.
Garrard, P., Maloney, L.M., Hodges, J.R., Patterson, K., 2004. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain 128 (2), 250–260.
Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., Gorno-Tempini, M.L., 2014. Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. Cortex 55, 122–129.
Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R.C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., et al., 2006. Mild cognitive impairment. Lancet 367 (9518), 1262–1270.
Goldberg, T.E., Harvey, P.D., Wesnes, K.A., Snyder, P.J., Schneider, L.S., 2015. Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. Alzheimer's Demen. Diagnos. Assessm. Dis. Monit. 1 (1), 103–111.
Goodglass, H., Kaplan, E., 1983. The Assessment of Aphasia and Related Disorders. Lea & Febiger, Philadelphia.
Grut, M., Fratiglioni, L., Viitanen, M., Winblad, B., 1993. Accuracy of the mini-mental status examination as a screening test for dementia in a Swedish elderly population. Acta Neurol. Scand. 87 (4), 312–317.
Hillis Trupe, E., Hillis, A., 1985. Paucity vs. verbosity: another analysis of right hemisphere communication deficits. Clin. Aphasiology 15, 83–96.
Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L., Ogar, J., 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In: Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology, pp. 27–36.
Kaufman, L., Rousseeuw, P.J., 2009. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.
Kavé, G., Goral, M., 2016. Word retrieval in picture descriptions produced by individuals with Alzheimer's disease.. J. Clin. Exp. Neuropsychol. 38 (9), 958–966.
Kokkinakis, D., Lundholm Fors, K., Björkner, E., Nordlund, A., 2017. Data collection from persons with mild forms of cognitive impairment and healthy controls—infrastructure for classification and prediction of dementia. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22–24 May 2017, Gothenburg, Sweden. Linköping University Electronic Press, pp. 172–182.
König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H., et al., 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's Demen. 1 (1), 112–124.
Lai, Y.-H., Pai, H.-H., Lin, Y.-T., 2009. To be semantically-impaired or to be syntactically-impaired: Linguistic patterns in Chinese-speaking persons with or without dementia.. J. Neurolinguist. 22 (5), 465–475.
Landfeldt, E., Söderbäck, E., 2009. Predicerar Skriftliga Bildbeskrivningar Demens? - En Retrospektiv Studie (In Swedish).
Laske, C., Sohrabi, H.R., Frost, S.M., López-de Ipiña, K., Garrard, P., Buscema, M., Dauwels, J., Soekadar, S.R., Mueller, S., Linnemann, C., et al., 2015. Innovative diagnostic tools for early detection of Alzheimer's disease. Alzheimer's Demen. 11 (5), 561–578.
Le, X., Lancashire, I., Hirst, G., Jokel, R., 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Lit. Linguist. Comput. 26 (4), 435–461.
Lind, M., Kristoffersen, K.E., Moen, I., Simonsen, H.G., 2009. Semi-spontaneous oral text production: Measurements in clinical practice. Clinical Linguistics and Phonetics 23 (13), 872–886.
MacWhinney, B., 2000. The CHILDES project: tools for analyzing talk. third Lawrence Erlbaum Associates, Mahwah, New Jersey.
MacWhinney, B., 2007. The Talkbank project. Creating and Digitizing Language Corpora. Springer, pp. 163–180.
Masrani, V., Murray, G., Field, T., Carenini, G., 2017. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In: Proceedings of BioNLP, pp. 232–237.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, pp. 3111–3119.

Nensa, F., Beiderwellen, K., Heusch, P., Wetter, A., 2014. Clinical applications of PET/MRI: current status and future perspectives. Diagnos. Interven. Radiol. 20 (5), 438–447.

Orimaye, S.O., Wong, J.S., Golden, K.J., Wong, C.P., Soyiri, I.N., 2017. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. BMC Bioinform. 18 (1), 34.

Orimaye, S.O., Wong, J.S.-M., Golden, K.J., 2014. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In: Proc. First Workshop on Computational Linguistics and Clinical Psychology (CLPsych). ACL, pp. 78–87.

