LREC 2016 Workshop

Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)

Date: Monday, 23rd of May 2016

PROCEEDINGS

Editor:

Dimitrios Kokkinakis

Papers from the LREC Workshop: RaPID-2016, 23 May 2016, Portorož Slovenia.
Workshop Proceedings published by Linköping University Electronic Press.

LREC 2016 Workshop: Proceedings
"Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric impairments (RaPID-2016)"
23 May 2016 – Portorož, Slovenia
Edited by Dimitrios Kokkinakis
http://spraakbanken.gu.se/eng/rapid-2016

Linköping Electronic Conference Proceedings, No. 128
ISSN: 1650-3686
eISSN: 1650-3740
ISBN: 978-91-7685-730-4
URL: http://www.ep.liu.se/ecp/contents.asp?issue=128

Acknowledgments: This work has received support from Riksbankens Jubileumsfond - The Swedish Foundation for Humanities and Social Sciences through the grant agreement no:NHS14-1761:1; the Centre for Ageing and Health (AGECAP) and the Speech and Language Processing for Assistive Technologies (SIG-SLPAT).

Papers from the LREC Workshop: RaPID-2016, 23 May 2016, Portorož Slovenia.
Workshop Proceedings published by Linköping University Electronic Press.

Workshop Programme

Afternoon session, Monday, 23rd of May 2016, Portorož, Slovenia

Time
14:00 – 14:10 – Welcome and Introduction

14:10-14:55 Invited keynote talk by Dr Peter Garrard, St George's, University of London: *Neurobehavioural disease signatures in language corpora*

15:00-16:00 Session A
Kathleen C. Fraser and Graeme Hirst, *Detecting semantic changes in Alzheimer's disease with vector space models*

Christine Howes, Mary Lavelle, Patrick G.T. Healey, Julian Hough and Rose McCabe: *Helping hands? Gesture and self-repair in schizophrenia*

16:00 – 16:30 Coffee break

16:30-17:50 Session B
Spyridoula Varlokosta, Spyridoula Stamouli, Athanassios Karasimos, Georgios Markopoulos, Maria Kakavoulia, Michaela Nerantzini, Aikaterini Pantoula, Valantis Fyndanis, Alexandra Economou and Athanassios Protopapas: *A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications*

Mariya Khudyakova, Mira Bergelson, Yulia Akinina, Ekaterina Iskra, Svetlana Toldova and Olga Dragoy: *Russian CliPS: a Corpus of Narratives by Brain-Damaged Individuals*

Maksim Belousov, Mladen Dinev, Rohan M. Morris, Natalie Berry, Sandra Bucci and Goran Nenadic: *Mining Auditory Hallucinations from Unsolicited Twitter Posts*

Christopher Bull, Dommy Asfiandy, Ann Gledson, Joseph Mellor, Samuel Couth, Gemma Stringer, Paul Rayson, Alistair Sutcliffe, John Keane, Xiaojun Zeng, Alistair Burns, Iracema Leroi, Clive Ballard and Pete Sawyer: *Combining data mining and text mining for detection of early stage dementia: the SAMS framework*

17:50 – 18:00 Closing remarks

Papers from the LREC Workshop: RaPID-2016, 23 May 2016, Portorož Slovenia.
Workshop Proceedings published by Linköping University Electronic Press.

Editor

Dimitrios Kokkinakis	University of Gothenburg, Sweden

Workshop Organizers/Organizing Committee

Dimitrios Kokkinakis	University of Gothenburg, Sweden
Graeme Hirst	University of Toronto, Canada
Natalia Grabar	Université de Lille, France
Arto Nordlund	The Sahlgrenska Academy, Sweden
Jens Edlund	KTH - Royal Institute of Technology, Sweden
Åsa Wengelin	University of Gothenburg, Sweden
Simon Dobnik	University of Gothenburg, Sweden
Marcus Nyström	University of Lund, Sweden

Workshop Programme Committee

Jan Alexandersson	DFKI GmbH, Germany
Jonas Beskow	KTH - Royal Institute of Technology, Sweden
Heidi Christensen	University pf Sheffield, UK
Simon Dobnik	University of Gothenburg, Sweden
Jens Edlund	KTH - Royal Institute of Technology, Sweden
Gerardo Fernández	Universidad Nacional del Sur, Argentina
Peter Garrard	St George's, University of London, UK
Kallirroi Georgila	University of Southern California, USA
Natalia Grabar	Université de Lille, France
Nancy L. Green	U. of North Carolina at Greensboro, USA
Katarina Heimann Mühlenbock	University of Gothenburg, Sweden
Graeme Hirst	University of Toronto, Canada
Kristy Hollingshead	Florida Institute for Human & Machine Cognition (IHMC), USA
William Jarrold	Nuance Communications, USA
Richard Johansson	University of Gothenburg, Sweden
Dimitrios Kokkinakis	University of Gothenburg, Sweden
Yiannis Kompatsiaris	Centre for Research & Technology Hellas, Greece
Alexandra König	Toronto Rehabilitation Institute, Canada
Peter Ljunglöf	Chalmers University of Technology, Sweden
Karmele López-de-Ipiña	U. of the Basque Country (UPV/EHU), Spain
Arto Nordlund	The Sahlgrenska Academy, Sweden
Marcus Nyström	University of Lund, Sweden
François Portet	Laboratoire d'informatique de Grenoble, France
Vassiliki Rentoumi	SKEL, NCSR Demokritos, Greece
Frank Rudzicz	University of Toronto, Canada
Paul Thompson	Dartmouth College, USA
Magda Tsolaki	Aristotle University of Thessaloniki, Greece
Spyridoula Varlokosta	National & Kapodistrian U. of Athens, Greece
Åsa Wengelin	University of Gothenburg, Sweden
Maria Wolters	University of Edinburgh, UK

Papers from the LREC Workshop: RaPID-2016, 23 May 2016, Portorož Slovenia.
Workshop Proceedings published by Linköping University Electronic Press.

Table of Contents

Papers from the LREC Workshop: RaPID-2016, 23 May 2016, Portorož Slovenia.
Workshop Proceedings published by Linköping University Electronic Press.

Author Index

Papers from the LREC Workshop: RaPID-2016, 23 May 2016, Portorož Slovenia.
Workshop Proceedings published by Linköping University Electronic Press.

Preface/Introduction

The purpose of the Workshop on "*Resources and ProcessIng of linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments*" (RaPID-2016) was to provide a snapshot view of some of the current technological landscape, resources, data samples and also needs and challenges in the area of processing various data from individuals with various types of mental and neurological health impairments and similar conditions at various stages; increase the knowledge, understanding, awareness and ability to achieve useful outcomes in this area and strengthen the collaboration between researchers and workers in the field of clinical/nursing/medical sciences and those in the field of language technology/computational linguistics/Natural Language Processing (NLP).

Although many of the causes of cognitive and neuropsychiatric impairments are difficult to foresee and accurately predict, physicians and clinicians work with a wide range of factors that potentially contribute to such impairments, e.g., traumatic brain injuries, genetic predispositions, side effects of medication, and congenital anomalies. In this context, there is new evidence that the acquisition and processing of linguistic data (e.g., spontaneous story telling) and extra-linguistic and production measures (e.g., eye tracking) could be used as a complement to clinical diagnosis and provide the foundation for future development of objective criteria to be used for identifying progressive decline or degeneration of normal mental and brain functioning.

An important new area of research in NLP emphasizes the processing, analysis, and interpretation of such data and current research in this field, based on linguistic-oriented analysis of text and speech produced by such a population and compared to healthy adults, has shown promising outcomes. This is manifested in early diagnosis and prediction of individuals at risk, the differentiation of individuals with various degrees of severity forms of brain and mental illness, and for the monitoring of the progression of such conditions through the diachronic analysis of language samples or other extra-linguistic measurements. Initially, work was based on written data but there is a rapidly growing body of research based on spoken samples and other modalities.

Nevertheless, there remains significant work to be done to arrive at more accurate estimates for prediction purposes in the future and more research is required in order to reliably complement the battery of medical and clinical examinations currently undertaken for the early diagnosis or monitoring of, e.g., neurodegenerative and other brain and mental disorders and accordingly, aid the development of new, non-invasive, time and cost-effective and objective (future) clinical tests in neurology, psychology, and psychiatry.

Papers were invited in all of the areas outlined in the *topics of interest* below particularly emphasizing multidisciplinary aspects of processing such data and also on the exploitation of results and outcomes and related ethical questions. Specifically, in the call for papers we solicited papers on the following topics:

- Building and adapting domain relevant linguistic resources, data, and tools, and making them available.
- Data collection methodologies.
- Acquisition of novel data samples, e.g. from digital pens (i.e., digital pen strokes) or keylogging and integrating them with data from various sources (i.e., information fusion).

Papers from the LREC Workshop: RaPID-2016, 23 May 2016, Portorož Slovenia.
Workshop Proceedings published by Linköping University Electronic Press.

- Guidelines, annotation schemas, and tools (e.g., for semantic annotation of data sets).
- Addressing the challenges of representation, including dealing with data sparsity and dimensionality issues, and feature combination from different sources and modalities,
- Adaptation of standard NLP tools to the domain.
- Syntactic, semantic, and pragmatic analysis of data, including modelling of perception (e.g., eye-movement measures of reading) and production processes (e.g., recording the writing process with digital pens, keystroke logging, etc.), use of gestures accompanying speech and non-linguistic behaviour.
- Machine learning approaches for early diagnosis, prediction, monitoring, classification, etc. of various cognitive, psychological, and psychiatric impairments, including unsupervised methods (e.g., distributional semantics).
- Evaluation of tools, systems, components, metrics, applications, and technologies that make use of NLP in the domain.
- Evaluation, comparison, and critical assessment of resources.
- Evaluation of the significance of extracted features.
- Involvement of medical professionals and patients and ethical questions.
- Deployment of resources.
- Experiences, lessons learned, and the future of NLP in the area.

Most of these topics lie at the heart of the papers that were accepted to the workshop which features 6 oral presentations.

We would like to thank all the authors who submitted papers, as well as the members of the Program Committee for the time and effort they contributed in reviewing the papers. We are also grateful to Dr Peter Garrard for accepting to give an invited talk at the workshop entitled: “Neurobehavioural disease signatures in language corpora”.

The Editor

Papers from the LREC Workshop: RaPID-2016, 23 May 2016, Portorož Slovenia.
Workshop Proceedings published by Linköping University Electronic Press.

Detecting semantic changes in Alzheimer's disease with vector space models

Kathleen C. Fraser, Graeme Hirst
Department of Computer Science
University of Toronto, Toronto, Canada
kfraser@cs.toronto.edu, gh@cs.toronto.edu

Abstract

Numerous studies have shown that language impairments, particularly semantic deficits, are evident in the narrative speech of people with Alzheimer's disease from the earliest stages of the disease. Here, we present a novel technique for capturing those changes, by comparing distributed word representations constructed from healthy controls and Alzheimer's patients. We investigate examples of words with different representations in the two spaces, and link the semantic and contextual differences to findings from the Alzheimer's disease literature.

Keywords: distributional semantics, Alzheimer's disease, narrative speech

1. Introduction

Vector space models of semantics have become an increasingly popular area of research in computational linguistics, with notable successes on tasks such as query expansion for information retrieval (Manning et al., 2008), synonym identification (Bullinaria and Levy, 2012), sentiment analysis (Socher et al., 2012), machine translation (Zou et al., 2013), and many others. Here we present a preliminary study on how we can use vector space models to detect semantic changes that may occur with Alzheimer's disease (AD).

The general idea is simple: if we construct two semantic spaces from two different corpora, we expect the differences between the spaces to be related to the differences between the corpora. If we generate word vectors from a corpus of text about cars, and another set of word vectors from a corpus about wildlife, we expect that the word *jaguar* will have two very different representations in these two spaces. If the dimensions are the same, we can measure the distance between the two vectors for *jaguar*, and we would expect to find that it is non-zero.

In this study, we fix the topic of the two corpora to be the same: each document in the two corpora is a description of the "Cookie Theft" picture shown in Figure 1. Rather, the difference is that the documents in one corpus were produced by people with AD, and the documents in the other corpus were produced by healthy, older controls. We suggest that differences in the vector representations trained on the two corpora will be due, at least in part, to the semantic impairment that often occurs with AD.

In the following sections, we first present a brief summary of the literature on language in AD as well as related computational work. We then describe our data and procedure, examine the differences between the two corpora using a simple vector representation, and present three methods to help interpret these differences, with specific examples from the narratives. We also discuss the limitations of this study and suggest ways to build on these preliminary results.

2. Background

A great deal of work has been undertaken studying the degradation of semantic processing in AD, of which we will only begin to scratch the surface in this discussion.

Semantic memory deficits have been widely reported, with AD patients having difficulty on naming tasks and often substituting high-frequency hypernyms or semantic neighbours for target words which cannot be accessed (Kempler, 1995; Giffard et al., 2001; Kirshner, 2012; Reilly et al., 2011). Numerous studies have reported a greater impairment in category naming fluency (e.g., naming animals or tools) relative to letter naming fluency (e.g., naming words that start with the letter R) (Salmon et al., 1999; Monsch et al., 1992; Adlam et al., 2006). As a result of word-finding difficulties and a reduction in working vocabulary, the language of AD patients can seem "empty" (Ahmed et al., 2013) and lacking coherence (Appell et al., 1982). In the famous "Nun Study" (Snowdon et al., 1996), it was shown that decreased idea density in writing produced in early life was associated with developing Alzheimer's disease decades later.

Specifically with regards to the "Cookie Theft" picture description task that we consider here, AD patients tend to show a reduction in the amount of information that is conveyed (Giles et al., 1996; Croisile et al., 1996; Lira et al., 2014). That is, they do not mention all the expected facts or inferences about the picture. Furthermore, these impairments are noticeable from a very early stage in the disease (Forbes-McKay and Venneri, 2005). Nicholas et al. (1985) found that AD patients mentioned roughly half of the expected information units, and produced a large number of deictic terms and indefinite terms (e.g. pronouns without antecedents). Ahmed et al. (2013) found that AD patients made fewer references to the people and their actions depicted in the picture than controls.

Recently, there has been some progress on automatically determining the information content of picture description narratives using computational techniques. Pakhomov et al. (2010) generated a list of expected information units and some of their lexical and morphological variants, then searched for matches. Hakkani-Tür et al. (2010) scored picture descriptions using information retrieval techniques

to match the narratives with a list of 35 key concepts.
In previous work, we used a combination of keyword-spotting and dependency parsing to identify relevant information units in "Cookie Theft" narratives (Fraser et al., 2015). However, accurately identifying atypical speech patterns will require accounting for not just *what* words are used, but *how* they are used. A better understanding of the semantic space and the different senses in which words are used will be a first step towards better models for detecting AD from speech.

3. Data

The narrative speech data were obtained from the Pitt corpus in the DementiaBank database[1] (MacWhinney, 2007). These data were collected between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh. Detailed information about the study cohort is available from Becker et al. (1994), and demographic information is given in Table 1. Unfortunately, the patient and control groups are not matched for age and education; the AD patients tend to be both older ($p < 0.01$) and less educated ($p < 0.01$), which is one limitation of this data set. There is no significant difference on sex ($p = 0.8$).
The language samples were elicited using the "Cookie Theft" picture description task from the Boston Diagnostic Aphasia Examination (BDAE) (Goodglass and Kaplan, 1983), in which participants are asked to describe everything they see going on in a picture. The stimulus picture is shown in Figure 1. The data were manually transcribed following the CHAT transcription protocol (MacWhinney, 2000).
Patients in the Pittsburgh study were diagnosed on the basis of their clinical history and their performance on neuropsychological testing, and the diagnoses were updated in 1992, taking into account any relevant information from the intervening years. Autopsies were performed on 50 patients, and in 43 cases the AD diagnosis was confirmed (86.0%) (Becker et al., 1994). A more recent study of clinical diagnostic accuracy in AD found that of 526 cases diagnosed as probable AD, 438 were confirmed as neuropathological AD post-mortem (83.3%) (Beach et al., 2012), suggesting that the DB diagnoses are generally as reliable as diagnoses made using present-day criteria.
We include 240 narratives from 167 participants diagnosed with possible or probable AD (average number of narratives per participant is 1.44, median is 1.0), and 233 narratives from 97 healthy, elderly controls (average 2.40, median 2.0). As shorthand throughout the paper, we refer to the set of narratives from participants with AD as the "AD corpus", and the set of narratives from healthy controls as the "CT corpus." In total, the AD corpus contains 31,906 words, and the CT corpus contains 27,620 words.

4. Differences in word representations between AD patients and controls

To compare the vector spaces directly, we require that the dimensions be identical (in interpretation as well as number). For this reason, we do not consider popular neural network models such as skip-gram and CBOW (Mikolov et al., 2013), whose resulting dimensions are not easily interpretable. Instead we consider a simple word-word co-occurrence model, in which the rows and columns represent words from the vocabulary, and the value of (r_i, c_i) is the number of times context word c_i appears near the given word r_i. We use a window size of three words on each side of the target word, with the exception of words at the beginning and end of narrative samples (i.e. the window is not permitted to overlap with the end of one sample and the beginning of the next). To reduce data sparsity, we consider only words that occur a minimum of 10 times in both the CT and AD corpora, and we lemmatize the words using NLTK's WordNet lemmatizer, after first tagging the words to increase lemmatization accuracy (Bird et al., 2009). After examining the frequency distribution of words in the two corpora, we decided not to remove any stop-words, as many of the highest frequency words are actually content words (e.g. *cookie*). Furthermore, we predict that common words such as prepositions and pronouns might show some variation in usage between the groups.
We stated above that the most likely reason for differences between the vector representations would be differences between the language of people with AD and healthy controls. Of course, another reason for differences could simply be random variation in word choice and speaking style between individuals, which may be a factor here given the relatively small size of the data set. To mitigate this effect, we adopt the following procedure:

1. First, split the CT corpus in half, create two co-occurrence matrices, and measure the cosine distance

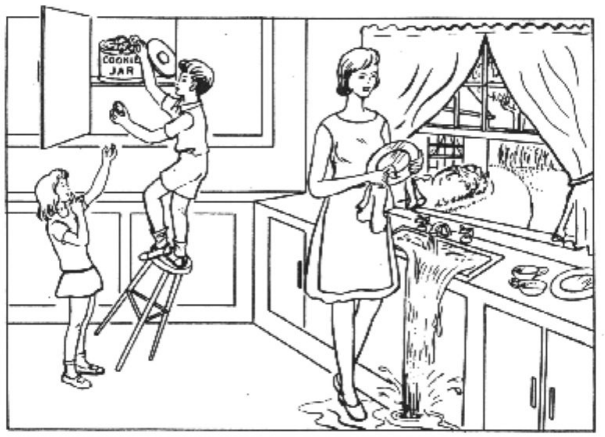


Figure 1: The Cookie Theft Picture (Goodglass and Kaplan, 1983).

	AD	Controls
	$n = 240$	$n = 233$
Age	71.8 (8.5)	65.2 (7.8)
Education	12.5 (2.9)	14.1 (2.4)
Sex (M/F)	82/158	82/151
MMSE	18.5 (5.1)	29.1 (1.1)

Table 1: Demographic information (mean and standard deviation).

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

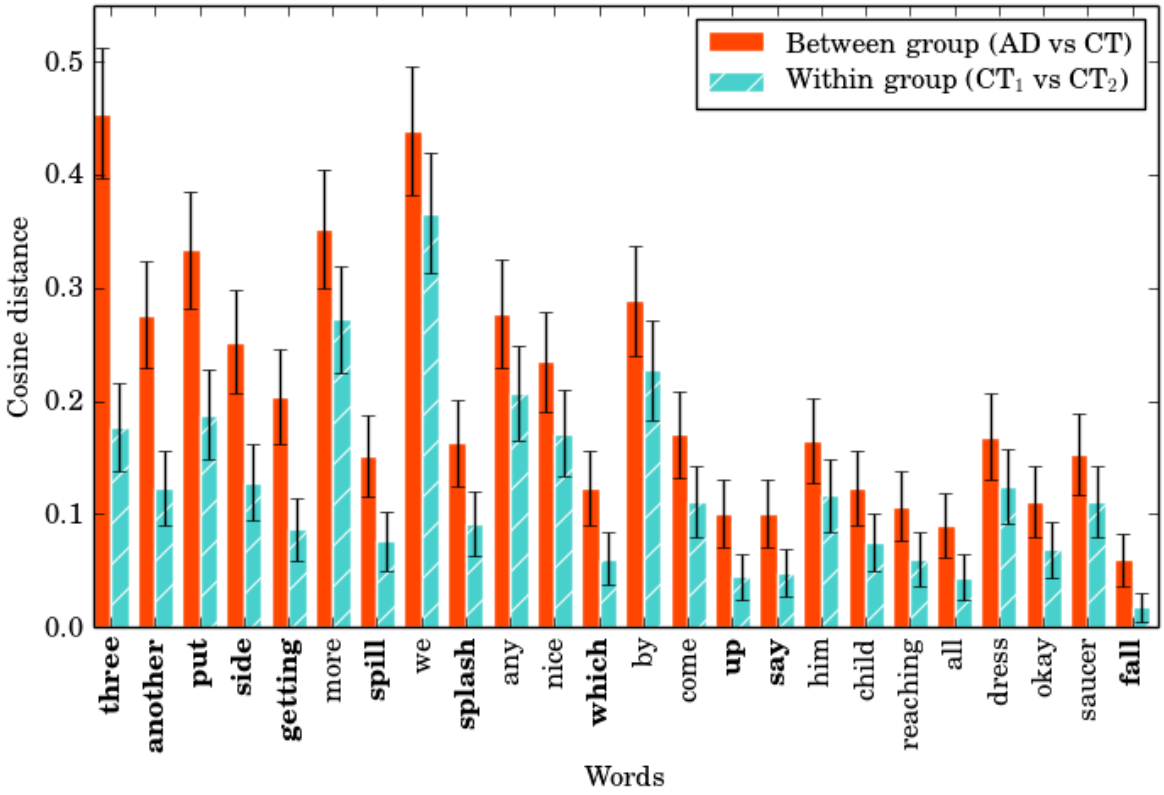


Figure 2: Cosine distances within the control group, and between the control and AD groups. Words marked in bold on the horizontal axis were selected for analysis in the next section (the difference between groups was greater than the error).

between the vector for word *w* in the first and the vector for word *w* in the second. This gives us an idea of the expected variation that occurs for each word.

2. Then, measure the cosine distance between the representation of word *w* trained on the CT corpus and trained on the AD corpus. This represents the variation of the word across the two groups.

3. Finally, only select a word for analysis if the distance *across* groups is greater than the distance *within* the control group; that is, if the variation between AD patients and controls is greater than the normal variation within healthy speakers. (Note that we do not consider the variation within the AD group in this calculation, as we want to measure whether the across-group variation is greater than *typical* variation.)

One difficulty that we encountered in performing this calculation was choosing an appropriate metric for measuring statistical significance in cosine differences. We experimented with partitioning the data into several folds, obtaining observations from each fold, and then testing for significance (as in Bullinaria and Levy (2012)), but concluded that our data set is simply too small for this method to be feasible. For lack of a better option, and given the close relationship between cosine similarity and correlation (Van Dongen and Enright, 2012), we instead computed the standard error for each cosine distance (treating each word as an observation). Figure 2 shows the cosine distances and errors for a subset of vectors, including all of those which were selected for analysis in the following section.

5. Interpretation of the differences

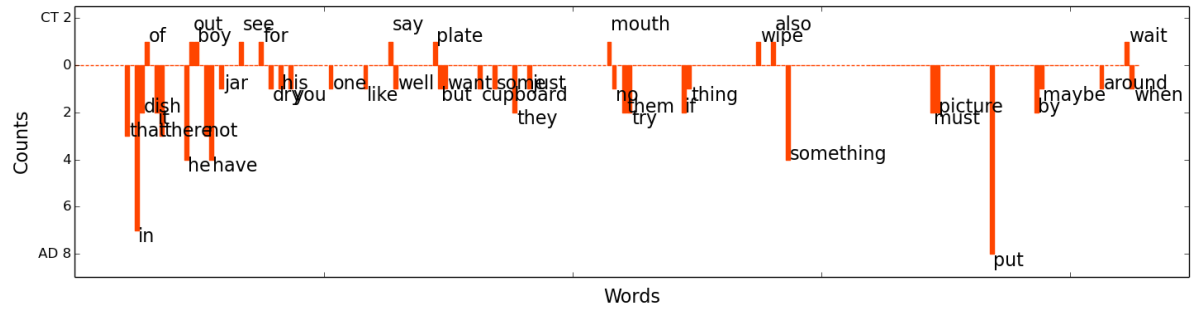
The method described above leaves us with a fairly small set of vectors that differ notably between the groups (i.e. only those 11 features shown in bold in Figure 2). The next question is: in what way are they different, and how does this relate to our knowledge of Alzheimer's disease? This proved to be a more difficult question to answer. In the following sections we present three different approaches, with illustrative examples for each.

5.1. Contextual differences

As a first step, we examined those dimensions which were non-zero in one group and zero in the other (i.e. context for a given word that appeared in one group but not the other). In the selected words, two different scenarios were observed: In the first case, the control participants used a number of context words not used by the AD participants. An example of this is shown in Figure 3a, for the word *another*. Two context words which occur fairly often with *another* in the control group are *he* and *window*. Some examples of these words in context (words inside the context window are italicized) are:

- And *he's getting another one out of* the cookie jar
- *He's handing another one to the* little girl
- *And there's another window and some* trees apparently

(a) another



(b) put

Figure 3: Differences in context for the words *another* and *put*. Counts above the horizontal axis indicate context words that occurred only in the control group; counts below the horizontal axis indicate words that occurred only the AD group.

- You can *see beyond that another window and um*

These examples demonstrate a certain attention to detail — to say that the boy is getting *another cookie*, one must first observe that he already has a cookie in his hand; the second window referenced in *another window* is a minor detail seen through the first, more prominent window. They also reflect an element of cohesion, in that *another window* makes reference to an earlier window mentioned by the speaker, and clarifies to the listener that this new reference to a window is distinct from the previous one. Prior work has shown that attention is often one of the first areas of cognition (after memory) to be affected in AD (Perry and Hodges, 1999), and that the narratives of people with AD tend to show a lack of cohesion (Chenery and Murdoch, 1994).