Pakhomov, S.V., Smith, G.E., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., Knopman, D.S., 2010. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. Cogn. Behav. Neurol. 23, 165–177.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pekkala, S., Wiener, D., Himali, J.J., Beiser, A.S., Obler, L.K., Liu, Y., McKee, A., Auerbach, S., Seshadri, S., Wolf, P.A., Au, R., 2013. Lexical retrieval in discourse: an early indicator of Alzheimer's dementia. Clin. Linguist. Phonet. 27 (12), 905–921.

Posner, H., Curiel, R., Edgar, C., Hendrix, S., Liu, E., Loewenstein, D.A., Morrison, G., Shinobu, L., Wesnes, K., Harvey, P.D., 2017. Outcomes assessment in clinical trials of Alzheimer's disease and its precursors: readying for short-term and long-term clinical trial needs. Innov. Clin. Neurosci. 14 (1,2), 22.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.P., 2013. The global prevalence of dementia: a systematic review and meta-analysis. Alzheimer's Demen. 9 (1), 63–75.

Prud'hommeaux, E., Roark, B., 2015. Graph-based word alignment for clinical language evaluation. Comput. Linguist. 41 (4), 549–578.

Reisberg, B., Gauthier, S., 2008. Current evidence for subjective cognitive impairment (SCI) as the pre-mild cognitive impairment (MCI) stage of subsequently manifest Alzheimer's disease. Int. Psychogeriat. 20 (1), 1–16.

Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C.A., Garrard, P., 2014. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. J. Alzheimer's Dis. 42 (S3), S3–S17.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., Kaye, J., 2011. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2081–2090.

Sirts, K., Piguet, O., Johnson, M., 2017. Idea density for predicting Alzheimer's disease from transcribed speech. In: Proceedings of the Twenty First Conference on Computational Natural Language Learning (CoNLL 2017), pp. 322–332.

Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y., 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. Proceedings of the International Conference on Learning Representations (ICLR).

Snowdon, D.A., Kemper, S.J., Mortimer, J.A., Greiner, L.H., Wekstein, D.R., Markesbery, W.R., 1996. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the Nun Study. J. Am. Med. Assoc. 275 (7), 528–532.

Stevens, S., 1985. The language of dementia in the elderly: a pilot study. Br. J. Dis. Commun. 20 (2), 181–190.

Stilwell, B.L., Dow, R.M., Lamers, C., Woods, R.T., 2016. Language changes in bilingual individuals with Alzheimer's disease. Int. J. Lang. Commun. Dis. 51 (2), 113–127.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., Pakaski, M., 2015. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front. Aging Neurosci. 7, 1–7.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., Asp, E., 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In: Proceedings of the IEEE International Conference on Mechatronics and Automation, pp. 1569–1574.

Toutanova, K., Klein, D., Manning, C., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL: HLT), pp. 252–259.

Trask, A., Michalak, P., Liu, J., 2015. Sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. arXiv:1511.06388.

Tyche, O., 2001. Subtila språkstörningar hos patienter med diagnosen MCI (Mild Cognitive Impairment) Del I: utifrån den tematiska bilden "Kakstölden" (In Swedish). Master's thesis, Karolinska institute.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., Stålhammar, J., Eckerström, M., Kern, S., Börjesson-Hanson, A., Carlsson, M., Olsson, E., Zetterberg, H., Blennow, K., Svensson, J., Öhrfelt, A., Bjerke, M., Rolstad, S., Eckerström, C., 2016. The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up.. J. Cereb. Blood Flow Metabol. Off. J. Int. Soc. Cereb. Blood Flow Metabol. 36 (1), 31–114.

Yancheva, M., Rudzicz, F., 2016. Vector-space topic models for detecting Alzheimer's disease.. In: Proceedings of the Fifty-Fourth Annual Meeting of the Association for Computational Linguistics (ACL), pp. 2337–2346.