In the second case, the AD participants use a number of context words that do not occur in the CT corpus. An example of this is illustrated in Figure 3b, for the word *put*. One of the most frequent context words in the figure is *put* itself. Control participants rarely repeat the same word within the 6-word context window, but it is not uncommon in the AD group. Another interesting context word is *in*. The controls do not tend to describe any of the actions in the Cookie Theft picture as *putting* something *in* something else. They use *put* to describe the action of the girl (e.g. the little girl is putting her finger to her mouth). On the other hand, some examples from the AD corpus include:

- he's *trying to put put put food in* that in that crocker jar
- has a cookie jar up *there he's putting cookies in and* the thing's falling over

These errors are similar to the "implausible details" that Croisile et al. (1996) found to occur more frequently in AD narratives than controls. The underlying explanation is unclear, although it could represent a breakdown in logic and understanding. It also demonstrates a potential pitfall of the keyword-spotting approach to scoring — a participant may mention the boy and the cookie jar, but the action connecting the two is also fundamentally important.

5.2. Vectors shifting in space

Another way of looking at these differences is to see how the words in question have moved in the vector space. To visualize the space in two dimensions, we use the method of t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008). The t-SNE method was proposed as a solution to the problem of visualizing high-dimensional data in two or three dimensions. It is capable of producing visualizations that reveal structure at both the local and global level, although the resulting dimensions are not generally interpretable (and therefore not labelled in the figures). The example word we consider here is *getting*. (Note that verbs ending in *-ing* are subject to a consistent issue in the tagging and lemmatizing pipeline which results in them not being reduced to the base form.) Figure 4 shows part of the two-dimensional representation of the word vector space. In many cases, the word representations in the AD and CT corpora lie very close to each other. However, in the case of *getting*, the vectors lie much further apart. Examining the surrounding vectors, it appears that *getting* is closer to *running*, *overflowing*, and

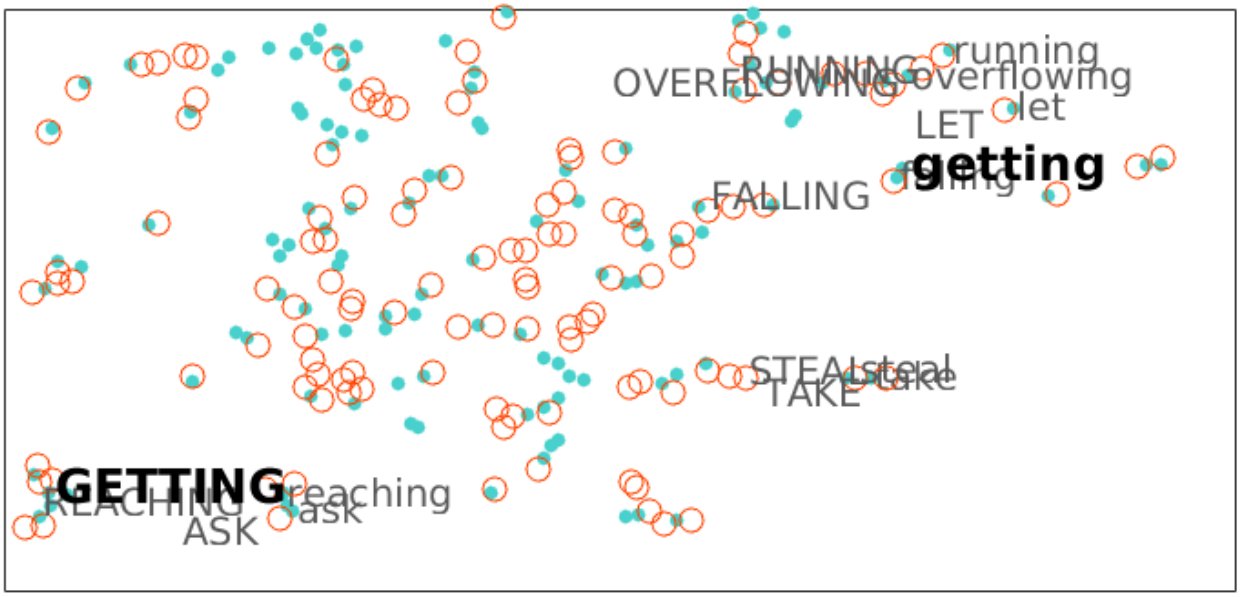


Figure 4: Two-dimensional visualization of the vector space using t-SNE. Word representations from the AD corpus are labelled with filled, green circles and lowercase labels; words from the CT corpus are labelled with open, orange circles and uppercase labels.

falling in the AD corpus, and closer to words like *reaching* and *ask* in the CT corpus. This is confirmed by comparing the cosine distances (Table 2).

The nearest neighbours of the vectors suggest that in the AD corpus, *getting* is used more in the context of the sink overflowing, while in the CT corpus *getting* is used in the context of the cookie theft. This is borne out in the data itself, as there is only one example in the control group of using *getting* in the context of the sink (*her foot is getting wet*) and the rest refer to the act of stealing the cookie. In the AD group there are a number of references to the sink context (e.g. *the floor is getting wet*, *mom is getting her foot wet*, *the water is getting over the sink*), as well as referring to stealing the cookie.

One explanation for this phenomenon could lie in the fact that *get* is a "light" verb, in that it does not convey very much semantic information about the action it describes (Breedin et al., 1998). Kim and Thompson (2004) showed that people with Alzheimer's disease produced more light verbs and fewer heavy verbs in a story-telling task, and were more impaired on retrieving heavy verbs than light verbs in a sentence completion task. Both AD and control speakers use *getting* in the sense of *getting a cookie*, which is the primary sense of the word *get*[2], meaning "to obtain or procure". However, we expect that AD speakers may also substitute light, easy-to-access verbs for more semantically appropriate verbs. This phenomenon is observed when, for example, an AD participant says *the water is getting over the sink* rather than *the water is flowing/overflowing/running over the sink*.

5.3. Cluster analysis

The example in the previous section illustrates how a single word can have multiple senses, and these senses can be distinguished by the different contexts in which they appear. This idea is the basis of the Distributional Hypothesis, stated perhaps most succinctly as, "you shall know a word by the company it keeps" (Firth, 1957). The difficulty of representing the different senses of a word was an issue in the early days of vector space representations, although numerous solutions have been proposed since (Reisinger and Mooney, 2010; Huang et al., 2012; Guo et al., 2014; Wu and Giles, 2015).

Words	Distance (AD corpus)	Distance (CT corpus)
getting, running	**0.225**	0.524
getting, overflowing	**0.224**	0.511
getting, falling	**0.191**	0.384
getting, reaching	0.395	**0.349**
getting, ask	0.588	**0.482**

Table 2: Cosine distances for the words in Figure 4 in each of the two vector representations (trained on AD corpus and CT corpus).

In this section, we perform an analysis based on methods for unsupervised word sense discovery. We take a step back, and rather than considering the final vector representation for a word, we look at all the context vectors that contribute to the final vector and perform cluster analysis on them. Different clusters will represent different contexts, and by assumption different word senses. We use k-means clustering with a Euclidean distance metric. The optimal number of clusters k is chosen manually, by the silhouette method (Rousseeuw, 1987).

Our example for this section is the word *three*. The clusters for $k = 5$ are shown in two-dimensions in Figure 5a. The lemmatized context words associated with each point are given in Figure 5b. Our interpretation of the clusters

[2] http://www.oed.com/view/Entry/77994

is subjective, but in general we see that both AD participants and controls use the word *three* to describe the three-legged stool (turquoise cluster), as well as the number of dishes and the number of people in the room (purple cluster). Of the smaller outlier clusters, two consist of only contexts from the AD corpus. The orange cluster is made up of examples where participants described three cups, which is not a semantically accurate representation of the picture (there are three dishes, but one is a plate). In the green cluster, a single AD participant repeats the word *three* (thus creating two instances with very similar contexts) and uses the context word *woman*, which is unusual (the unlemmatized transcript reads *one two three three women*). Interestingly, although they were placed in a separate cluster, we also see this "counting" use of *three* in context vectors 4 and 7 (both from the AD corpus). Finally, the red cluster consists of two cases of repeating the word *three* in the context of the stool — one from the CT corpus and one from the AD corpus.

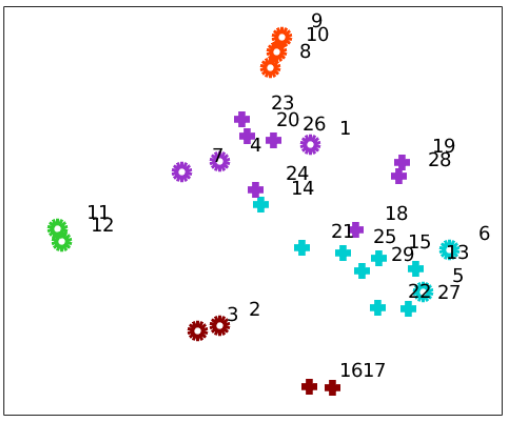
To summarize, in this example we see word senses (contexts) that are used by both AD participants and controls, and then we see other rare senses that appear only in the AD corpus. In this particular case, those rare senses correspond to semantic errors, although more work will be needed to see if that result generalizes to other words.

6. Limitations

Models based on raw word co-occurrence counts are perhaps the most basic distributional models of semantics, and it is known that performance is usually improved by (a) transforming the raw counts using methods like positive pointwise mutual information, or (b) learning predictive models using neural networks (Baroni et al., 2014). However, as the model increases in complexity, we face the issue of whether the Alzheimer's model and control model can still be compared directly. The literature on comparing different semantic spaces is relatively sparse, with some exceptions (Zuccon et al., 2009). In future work we plan to build on the baseline we have presented here by exploring different vector space models and methods for comparing them.

Furthermore, certain aspects of our methodology, such as choosing the optimal number of clusters, involve human intervention and have some degree of subjectivity. Work on automatically choosing k and then evaluating the purity of clusters and identifying outlier clusters is currently under way.

Moving away from computational details and looking at the big picture, it is clear that this study faces the same problem as many others: trying to study individual variation at the population level. We do not expect that every person with AD will say the boy is putting cookies into the jar, or that there are three cups on the counter. Rather, we expect that most people with AD will start to make semantic errors of some kind. In doing this analysis, we have picked up on semantic errors that we did not find using our previous approach (Fraser et al., 2015), and which have not been reported, to our knowledge, in any previous work using this data set. However, our approach here was similar in some ways to a case study, where we dug deep into a few representative examples. The true value will lie in scaling our methods to detect and count general semantic irregularities, which can then be used as input to a system for screening, longitudinal assessment, or diagnostic support.



(a) Two-dimensional visualization of the clusters using t-SNE. Circles represent contexts from the AD corpus, while plus signs represent contexts from the CT corpus.

Purple	1	okay there be person in the
	4	uh two dish dish sit on
	7	have two two four five door
	19	oh I see people in there
	20	dish there be dish set on
	23	flower there be dish leave to
	24	of the on of the cupboard
	26	back there be dish on the
	28	can see those thing
Red	2	xxx doing xxx three legged stool
	3	doing xxx three legged stool I
	16	fall off a three prong stool
	17	off a three prong stool his
Turquoise	5	stand on a legged stool and
	6	depart from that legged stool he
	13	a stool a legged stool and
	14	upend on the legged stool uh
	15	be a a legged stool and
	18	it be a it be a
	21	his sister the legged stool be
	22	be on a legged stool which
	25	stool be a legged stool um
	27	up on a legged stool which
	29	it be a legged stool and
Orange	8	running there be cup and uh
	9	the floor and cup three bowl
	10	and three cup bowl there
Green	11	be one two three woman has
	12	one two three woman has some

(b) The lemmatized contexts corresponding to each point.

Figure 5: Clustering of contexts for the word *three*.

7. Conclusion

We have presented preliminary results showing that the changes in word usage that occur in Alzheimer's disease

can be detected through analysis of the resulting semantic space. We examined these differences through visual analysis of the vectors themselves, two-dimensional representations of the vector spaces, and cluster analysis of the individual context vectors. Many of the differences are consistent with previous work on language changes in AD. Future work will focus on how these methods can be applied to automated scoring of the picture description task, or generating meaningful features for a diagnostic classifier.

8. Acknowledgements

The authors would like to thank Tong Wang and Maria Yancheva for their helpful discussions. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant number RGPIN-2014-06020.

9. References

Adlam, A.-L. R., Bozeat, S., Arnold, R., Watson, P., and Hodges, J. R. (2006). Semantic knowledge in mild cognitive impairment and mild Alzheimer's disease. *Cortex*, 42(5):675–684.

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

Appell, J., Kertesz, A., and Fisman, M. (1982). A study of language functioning in Alzheimer patients. *Brain and language*, 17(1):73–91.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–247.

Beach, T. G., Monsell, S. E., Phillips, L. E., and Kukull, W. (2012). Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. *Journal of Neuropathology & Experimental Neurology*, 71(4):266–273.

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

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Breedin, S. D., Saffran, E. M., and Schwartz, M. F. (1998). Semantic factors in verb retrieval: An effect of complexity. *Brain and Language*, 63:1–31.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3):890–907.

Chenery, H. J. and Murdoch, B. E. (1994). The production of narrative discourse in response to animations in persons with dementia of the Alzheimer's type: Preliminary findings. *Aphasiology*, 8(2):159–171.

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

Firth, J. R. (1957). A synopsis of linguistic theory. *Studies in Linguistic Analysis*, pages 1–32.

Forbes-McKay, K. E. and Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences*, 26:243–254.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2015). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Giffard, B., Desgranges, B., Nore-Mary, F., Lalevée, C., de la Sayette, V., Pasquier, F., and Eustache, F. (2001). The nature of semantic memory deficits in Alzheimer's disease. *Brain*, 124(8):1522–1532.

Giles, E., Patterson, K., and Hodges, J. R. (1996). Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: missing information. *Aphasiology*, 10(4):395–408.

Goodglass, H. and Kaplan, E. (1983). *Boston diagnostic aphasia examination booklet*. Lea & Febiger Philadelphia, PA.

Guo, J., Che, W., Wang, H., and Liu, T. (2014). Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 497–507.

Hakkani-Tür, D., Vergyri, D., and Tür, G. (2010). Speech-based automated cognitive status assessment. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 258–261.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 873–882. Association for Computational Linguistics.

Kempler, D. (1995). Language changes in dementia of the Alzheimer type. *Dementia and Communication*, pages 98–114.

Kim, M. and Thompson, C. K. (2004). Verb deficits in Alzheimer's disease and agrammatism: Implications for lexical organization. *Brain and Language*, 88(1):1–20.

Kirshner, H. S. (2012). Primary progressive aphasia and Alzheimer's disease: brief history, recent evidence. *Current neurology and neuroscience reports*, 12(6):709–714.

Lira, J. O. d., Minett, T. S. C., Bertolucci, P. H. F., and Ortiz, K. Z. (2014). Analysis of word number and content in discourse of patients with mild to moderate Alzheimer's disease. *Dementia & Neuropsychologia*, 8(3):260–265.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs*. Lawrence Erlbaum Associates.

MacWhinney, B. (2007). The Talkbank Project. In Beal, J., Corrigan, K., and Moisl, H. L., editors, *Creating and Digitizing Language Corpora*, pages 163–180. Springer.

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008).

Introduction to Information Retrieval, volume 1. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., and Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of Neurology*, 49(12):1253–1258.

Nicholas, M., Obler, L. K., Albert, M. L., and Helm-Estabrooks, N. (1985). Empty speech in Alzheimer's disease and fluent aphasia. *Journal of Speech, Language, and Hearing Research*, 28(3):405–410.

Pakhomov, S. V., Smith, G. E., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., and Knopman, D. S. (2010). Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23:165–177.

Perry, R. J. and Hodges, J. R. (1999). Attention and executive deficits in Alzheimer's disease. *Brain*, 122(3):383–404.

Reilly, J., Troche, J., and Grossman, M. (2011). Language processing in dementia. *The handbook of Alzheimer's disease and other dementias*, pages 336–368.

Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 109–117. Association for Computational Linguistics.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Salmon, D. P., Butters, N., and Chan, A. S. (1999). The deterioration of semantic memory in Alzheimer's disease. *Canadian Journal of Experimental Psychology*, 53(1):108.

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

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Van Dongen, S. and Enright, A. J. (2012). Metric distances derived from cosine similarity and Pearson and Spearman correlations. *arXiv preprint arXiv:1208.3145*.

Wu, Z. and Giles, C. L. (2015). Sense-aaware semantic analysis: A multi-prototype word representation model using wikipedia. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-15)*, pages 2188–2194.

Zou, W. Y., Socher, R., Cer, D. M., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1393–1398.

Zuccon, G., Azzopardi, L. A., and Van Rijsbergen, C. (2009). Semantic spaces: Measuring the distance between different subspaces. In *The Proceedings of the 3rd International Symposium on Quantum Interaction*, pages 225–236.

Helping hands? Gesture and self-repair in schizophrenia

Christine Howes, Mary Lavelle, Patrick G.T. Healey, Julian Hough, Rose McCabe
University of Gothenburg, King's College London, Queen Mary University of London, Bielefeld University, Exeter University
christine.howes@gu.se, mary.lavelle@kcl.ac.uk, p.healey@qmul.ac.uk, julian.hough@uni-bielefeld.de, r.mccabe@exeter.ac.uk

Abstract

Successful social encounters require mutual understanding between interacting partners, and patients with schizophrenia are known to experience difficulties in social interaction. Several studies have shown that in general people compensate for verbal difficulties by employing additional multimodal resources such as hand gesture. We hypothesise that this will be impaired in patients with schizophrenia, and present a preliminary study to address this question. The results show that during social interaction, schizophrenia patients repair their own speech less. In addition, although increased hand gesture is correlated with increased self-repair in healthy controls, there is no such association in patients with schizophrenia, or their interlocutors. This suggests that multimodal impairments are not merely seen on an individual level but may be a feature of patients' social encounters.

Keywords: Gesture, Self-repair, Schizophrenia

1. Introduction

Many patients with schizophrenia experience difficulty engaging in successful social interaction. This difficulty presents prior to the onset of defining symptoms of schizophrenia, such as hallucinations or delusions, is persistent and stable over time, and is associated with patients' poorer prognosis (Addington and Addington, 2008; Monte et al., 2008).

Successful social encounters require mutual understanding between interacting partners. To achieve this, conversational partners must monitor their own and their interlocutors' behaviour for potential misunderstandings, and attempt to address them as they arise. One way in which this can be done is self-repair (Schegloff et al., 1977), where the speaker identifies, and repairs or revises, their own speech as it is being produced.

The presence and amount of repair used by patients with schizophrenia may be indicative of some of the specific difficulties patients have in interacting with others. Research shows that, for non-clinical participants, the presence of repair can aid comprehension (Brennan and Schober, 2001) and that when verbal difficulties are encountered people may compensate by using additional multimodal resources such as hand gesture (Seyfeddinipur and Kita, 2014; Healey et al., 2015) and head nods (Healey et al., 2013).

In the psychiatric domain, levels of repair have been found to be associated with verbal hallucinations, and patient adherence to treatment (Leudar et al., 1992; McCabe et al., 2013). In addition, patients with schizophrenia are known to use fewer repairs in their talk (Leudar et al., 1992; Caplan et al., 1996), however, these findings are based on instruction giving or narrative tasks. Although these tasks ostensibly involve interaction in that the talk is designed for a listener, they tend to be monologic in practice. It is unclear if patients' performance on such tasks reflects their ability to interpret and respond to others during more typical social interactions.

Self-repair is often characterised as being a response to noticing and correcting errors via a self-monitoring process (Levelt, 1983), and patients with schizophrenia are known to have difficulty monitoring their own behaviour (Johns et al., 2001). However, some self-repair is interactive, triggered by feedback from one's interlocutors or indicative of audience design (Goodwin, 1979). Self-repairs of this type may be an indicator of a person's engagement in a task, or need for clarity, for example, there are known to be more self-repairs from instruction givers in the Map Task (Colman and Healey, 2011) who have to describe a route carefully for a follower who does not have visual access to the route, but must draw it as accurately as possible on their own map. It is unclear whether one or both of these factors are responsible for the reduced levels of self-repair seen in patients.

Patients with schizophrenia are also known to display fewer hand gestures when speaking (Lavelle et al., 2013a), and have mismatched gesture use and speech (Millman et al., 2014). Furthermore, studies have identified that the presence of a patient with schizophrenia in an interaction influences the nonverbal behavior of their interacting partners, both in clinical contexts (Lavelle et al., 2015) and during first meetings with healthy controls, when the patient's diagnosis is undisclosed (Lavelle et al., 2013a; Lavelle et al., 2014). This suggests that patients' atypical patterns of participation in social interactions involve deficits in the interaction of verbal and non-verbal behaviours, with interaction itself playing a crucial role. We are therefore interested in investigating whether patients with schizophrenia compensate for verbal difficulties by using gesture in the same ways as healthy controls (Seyfeddinipur and Kita, 2014; Healey et al., 2013), and whether their interlocutors also modify their own verbal and non-verbal behaviours in interactions with patients.

This study aims to address the following questions.

1.1. Research questions

Compared to healthy control conversational groups and their healthy conversational partners:

1. Do patients with schizophrenia use less self-repair and gesture during conversation?
2. Is their use of self-repair associated with their use of gesture?

2. Methods

2.1. Participants

The data analysed in this study consists of transcripts and motion captured data of twenty patient interactions, involving one patient conversing with two healthy controls who were unaware of the patient's diagnosis, and twenty control interactions (with 3 healthy participants). Due to technical issues one patient interaction and one control conversation could not be transcribed and are excluded from the analysis.

Patients were taking anti-psychotic medication which fell within the low dose range (Chlorpromazine equivalents 50-200mg/day). Patients presenting with motor side effects from antipsychotic medication were excluded based on clinicians' assessment. Patients' symptoms were assessed using the Positive And Negative Symptom Scale for Schizophrenia (Kay et al., 1987).

Patients displayed relatively low PANSS scores for both positive symptoms ($M = 15.8; sd = 6.76$), which are additional features that occur with the onset of the disorder such as hallucinations or delusional beliefs, and negative symptoms ($M = 9.95; sd = 3.36$), which represent a reduction in usual function such as social withdrawal, diminished affect, apathy and anhedonia.

2.2. Ethics

All procedures were approved by a NHS Research Ethics Committee in the UK (07/H0711/90). All participants gave written informed consent and were free to withdraw at any time. Patients were recruited at routine psychiatric outpatient clinics under supervision of their psychiatrist, on the basis of a diagnosis of schizophrenia. 25% of all patients approached agreed to participate. Patients presenting with motor side effects from antipsychotic medication were excluded based on a clinician's assessment. Non-fluent English speakers were also excluded.

2.3. Procedure

Participants were brought into the laboratory in threes and seated in a triangular formation so that each participant had good visual access to each of the others (see Figure 1). The researcher read aloud a fictional moral dilemma, the 'balloon task' (see section 2.4. for details), which has been used for studying many aspects of dialogue, and is known to stimulate discussion (Howes et al., 2011). The group was provided with an opportunity to ask questions before the researcher left the interaction space and the task began. Interactions ended when participants reached a joint decision. Groups that failed to reach agreement had their interaction terminated at approximately 450 seconds (7 minutes 30 seconds).

All interactions were recorded in a human interaction laboratory fitted with an optical based Vicon motion-capture system, consisting of 12 infrared cameras and Vicon iQ software. Participants wore a top and a cap with 27 reflective markers attached. Cameras detected the markers at 60 frames per second, resulting in a highly accurate 3D representation of participants' movements over time (see Figures 1 and 2).

Figure 1: 2-dimensional image of participants engaged in triadic interaction, wearing the reflective markers

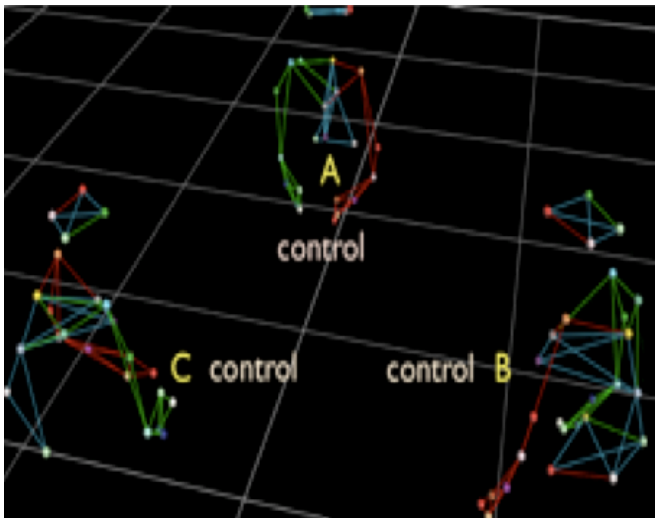


Figure 2: The wire frame representation of the interaction in 3-dimensional space

2.4. Task

The *balloon task* is an ethical dilemma requiring agreement on which of four passengers should be thrown out of a hot air balloon, which is losing height and about to crash into some mountains killing all on board unless one of them jumps to their certain death in order to save the other three. The four passengers are described to the participants as follows:

Dr. Robert Lewis - a cancer research scientist, who believes he is on the brink of discovering a cure for most common types of cancer.

Mrs. Susanne Harris - who is not only widely tipped as the first female MP for her area, but is also over the moon because she is 7 months pregnant with her second child.

Mr. William Harris - husband of Susanne, who he loves very much, is the pilot of the balloon, and the only one on board with balloon flying experience.

Miss Heather Sloan - a 14 year-old music prodigy, considered by many to be a "twenty first century Mozart".

Participants were instructed to debate the reasons for and against each person being saved, and reach mutual agreement about who should jump.

2.5. Analysis

2.5.1. Self-repair

Participants' speech was transcribed in ELAN. Self-repairs were annotated using STIR (STrongly Incremental Repair detection) (Hough and Purver, 2014); which automatically detects speech repairs on transcripts. STIR, which is trained on the Switchboard corpus (Godfrey et al., 1992) has previously been shown to be applicable to therapeutic dialogue, with high rates of correlation to human coders in terms of self-repair rate (Howes et al., 2014). The self-repair rate per word was calculated for each individual participant as the total number of self-repairs produced divided by the total number of words spoken.

2.5.2. Gesture

An index of gesture was derived from participants' hand movements using the 3D motion capture data. Gestures were identified as hand movement speeds greater than one standard deviation above an individual's mean hand movement speed thus giving a measure that was sensitive to individual variation in baseline hand movement (following (Lavelle et al., 2012)). The presence of gesture was assessed on a frame by frame basis and the percentage of frames spent gesturing was identified for each individual. This means we are looking at overall levels of hand movement (calculated for each individual), and not specific gestures or gesture types. This has the advantage of being calculable automatically from the motion capture data, but may also include movements that are not typically counted as gestures, such as brushing one's hair out of one's eyes.

3. Results and Discussion

3.1. Self-repair

	N	M (sd)	β	SE	χ^2	p
Patient	19	0.01 (0.01)	-0.02	0.004	12.59	<0.001
HP partner	38	0.02 (0.02)	-0.01	0.004	2.78	0.1
Controls	57	0.03 (0.02)				

Table 1: Repair rate

A mixed models regression analysis, adjusting for triadic group, age and gender, identified that healthy participants in the control groups used significantly more self-repair than schizophrenia patients (χ^2_1=12.59, 95% CI -0.02 to -0.01, p<0.001), as shown in Figure 3. The amount of self-repair produced by the healthy participants in the patient groups was numerically higher than that of their patient interlocutors and lower than that of the participants in the control groups, suggesting that there may be some modification of self-repair behaviour when interacting with a patient. However, neither of these differences were statistically significant (see Table 1), possibly due to lack of power in looking only at the mean figure per participant, and the variability of repair rates by person. Future work would investigate the levels of self-repair in a more fine-grained way, at the level of the utterance, which would allow us to look at these potentially relevant differences more precisely.

That patients repair their own speech less in a social interactive setting could be down to a number of factors, which cannot be decided between based on the current results. Deficits in both self-monitoring and audience design may be factors for patients. However, self-monitoring cannot explain the somewhat lower frequency of self-repair exhibited by patients' healthy partners, so is likely to be only part of the story. The possibility that patients' healthy partners have reduced self-repair (though not reaching significance in this study) could indicate that they are less engaged in the interaction – consistent with the finding that interacting with a patient (whilst unaware of their diagnosis) also affects subsequent ratings of rapport (Lavelle et al., 2014).

3.2. Gesture

Mixed models regression analyses, adjusting for triadic group age and gender, reveled that patient did not significantly differ from control participants in terms of their overall rates of gesture during the interaction (see Table 2). However, patients did use significantly fewer hand gestures when speaking (Table 3).

	N	M (sd)	β	SE	χ^2	p
Patient	19	7.2 (2.8)	0.21	0.84	0.06	0.8
HP partner	38	7.7 (3.0)	0.55	0.63	0.77	0.38
Controls	57	7.2 (3.0)				

Table 2: Overall gesture rate

	N	M (sd)	β	SE	χ^2	p
Patient	19	12.5 (2.8)	-5.84	2.91	4.01	0.05
HP partner	38	13.1 (3.0)	-3.29	1.99	2.78	0.1
Controls	57	16.5 (3.0)				

Table 3: Gesture rate while speaking

3.3. Gesture and self-repair

Partial correlations, adjusting for the amount of speech (see Figure 4), revealed that, in control group participants, increased self-repair was associated with increased overall gesture ($Rho_{48} = 0.33, p = 0.02$). In contrast, self-repair rates were not associated with gesture use in patients with schizophrenia ($Rho_{15} = -0.03, p = 0.91$), or their conversational partners ($Rho_{33} = -0.16, p = 0.40$).

These results indicate that in normal conversation between healthy participants, the amount of self-repair is positively correlated with gesture. Participants who are doing more repair, which may be due to discovering potential errors

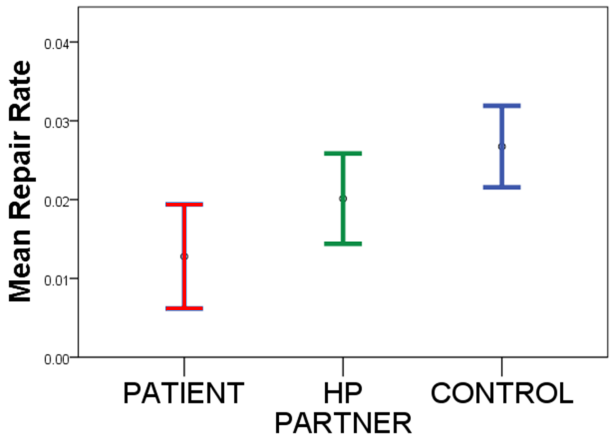


Figure 3: Mean repair rate per participant

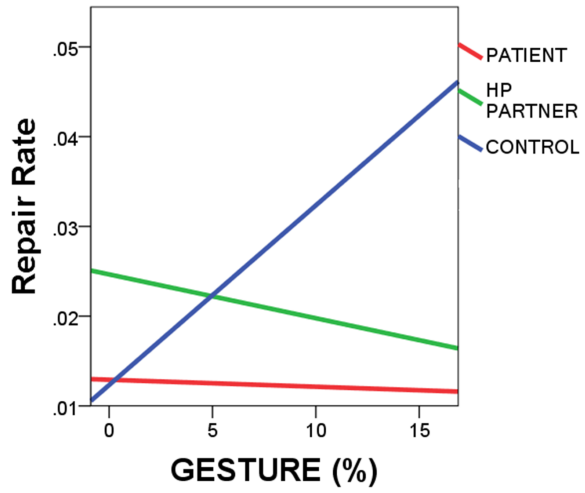


Figure 4: Correlation of mean repair rate and proportion of time spent gesturing

by self-monitoring or because they are tailoring their talk to their audience, are also utilising more multimodal resources in their interactions. Although this is at the level of the participant, so is a broad brush measure, it is consistent with previous findings (Seyfeddinipur and Kita, 2014; Healey et al., 2013). Contrarily, and in addition to the overall reduced levels of self-repair, there is no such relationship between self-repair levels and gesture in the dialogues including a patient. This holds both for patients, for whom a disconnect between communication modalities has been previously observed (Millman et al., 2014), but also, more surprisingly, also for their healthy interlocutors for whom no such disconnect would be expected.

4. Conclusions

During social interaction, schizophrenia patients repair their own speech less, and make less use of hand gesture when repair is required. In line with previous studies (Johns et al., 2001), these findings may reflect patients' difficulty monitoring their own behaviour. However, when self-repair does occur, patients are not employing other compensatory nonverbal modalities to assist with the difficulty. This may reflect a disconnect between communication modalities in patients with schizophrenia, however, this may not be entirely explained by impairments in self-monitoring, as patients' healthy interlocutors seem to also display reduced association between self-repair and gesture. Previous studies have identified that the degree of coordination between speech and nonverbal behaviour is impaired in schizophrenia. Furthermore this impairment is also visible in those interacting with the schizophrenia patient (Ellgring, 1986; Lavelle et al., 2013b). This suggests that the relationship between self-repair and gesture is also affected by elements of interaction, such as audience design or engagement, which may or may not also contribute to the difficulties displayed by patients.

Although the impact of patients symptoms were not explored in the current study previous findings suggest that they may have an influence on patients' gesture use (Lavelle et al., 2013a). This should be explored in studies with larger sample sizes where patients could be distinguished in terms of their symptom profiles.

Even though this is a very broad brush picture of the relationship between self-repair and gesture in patients with schizophrenia, as it is by participant over the whole conversation, this preliminary study indicates that combining automatically derivable data from transcripts and motion capture data offers a fruitful line of research in investigating the difficulties experienced by patients in social interaction. These automatic measures, while crude, do give an indication that these are areas in which patients' behaviours do not follow typical patterns which may be picked up on – if unconsciously – by their interlocutors, and contribute to the social exclusion experienced by patients. In future work, we will extend the existing study to look at the data at the level of the utterance, using cross-correlational techniques such as those in Healey et al (2013). This study also suggests looking more closely at both gesture and repair. In both cases, this may involve using more time intensive annotation methods to identify differences in the types of gesture and repair used (Colman and Healey, 2011; Healey et al., 2015), but the workload could be reduced by using automatic methods such as those outlined here to target particular utterances where the differences are apparent.

Overall, the ability to self-monitor and flexibly modify speech during conversation appears to be impaired in schizophrenia. This may make achieving mutual-understanding more difficult, contributing to the debilitating social deficits experienced by this patient group.

5. Acknowledgements

The data was collected as part of Lavelle's Ph.D. funded by the Engineering and Physical Sciences Research Council Doctoral Training Programme (EP/P502683/1).

Hough is supported by the Deutsche Forschungsgemeinschaft (DUEL project, grant SCHL 845/5-1) and the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University.

6. Bibliographical References

Addington, J. and Addington, D. (2008). Social and cognitive functioning in psychosis. *Schizophrenia research*, 99(1):176–181.

Brennan, S. and Schober, M. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

Caplan, R., Guthrie, D., and Komo, S. (1996). Conversational repair in schizophrenic and normal children. *Journal of the American Academy of Child & Adolescent Psychiatry*, 35(7):950 – 958.

Colman, M. and Healey, P. G. T. (2011). The distribution of repair in dialogue. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 1563–1568, Boston, MA.

Ellgring, H. (1986). Nonverbal expression of psychological states in psychiatric patients. *European archives of psychiatry and neurological sciences*, 236(1):31–34.

Godfrey, J. J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE ICASSP-92*, pages 517–520, San Francisco, CA.

Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas, editor, *Everyday Language: Studies in Ethnomethodology*, pages 97–121. Irvington Publishers, New York.

Healey, P. G. T., Lavelle, M., Howes, C., Battersby, S., and McCabe, R. (2013). How listeners respond to speaker's troubles. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, Berlin, July.

Healey, P. G. T., Plant, N., Howes, C., and Lavelle, M. (2015). When words fail: Collaborative gestures during clarification dialogues. In *2015 AAAI Spring Symposium Series: Turn-Taking and Coordination in Human-Machine Interaction*.

Hough, J. and Purver, M. (2014). Strongly incremental repair detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics.

Howes, C., Purver, M., Healey, P. G. T., Mills, G. J., and Gregoromichelaki, E. (2011). On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, 2(1):279–311.

Howes, C., Hough, J., Purver, M., and McCabe, R. (2014). Helping, I mean assessing psychiatric communication: An application of incremental self-repair detection. In *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialWatt)*, pages 80–89, Edinburgh.

Johns, L. C., Rossell, S., Frith, C., Ahmad, F., Hemsley, D., Kuipers, E., and McGuire, P. (2001). Verbal self-monitoring and auditory verbal hallucinations in patients with schizophrenia. *Psychological medicine*, 31(04):705–715.

Kay, S. R., Flszbein, A., and Opfer, L. A. (1987). The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, 13(2):261.

Lavelle, M., Healey, P. G. T., and McCabe, R. (2012). Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia Bulletin*.

Lavelle, M., Healey, P. G., and McCabe, R. (2013a). Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia bulletin*, 39(5):1150–1158.

Lavelle, M., Howes, C., Healey, P. G. T., and McCabe, R. (2013b). Speech and hand movement coordination in schizophrenia. In *Proceedings of the TiGeR Tilberg gesture research meeting*, Tilburg, June.

Lavelle, M., Healey, P. G., and McCabe, R. (2014). Participation during first social encounters in schizophrenia. *PloS one*, 9(1).

Lavelle, M., Dimic, S., Wildgrube, C., McCabe, R., and Priebe, S. (2015). Non-verbal communication in meetings of psychiatrists and patients with schizophrenia. *Acta Psychiatrica Scandinavica*, 131(3):197–205.

Leudar, I., Thomas, P., and Johnston, M. (1992). Self-repair in dialogues of schizophrenics: Effects of hallucinations and negative symptoms. *Brain and Language*, 43(3):487 – 511.

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

McCabe, R., Healey, P. G. T., Priebe, S., Lavelle, M., Dodwell, D., Laugharne, R., Snell, A., and Bremner, S. (2013). Shared understanding in psychiatrist-patient communication: Association with treatment adherence in schizophrenia. *Patient Education and Counselling*.

Millman, Z. B., Goss, J., Schiffman, J., Mejias, J., Gupta, T., and Mittal, V. A. (2014). Mismatch and lexical retrieval gestures are associated with visual information processing, verbal production, and symptomatology in youth at high risk for psychosis. *Schizophrenia Research*, 158(1-3):64 – 68.

Monte, R. C., Goulding, S. M., and Compton, M. T. (2008). Premorbid functioning of patients with first-episode nonaffective psychosis: a comparison of deterioration in academic and social performance, and clinical correlates of premorbid adjustment scale scores. *Schizophrenia research*, 104(1):206–213.

Schegloff, E., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Seyfeddinipur, M. and Kita, S. (2014). Gestures and self-monitoring in speech production. In *Annual Meeting of the Berkeley Linguistics Society*, volume 27, pages 457–464.

A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications

Spyridoula Varlokosta[1], Spyridoula Stamouli[1,2], Athanasios Karasimos[1,3], Georgios Markopoulos[1], Maria Kakavoulia[4], Michaela Nerantzini[1,5], Aikaterini Pantoula[1], Valantis Fyndanis[1,6], Alexandra Economou[1], Athanassios Protopapas[1]

[1] National and Kapodistrian University of Athens

[2] Institute for Language and Speech Processing / "Athena" Research Center

[3] Academy of Athens

[4] Panteion University of Social and Political Sciences

[5] Northwestern University

[6] University of Oslo

Address: University of Athens, Philology/Linguistics, Panepistimioupoli Zografou, Athens, 15784 Greece.

E-mail: svarlokosta@phil.uoa.gr, pstam@ilsp.gr, akarasimos@academyofathens.gr, gmarkop@phil.uoa.gr, markak@panteion.gr, nmixaela@gmail.com, aikaterini.pantoula@gmail.com, valantis.fyndanis@iln.uio.no, aoikono@psych.uoa.gr, aprotopapas@phs.uoa.gr

Abstract

In this paper, the process of designing an annotated Greek Corpus of Aphasic Discourse (GREECAD) is presented. Given that resources of this kind are quite limited, a major aim of the GREECAD was to provide a set of specifications which could serve as a methodological basis for the development of other relevant corpora, and, therefore, to contribute to the future research in this area. The GREECAD was developed with the following requirements: a) to include a rather homogeneous sample of Greek as spoken by individuals with aphasia; b) to document speech samples with rich metadata, which include demographic information, as well as detailed information on the patients' medical record and neuropsychological evaluation; c) to provide annotated speech samples, which encode information at the micro-linguistic (words, POS, grammatical errors, clause types, etc.) and discourse level (narrative structure elements, main events, evaluation devices, etc.). In terms of the design of the GREECAD, the basic requirements regarding data collection, metadata, transcription, and annotation procedures were set. The discourse samples were transcribed and annotated with the ELAN tool. To ensure accurate and consistent annotation, a Transcription and Annotation Guide was compiled, which includes detailed guidelines regarding all aspects of the transcription and annotation procedure.

Keywords: aphasia, aphasic discourse, annotated corpus

1. Introduction

Aphasia is defined as a language disorder following a focal damage to the left cerebral hemisphere caused either by a cerebral vascular accident (CVA), a traumatic brain injury (TBI), an infection, such as encephalitis, or as the result of the existence or the removal of a brain tumor (De Roo, 1999: 1; Mesulam, 2000: 296). Aphasia is typically restricted to language impairments in the absence of any other general cognitive impairment or dementia (Obler & Gjerlow, 1999: 38). Deficits in aphasia can potentially affect speech production and comprehension in both oral and written language forms, and at all linguistic levels (i.e., phonological, morphological, syntactic, and semantic), to varying degrees depending on the site and the severity of the brain injury (Harley, 2001: 23); from mild, to moderate and severe disorders.

Although in the aphasiological literature many different types of aphasia have been described, the most widespread classification identifies two basic categories: non-fluent aphasia or Broca's aphasia and fluent aphasia or Wernicke's aphasia, each one of which has been associated with different neurological characteristics, as for the locus and the extent of the lesion, and different linguistic characteristics.

Studies on speakers with aphasia conducted over the past 40 years have emphasized the clinical importance of the study of discourse production (e.g. Berko-Gleason et al., 1980; Nicholas & Brookshire, 1993; Olness & Ulatowska, 2011; Saffran, Berndt & Schwartz, 1989; Ulatowska, North & Macaluso-Haynes, 1981; Ulatowska et al., 1983; Vermeulen, Bastiaanse & van Wageningen, 1989; see Armstrong, 2000, for an overview of the literature). Since people with aphasia experience particular difficulties in their everyday communication, the study of their abilities at the discourse level is considered as a natural and objective method for assessing the communicative effectiveness of these individuals in their everyday life. More specifically, the study of discourse production can contribute to the diagnosis of the type of aphasia, to a more accurate identification of the communication impairments of patients, to the design of a more effective treatment as well as to the evaluation of patients' response to treatment (Wright, 2011).

Despite the fact that there is a large body of literature on the characteristics of aphasic discourse in many languages, which includes studies following different methodological approaches, theoretical frameworks, and analytical perspectives, there is a considerable lack of available resources to allow the systematic study of aphasic discourse in a comparable and replicable way across languages. The available corpora of aphasic discourse -constructed with the use of corpus linguistic techniques and providing systematic methods for the transcription, annotation, and analysis- are the Corpus of Dutch Aphasic Speech (CoDAS Westerhout & Monachesi, 2006), the Cambridge Cookie-Theft Corpus (Williams et al., 2010),

and the AphasiaBank (MacWhinney et al., 2011, 2012). Each one of them has contributed from a different perspective to the process of enriching the existing methods and data for the study of aphasic discourse, and, consequently, to the advancement of research in this area. CoDAS comprises a pilot study of six aphasic speakers with two levels of annotation, an orthographic-phonetic transcription and a Part-Of-Speech (POS) tagging. The Cambridge Cookie-Theft Corpus contains transcriptions of spontaneous speech and single-picture descriptions elicited with the cookie-theft picture. The study includes data from approximately 87 brain-damaged patients in comparison to a group of 227 healthy individuals. A total of 1331 utterances are time-stamped and annotated on the phonological level following an XML-based TEI schema. AphasiaBank is a multimedia database with video and speech annotated transcriptions of approximately 180 speakers with aphasia and 140 non brain-damaged controls in a variety of communicative tasks and interactions. The transcriptions are based on the CHAT format and coded for analysis with specific CLAN programs. A multi-level annotation produces a language profile that includes word-level and utterance-level morphosyntactic errors.

The Greek Corpus of Aphasic Discourse (GREECAD) is the outcome of a research action under the large scale multidisciplinary project "THALES-Levels of impairment in Greek aphasia: relationship with processing deficits, brain region, and therapeutic implications". The aim of this research action was the collection, annotation, documentation, and linguistic analysis of spoken discourse of Greek speakers with aphasia.

The development of the GREECAD had to meet the following requirements: a) to include a rather homogeneous sample of Greek as spoken by individuals with mild non-fluent aphasia; b) to document speech samples with rich metadata, which include demographic information, information on the patients' medical record, as well as their speech and language therapy and neuropsychological evaluation; c) to provide annotated speech samples which encode properties of speech as well as linguistic information at the micro-linguistic (words, POS, grammatical, semantic, and phonological errors, clause types, etc.) and discourse level (narrative structure units, main events, evaluation devices, etc.).

In this paper, the process of designing the GREECAD is presented, regarding data collection, transcription, and annotation of speech samples. Given that resources of this kind are quite limited, a major aim of the development of the GREECAD was to provide a set of specifications which could serve as a methodological basis for the development of other relevant corpora, and, therefore, to contribute to future research in this area.

2. Data Collection

Among the various discourse types that have been studied, narrative discourse has attracted more attention in aphasia research, mainly because the abstract narrative schema provides an objective framework for the analysis of speakers' productions and their comparison to healthy controls. Therefore, a protocol of four narrative tasks (Kakavoulia et al., 2014) was developed to elicit spoken discourse samples from Greek speakers with aphasia. Previous research (Doyle et al., 1998) shows that the discourse produced by speakers with aphasia is influenced by the characteristics of elicitation tasks, such as the type of stimuli and the modality of presentation, as well as by the cognitive and linguistic requirements of the tasks, depending on the particular clinical characteristics of each individual. Thus, it was decided that the protocol should include different narrative genres (personal narrative, third person narrative, fairy tale, etc.) and different elicitation techniques (McNeil et al., 2007; Menn, Ramsberger & Helm-Estabrooks, 1994, Nicholas & Brookshire, 1995; Ulatowska et al., 1983), which provide different degrees and types of support to the participants in order to compensate for the cognitive and linguistic demands of each task. Personal narratives were chosen because they elicit more natural speech data characterized by extensive use of evaluative devices (Ulatowska et al., 2006). More constrained elicitation tasks, such as story retelling and picture elicitation, were also employed to ensure more controlled discourse samples. More specifically, the protocol includes the following tasks:

Task 1: Unaided production of a personal narrative ("stroke story"). The individuals with aphasia narrate the incident of their stroke story, while the control group (people who have suffered a heart attack, see Section 3, Participants) narrate the heart attack incident.

Task 2: Production of an unknown story based on a 6-picture series ("the party"). The participant narrates a short, simple story shown in the pictures presented to her/him by the researcher. Linguistic demands are high, since the participant has to generate the story events and the narrative structure from the pictures, but there are no memory requirements.

Task 3: Retelling of an unknown story, aided by a 5-picture series ("the ring"). The participant listens to a recorded story, which has the structure of a traditional fairy tale. The story is quite lengthy; it has many episodes and a complex plot, characteristics which increase the linguistic and cognitive demands of the task. At the same time, five pictures depicting important events of the story are presented to her/him. After listening to the story, the participant has to retell it using the pictures. Visual support is expected to compensate for the increased linguistic and cognitive demands of the task.

Task 4: Familiar story retelling ("hare and tortoise" Aesop's fable). The participant listens to a recorded narration of the fable and afterwards she/he has to retell the story to the researcher. No visual support is used. Memory load is increased, since the participant has to retain story elements and their temporal order. However, this demand is compensated by the fact that the story is already familiar to the speaker.

3. Participants

The GREECAD contains spoken discourse samples

elicited from Greek-speaking individuals with mild non-fluent aphasia and controls matched for age and level of education to the aphasic speakers (Table 1).

	N	Age range (years)	Sex		Education	
			M	F	Years	N
Speakers with aphasia	18	39-67	15	3	6	3
					9	1
					12	5
					over 12	9
Control group	7	43-71	7	0	9	2
					12	2
					over 12	3

Table 1: Characteristics of participants

The control group comprises individuals who have suffered a heart attack. The choice of these individuals as a control group was made to ensure the comparability of the personal narrative samples (Task 1) in terms of their textual characteristics. Therefore, in accordance to the stroke story production by speakers with aphasia, the control group participants narrated their "heart attack story". Heart attack is a similar traumatic experience to the stroke, with a comparable informational content and event sequence (initial symptoms, reaction from the part of the patient and relatives, medical diagnosis and intervention, outcomes). Although a few differences in the vocabulary were expected between the two versions of the personal narrative, mainly with respect to specific symptoms or medical treatment, their overall linguistic, structural and informational similarities were considered as more useful for comparison between the two groups.

Ethical approval was obtained by the Ethics Committee of the hospitals, medical and rehabilitation centres involved in the project. Patients received written information about the study and were asked to provide full informed consent.

4. The GREECAD Corpus

The spoken discourse samples of speakers with aphasia and those of the control group were manually transcribed and annotated. The result of this process was the compilation of the GREECAD. Speakers with aphasia are currently represented in the corpus with 72 transcripts, while the control group with 28 transcripts. Table 2 shows the total number (N) of transcripts, tokens and clauses per group. Specific measurements on tokens and clauses, besides total count, include the statistical mean, as well as the minimum and maximum value per speaker in the corresponding group. The corpus is still being enriched with new data collected from individuals with aphasia and controls.

The discourse samples are documented with rich metadata which include demographic information about the participants, as well as detailed information on the patients' medical record, including the type of aphasia, and their speech and language therapy and neuropsychological evaluation (e.g. their scores on the Boston Diagnostic Aphasia Examination, Greek version: Papathanassiou et al., 2008, and the Boston Naming Test, Greek version: Simos, Kasselimis & Mouzaki, 2011).

	Transcripts	Tokens				Clauses			
	N	N	Mean	Min	Max	N	Mean	Min	Max
Speakers with aphasia	72	4643	64.5	14	117	1158	16.1	4	36
Control group	28	4006	143.1	68	208	871	31.1	12	56
Total	100	8649				2039			

Table 2: The GREECAD Corpus

5. Transcription and Annotation

Discourse samples were manually transcribed and annotated using the ELAN transcription and annotation tool (Wittenburg et al., 2006). Transcription was orthographic using the Greek alphabet. A transcription protocol was designed to encode the necessary information for linguistic analysis, excluding detailed phonological information. Specific conventions were used for unintelligible words and neologisms, while no special symbols were used. The transcripts were time-aligned with the audio files at utterance, clause, and word level.

Annotation was carried out to encode linguistic information of the patients' discourse at various levels. A structured, multi-tiered annotation scheme was designed in order to include all the parameters of spoken discourse under investigation. These parameters include speech and non-speech events (e.g. vowel and consonant lengthening, pauses, filled gaps, laughter, etc.), micro-linguistic features (words, POS, grammatical, semantic, and phonological errors, clause types, etc.), as well as discourse features (narrative structure units, main events, evaluation devices). The annotated corpus is available in XML / EAF format, which allows the future analysis of data with automatic computational linguistic techniques. It was based on the Formal Framework for Linguistic Annotations (Bird & Liberman, 1999; Ide & Suderman, 2007, 2014) and the template is governed by token-based, type-based, and graph-based hierarchy.

For ensuring accurate and consistent transcription and annotation, a set of explicit and clear procedures was established, together with detailed guidelines for the annotators, which comprised a Transcription and Annotation Guide (Varlokosta et al., 2013). The annotators were graduate or postgraduate students of Linguistics. They were divided into small groups of 2-3 annotators. Each group annotated only certain tiers in each narrative, according to their specialisation, experience and interests, and not all tiers. Annotation tiers were assigned to groups as follows:
Group 1: Transcription (Researcher – Patient)
Group 2: Processed Transcription – Utterances (limits) – Clauses (limits)
Group 3: Events
Group 4: Clauses (tagging: types, grammaticality, completeness)
Group 5: Words (limits, POS tagging, counting)
Group 6: Errors (tagging: phonological, morphosyntactic, lexical/semantic errors, paraphrases)
Group 7: Reformulations
Group 8: Narrative annotation (narrative structure elements, main events, evaluation devices)
It should be noted that in some cases there was a single annotator in each group (e.g. group 1, 2, 3, 7). A two-person leading team was appointed to train and coordinate the groups of annotators. The annotation leaders were experts in Corpus Linguistics, experienced in data collection and processing with the use of the ELAN tool. This team trained each individual group in the annotation of the specific tiers they were assigned to. After training, a pilot phase was carried out, including two phases: a) initially, annotators were given a file in which they annotated their tiers in collaboration with one of the trainers, who helped them and resolved any query on the spot; b) subsequently, annotators annotated another file on their own, using the Transcription and Annotation Guide. Their annotations were checked by their trainer, who gave feedback regarding problematic issues. Phase b was repeated as many times as needed to ensure agreement of an acceptable level between the annotator and the trainer (above 90%). It should be noted that some annotations were more difficult than others (e.g. setting the utterance limits or tagging error types at word level), which led to more repetitions of phase b until the annotator and the trainer reached an agreement. This procedure highlighted the need for more explicit and detailed criteria for annotating these particularly difficult tiers. The pilot phase was carried out with each new annotator who entered a group. Before marking each file as "complete", a checking phase was carried out, during which all annotations were checked by the leading team, who made the necessary corrections. Each member of the leading team was responsible for checking specific tiers. During the checking phase, the leading team provided feedback to the annotators, including new guidelines, if needed. All the new instructions and modifications regarding the annotation scheme, the guidelines, as well as specific annotation criteria that came up during the pilot and the checking phase were integrated into the Transcription and Annotation Guide. Moreover, during file processing, the annotators were in direct and constant contact with the leading team for questions and instructions. Finally, it is worth noting that most of the times the members of each group were working together, as a team, and not individually.

5.1 Annotation Scheme

Figure 1 shows tier dependencies of the multi-tiered annotation scheme:

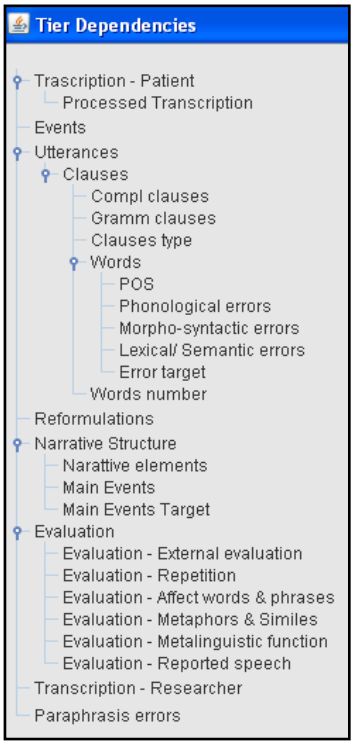


Figure 1: Tier dependencies of the annotation scheme

A set of detailed criteria regarding accurate transcription, definition of speech segment boundaries (utterances, clauses, and words), identification of each annotation category, and assignment of a valid value at each one were provided to the annotators. In the following sections, a brief description of the main annotation tiers is provided.

5.1.1. Patient transcription

This group includes two transcription tiers: the primary or "rough" transcription (parent tier) and the secondary, processed transcription (child tier). The first one contains anything which has been uttered by the participant, orthographically transcribed in the Greek alphabet. Processed transcription is the result of "cleaning up" the primary transcription of: a) repetitions: all but the final occurrence of a repeated word, phrase or segment were eliminated, excluding repetition for emphasis, b) self-corrections, c) formulaic phrases, d) one-word replies, e) parts of discourse irrelevant to the narrative content. Processed transcription provides the basis for the linguistic measurements of participants' discourse (number of utterances, clauses, words), as well as for the measurements of verbal flow (words / minute), verbal disruption, syntactic complexity, and narrative macrostructure.

5.1.2. Speech and non-speech events

In this tier the events of spoken discourse are annotated, such as vowel and consonant lengthening, silence, pauses (longer than 0.5 sec), noise, filled gaps, etc. This tier refers to the primary, "rough" transcription tier.

5.1.3. Reformulations

This tier contains: a) self-corrections at word level (e.g. "i podherta tis i p i mitria tis eklepse to dhaxtilidhi", transl. "her (neologism targeting the word 'stepmother') her p her stepmother stole the ring"), which can be phonological, lexical or morphological; b) repetitions (e.g. "pire pire to to dhepidhoni", transl. "he got he got the the (neologism targeting the word 'ring')"). Repetitions which are used for emphasis and serve an evaluative function in the narrative are not tagged as reformulations (e.g. "pias' tin Eleni pias' tin" transl. "catch Helen catch her").

5.1.4. Utterances

This tier includes two child tier groups: clauses and words. The term "utterance" is used as equivalent to the term "sentence", adopting the dominant view in linguistics that a sentence can consist of one or more clauses. The terms "utterance" and "sentence" are often used interchangeably in aphasia research (Faroqi-Shah & Thompson, 2007; Fyndanis, Varlokosta & Tsapkini, 2012; Wang, Yoshida & Thompson, 2014), following mainly Saffran, Berndt and Schwartz (1989: 471), who identify a set of certain structural types of sentences as utterances. Following Thompson et al. (1995), we used a combination of prosodic and semantic criteria to determine utterance boundaries. Utterance is defined as the speech section which follows and precedes silence, coincides with an intonational curve, and corresponds to a coherent meaningful unit of discourse. In cases where the aphasic speech was so fragmented that an intonational curve was difficult to identify, semantic criteria (coherence and completeness of meaning) were mainly used to define the utterance boundaries.

The tier of clauses is a child tier to the one of utterances. The presence of a verb was used to determine a clause. However, it should be noted that in aphasia verbs are prone to omission. Therefore, the presence of an overt subject or object (or of both) could also be employed as a sufficient criterion to identify a clause. Each clause was further annotated as:

a) complete or incomplete: an incomplete clause is a clause that lacks some basic arguments or is abandoned before its meaning is completed. For example, the clause "ksafnika kapu pidhi egho... ox thee mu" (transl. "suddenly somewhere because I... oh my god"), was annotated as incomplete.
b) grammatical or ungrammatical: ungrammatical is a clause which contains grammatical errors at word level or lacks basic arguments. For example, the clause "pefto to aftocinito" (transl. "I fall the car") was annotated as ungrammatical, due to the omission of the preposition "apo" (transl. "from").

Clauses were further annotated for their type. Values for clause types include independent clause, elliptical clause, as well as all types of subordinate clauses (clauses of time, cause, result, purpose, conditional, relative, etc.) and verb complements.

The tier of words is a child tier to the one of clauses. Number of words in each clause is indicated and each word is further annotated with respect to its POS and to the phonological, morphosyntactic, and lexical/semantic errors it might contain.

Phonological errors are errors of phoneme omission (e.g. "cek" instead of "ce*i*k", transl. "cake"), substitution (e.g. "*j*elona" instead of "*ch*elona", transl. "tortoise"), addition (e.g. "s*e*trono" instead of "strono", transl. "spread"), etc.

Morphosyntactic errors are errors of omission (e.g. "laghos" instead of "*o* laghos", transl. "hare" instead of "the hare") or substitution (e.g. "theli *pu* pai" instead of "theli *na* pai", transl. "wants that go" instead of "wants to go") of free morphemes, as well as errors of agreement, such as number, case, and gender agreement between article and noun (e.g. "*ton* (def art masc) dhaxtilidhi (N neut)" instead of "*to* (def art neut) dhaxtilidhi", transl. "the ring"), incorrect choice of aspect (e.g. "tre*k*si" (pfv asp) instead of "tre*ch*i" (ipvf asp), transl. "runs"), tense (e.g. "pigha", transl. "I went" instead of "pijeno", transl. "I go"), case (e.g. "*to* ipe" (clit pro acc) instead of "*tu* ipe" (clit pro gen), transl. "told him"), etc.

Lexical/semantic errors include cases such as: a) neologisms, either retaining the morpho-phonological structure of Greek words (e.g. "dheklidhoni" instead of "dhaxtilidhi", transl. "ring"), or not retaining it, so the word's grammatical category is unspecified (e.g. "idhesofoliberi", target word: unknown); b) production of words which have a phonological (e.g. "sidhora" instead of "simera", transl. "today") or semantic (e.g. "aschimi" instead of "omorfi", transl. "ugly" instead of "pretty") relationship with the target word.

Regarding word counting, the criteria proposed by Nicholas and Brookshire (1993) were followed: to be counted as words, lexical items have to be intelligible in context but not necessarily complete, accurate and relevant to the story. For example, the word "pipidhizis" which is a neologism (target word: "titivizis", transl. "you chirp") was counted as a word, even not phonologically accurate, while the word segment "che" (probably targeting the word "chelona", transl. "tortoise") was not counted as a word.

5.1.5. Narrative structure

This group of annotation tiers refers to the analysis of narrative discourse at the level of macrostructure. More specifically, it includes: a) a tier where the components of narrative structure are annotated ("narrative elements") and b) a tier where the main informational units of discourse, the story's "main events" are annotated ("main events").

The structural components of the elicited narratives are annotated on the basis of the Labovian model of narrative structure (Labov, 1972; Labov & Waletzky, 1967), which includes the following structural units:

a) Abstract: A single or multi-clause unit which informs the addressee on what the story is about (e.g. "theli na mas pi to paramithi ti ti chelona me to lagho", transl. "the fairy tale wants to tell us (about) the tortoise with the hare").

b) Orientation: The setting of the story, informing the addressee on the main characters (who), the place (where), and the time (when) of the story (e.g. "lipon i chelona ena proino itan sto dh dhasos", transl. "well, in the morning the tortoise was in the woods").
c) Complication: A sequence of events describing a 'problem', an unexpected complication for one or more of the main characters, his/their response(s) to the problem and his/their plan of action and attempts to resolve it. The events are leading to the climax or high point of the narrative.
d) Resolution: The part of the narrative which describes the outcome of the character(s)' attempts to resolve the problem, leading to the narrative's closure (e.g. "i chelona itane sto dhelos. ce itane medh me echi nicisi. ce exase o loghos o laghos", transl. "the hare was at the end. and she was with, she has won. and the hare lost.").
e) Coda: A unit linking the narrative to the present time (e.g. "And they lived happily ever after").
Furthermore, for measuring the stories' informational content, the number of "main events" was used as an indicator (Capilouto, Wright & Wagowich, 2006; Wright et al., 2005). Main events are defined as single or multi-clause units of a story, each one referring to a significant event of the story, which, at the same time, is independent of the other story events. Main events usually include one or more associated events and the temporal and causal relationships between them. Stories of tasks 2, 3 and 4 had a predefined number of main events. For example, the main events of the "hare and tortoise" story are the following:
1. The hare is going for a walk in the woods looking for food.
2. He meets the tortoise and thinks her slow walking is very funny.
3. The hare laughs at the tortoise and she challenges him to a race.
4. The hare finds her proposal very funny but he accepts the challenge.
5. They appoint the fox, the smartest animal, as the referee, and the race begins the following morning, when all the animals are gathered to watch it.
6. The hare decides to take a nap because he is confident that he can cover the distance from the tortoise very easily as soon as he gets up.
7. The tortoise keeps walking.
8. The hare sleeps for a little longer and when he wakes up he starts running.
9. He finds it strange not to see the tortoise anywhere but he thinks she gave up the race.
10. When the hare reaches the finishing line, he sees that the tortoise has finished first.
In order to be tagged as a main event, a textual unit should include the respective event or sequence of associated events as well as the relationship between them. For example, the following part of a story produced by a speaker with aphasia was tagged as main event no 1: "vjice o jo laghos ce vj vj vji vji vjice o laghos ce zjicise to xajito tou" (transl. "the hare went out and looked for his food.").

5.1.6. Evaluation

The evaluation tier is not embedded in the narrative structure group of tiers, since evaluative devices might cross the boundaries of stories' structural components. The category of evaluation includes linguistic devices which indicate the emotional and cognitive status of the narrator and his attitude towards the story events and characters. Evaluation expresses the narrator's involvement in the narrative and constitutes a second narrative layer, which transforms a simple sequence of events to a worth-telling story. In line with previous studies on evaluation of aphasic discourse (Armstrong, 2005; Armstrong & Ulatowska, 2007, 2006; Ulatowska et al., 2006, 2011), the following evaluative features are annotated:
a) external evaluation: narrator's comments, sometimes directly addressed to story recipient (e.g. "katalavenis?", transl. "do you understand?")
b) repetition of words or phrases for emphasis (e.g. "posa atoma itane, para pola itane", transl. "so many people were there, a lot of people were there")
c) words and phrases indicating emotional state, as well as inherently evaluative lexical items (e.g. "iche nevriasi", transl. "he was upset", "eftixos", transl. "fortunately")
d) metaphors and similes (e.g. "san salighari", transl. "like a snail")
e) metalinguistic function: narrator's comments on his own speech (e.g. "edho itane... *dhe thimame*", transl. "here there was... *I don't remember*", "laghos fajito chelona... *pos to len...* to...", transl. "(the) hare (was looking for) food (the) tortoise... *how is it called...* the...")
f) reported speech (direct and indirect) (e.g. "tha pas ghrighora ston aghona? tha pao ghrighora", transl. "are you going to run fast at the race? I will run fast").

6. Conclusion

The GREECAD is the first systematic attempt to develop an annotated corpus of aphasic discourse for Greek. The annotated transcripts of individuals with aphasia and healthy controls included in the current version of the corpus are being analyzed in terms of a set of measures, such as: a) verbal production and verbal flow (number of utterances, sentences and words, MLU, words/minute), b) syntactic complexity and grammaticality (number of conjunctions/total number of words, number of grammatical clauses/total number of clauses, number of subordinate clauses/total number of clauses, noun/verb ratio, number of errors/total number of words, etc.), c) verbal disruption: self-corrections, repetitions, abandoned clauses, gap-fillers, formulaic expressions, d) narrative structure (number of main events, narrative structure units, number of clauses/unit, evaluative devices by category, etc.). The main aims of the current studies being conducted or future ones are: a) to identify specific impairments at grammatical, lexical, and discourse level in the speech production of Greek-speaking individuals with aphasia, which could contribute to more effective evaluation, treatment, and assessment of treatment outcomes; b) to evaluate the overall communication abilities of speakers

with aphasia in their everyday lives, using narrative discourse as an objective communicative condition. Initial findings show differences in the group of individuals with aphasia compared to the control group regarding verbal production and flow, verbal disruption, grammatical accuracy, and syntactic complexity. However, narrative measures show a relative preservation of communication skills at the discourse level, since speakers with aphasia are able to produce the main informational content of a narrative and retain the main elements of narrative macrostructure despite their impairments at the microlinguistic level (Stamouli & Karasimos, 2015).

It is worth noting that the annotation scheme designed for the development of the GREECAD has been proven functional, flexible and broad, allowing the extended linguistic annotation of discourse samples in a consistent way.

An unrestricted online version of the GREECAD is not yet available. However, as soon as the ongoing studies of the research team are completed, free access to the fully annotated corpus with the XML metadata will be provided for research purposes and the corpus will be shared as a language specific resource to the META-SHARE [1] open-source repository.

The availability of the annotated aphasic discourse samples to the research community in combination with the rich metadata that accompany them, are expected to increase interest in the linguistic study of aphasia in Greek and to support the interdisciplinary study of aphasia, thereby contributing to a deeper and broader investigation of the complex phenomenon of aphasia.

7. Acknowledgements

This research action was co-financed by the European Union (European Social Fund) and Greek national funds through the operational program "Education and Lifelong Learning" of the National Strategic Reference Framework, research program THALES-UoA "Levels of impairment in Greek aphasia: Relationship with processing deficits, brain region, and therapeutic implications" (PI: Spyridoula Varlokosta).

We are grateful to Constantin Potagas, Ilias Papathanasiou, Georgia Kolintza, and Ioannis Evdokimidis for patient referral, to Margarita Atsidakou, Eva Efstratiadou, Anastasia Archonti, Ariadni Chatzidaki, Christoforos Routsis, Lina Chatziantoniou, and Dimitrios Kasselimis for clinical testing, to Athina Kontostavlaki for the development of the project Database, to Eleni Konsolaki for the coordination of the Annotators Team, to Aikaterini Pantoula, Pola Drakopoulou, and Sofia Apostolopoulou for the data collection, and to Christina Alexandri, Sofia Apostolopoulou, Dimitra Arfani, Kelly Atsali, Alexandros Bakogianis, Sofia Banou, Konstantina Daraviga, Vassiliki Dechounioti, Evi Doli, Pola Drakopoulou, Martha Drouga, Popi Eldahan-Apergi, Spyridoula Gasteratou, Mariana Georgouli, Aliki Gotsi, Nasia Gouma, Konstantina Goni, Chara Gourgouleti, Dimitris Katsimpokis, Giannis Kritikos, Alexandra Kostaletou, Panayotis Kounoudis, Anna Michalaki, Sofia Pagoni, Panagiotis Panagopoulos, Kalliopi Papadodima, Aikaterini Pantoula, Stamatina Remoundou, Maria Saridaki, Alexandros Tsaboukas, and Ilias Valaskatzis for their valuable work in the transcription and annotation of the speech samples. Last, we would like to thank the six anonymous reviewers for their constructive comments, which helped us improve the paper.

[1] URL: http://www.meta-share.eu

8. Bibliographical References

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology*, 14, pp. 875-892.

Armstrong, E. & Ulatowska, H. K. (2007). Making stories: Evaluative language and the aphasia experience. *Aphasiology*, 21, pp. 763-774.

Armstrong, E. & Ulatowska, H. K. (2006). Stroke stories: Conveying emotive experiences in aphasia. In M. J. Ball & J. S. Damico (Eds.), *Clinical Aphasiology: Future Directions*. Hove, UK: Psychology Press.

Armstrong, E. (2005). Expressing opinions and feelings in aphasia: Linguistic options. *Aphasiology*, 19, pp. 285-296.

Berko-Gleason, J., Goodglass, H., Obler, L., Green, E., Hyde, M. & Weintraub, S. (1980). Narrative strategies of aphasics and normal-speaking subjects. *Journal of Speech and Hearing Research*, 23, pp. 370-382.

Bird, S. & Liberman, M. (1999). A Formal Framework for Linguistic Annotation. Technical Report (MS-CIS-99-01). University of Pennsylvania.

Capilouto, G. J., Wright, H. H. & Wagovich, S. A. (2006). Reliability of main event measurement in the discourse of individuals with aphasia. *Aphasiology*, 20, pp. 205-216.

De Roo, E. (1999). *Agrammatic Grammar: Functional Categories in Agrammatic Speech*. Hague: Theseus.

Doyle, P. J., McNeil, M. R., Spencer, K. A., Goda, A. J., Cottrell, K. & Lustig, A. P. (1998). The effects of concurrent picture presentations on retelling of orally presented stories by adults with aphasia. *Aphasiology*, 12, pp. 561-574.

Faroqi-Shah, Y. & Thompson, C. K. (2007). Verb inflections in agrammatic aphasia: Encoding of tense features. *Journal of Memory and Language*, 56, pp. 129-151.

Fyndanis, V., Varlokosta, S., & Tsapkini, K. (2012). Agrammatic production: Interpretable features and selective impairment in verb inflection. *Lingua*, 122, pp. 1134-1147.

Harley, T. (2001). *The Psychology of Language: From Data to Theory* (2nd edition). New York: Psychology Press.

Ide, N. & Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*. Stroudsburg, PA: Association for Computational Linguistics, pp. 1-8.

Ide, N. & Suderman, K. (2014). The linguistic annotation framework: A standard for annotation interchange and merging. *Language Resources and Evaluation*, 48, pp.

395-418.

Kakavoulia, M., Stamouli, S., Foka-Kavalieraki, P., Economou, A., Protopapas, A. & Varlokosta, S. (2014). A battery for eliciting narrative discourse by Greek speakers with aphasia: Principles, methodological issues, and preliminary results [in Greek]. *Glossologia*, 22, pp. 41-60.

Labov, W. (1972). *Language in the Inner City*. Philadelphia: The University of Pennsylvania Press.

Labov, W. & Waletsky, J. (1967). Narrative analysis. In J. Helm (Ed.), *Essays in the Verbal and Visual Arts*. Seattle: University of Seattle Press, pp. 12-44.

MacWhinney, B., Fromm, D., Forbes, M. & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, 25, pp. 1286-1307.

MacWhinney, B., Fromm, D., Holland, A. & Forbes, M. (2012). AphasiaBank: Data and methods. In N. Mueller & M. Ball (Eds.), *Methods in Clinical Linguistics*. New York: Wiley, pp. 31-48.

McNeil, M. R., Sung, J. E., Yang, D., Pratt, S. R., Fossett, T. R. D., Pavelko, S. & Doyle, P. J. (2007). Comparing connected language elicitation procedures in person with aphasia: Concurrent validation of the Story Retell Procedure. *Aphasiology*, 21, pp. 775-790.

Menn, L., Ramsberger, G. & Helm-Estabrooks, N. (1994). A linguistic communication measure for aphasic narratives. *Aphasiology*, 8, pp. 315-342.

Mesulam, M. M. (2000). *Principles of Behavioral and Cognitive Neurology* (2nd edition). New York: Oxford University Press.

Nicholas, L. E. & Brookshire, R. H. (1995). Presence, completeness and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech and Hearing Research*, 38, pp. 145-156.

Nicholas, L. E. & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research*, 36, pp. 338-350.

Obler, L. K. & Gjerlow, K. (1999). *Language and the Brain*. Cambridge: Cambridge University Press.

Olness, G. S. & Ulatowska, H. K. (2011). Personal narratives in aphasia: Coherence in the context of use. *Aphasiology*, 25, pp. 1393-1413.

Papathanassiou, E., Papadimitriou, D., Gavrilou, V. & Michou, A. (2008). Normative data for the Boston Diagnostic Aphasia Battery in Greek: Gender and age effects [in Greek]. *Psychology*, 15, pp. 398-410.

Saffran, E. M., Sloan-Berndt, R. & Schwartz, M. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, 37, pp. 440-479.

Simos, P.G., Kasselimis, D. & Mouzaki, A. (2011). Age, gender, and education effects on vocabulary measures in Greek. *Aphasiology*, 25, pp. 492-504.

Stamouli, S. & Karasimos, A. (2015). The Greek Corpus of Aphasic Discourse. Oral presentation at the workshop on the "Interdisciplinary Study of Aphasia" Thales Project. University of Athens. Athens, 27 June 2015.

Thompson, C. K., Shapiro, L. P., Tait, M. E., Jacobs, B. J., Schneider, S. L. & Ballard, K. J. (1995). A system for the linguistic analysis of agrammatic language production. *Brain and Language*, 51, pp. 124-127.

Ulatowska, H. K., Freedman-Stern, R., Doyel, A. W., Macaluso-Haynes, S. & North.A. (1983). Production of narrative discourse in aphasia. *Brain and Language*, 19, pp. 317-334.

Ulatowska, H. K., North, A. J. & Macaluso-Haynes, S. (1981). Production of narrative and procedural discourse in aphasia. *Brain and Language*, 13, pp. 345-371.

Ulatowska, H. K., Olness, G. S., Keebler, M. & Tillery, J. (2006). Evaluation in stroke narratives: A study in aphasia. *Brain and Language*, Special Issue Academy of Aphasia 2006 Program, 99 (1-2), pp. 51-52.

Ulatowska, H. K., Reyes, B. A., Santos, T. O. & Worle, C. (2011). Stroke narratives in aphasia: The role of reported speech, *Aphasiology*, 25, pp. 93-105.

Varlokosta, S., Karasimos, A., Stamouli, S., Kakavoulia, M., Markopoulos, G., Goutsos, D., Fyndanis, V., Neranztini, M. & Pantoula, A. (2013). Greek Corpus of Aphasic Discourse. Transcription and Annotation Guide [in Greek]. Athens: National and Kapodistrian University of Athens.

Vermeulen, J., Bastiaanse, R. & van Wageningen, B. 1989. Spontaneous speech in aphasia: A correlational study. *Brain and Language*, 36, pp. 252-274.

Wang, H., Yoshida, M. & Thompson, C. K. (2014). Parallel functional category deficits in clauses and nominal phrases: The case of English agrammatism. *Journal of Neurolinguistics*, 27, pp. 75-102

Westerhout, E. & Monachesi, P. (2006). A pilot study for a Corpus of Dutch Aphasic Speech (CoDAS). In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), pp. 1648-1653.

Williams, C., Thwaites, A., Buttery, P., Geertzen, J., Randall, B., Shafto, M., Devereux, B. & Tyler, L. (2010). The Cambridge Cookie-Theft Corpus: A corpus of directed and spontaneous speech of brain-damaged patients and healthy individuals. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), pp. 2824-2830.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation, pp. 1556-1559. (URL: http://tla.mpi.nl/tools/tla-tools/elan/, Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands).

Wright, H. H. (2011). Discourse in aphasia: An introduction to current research and future directions. *Aphasiology*, 25, pp. 1283-1285.

Wright, H. H., Capilouto, G. J., Wagovich, S. A., Cranfill, T. & Davis, J. (2005). Development and reliability of a quantitative measure of adults' narratives. *Aphasiology*, 19, pp. 263-273.

Russian CliPS: a Corpus of Narratives by Brain-Damaged Individuals

Mariya Khudyakova, Mira Bergelson, Yulia Akinina,

Ekaterina Iskra, Svetlana Toldova, Olga Dragoy

National Research University Higher School of Economics
20 Myasnitskaya Ulitsa, Moscow 101000, Russia
E-mail: mariya.kh@gmail.com, mirabergelson@gmail.com, julia.akinina@gmail.com,
ekaterina.iskra@gmail.com, toldova@yandex.ru, olgadragoy@gmail.com

Abstract

In this paper we present a multimedia corpus of Pear film retellings by people with aphasia (PWA), right hemisphere damage (RHD), and healthy speakers of Russian. Discourse abilities of brain-damaged individuals are still under discussion, and Russian CliPS (Clinical Pear Stories) corpus was created for the thorough analysis of micro- and macro-linguistic levels of narratives by PWA and RHD. The current version of Russian CliPS contains 39 narratives by people with various forms of aphasia due to left hemisphere damage, 5 narratives by people with right hemisphere damage and no aphasia, and 22 narratives by neurologically healthy adults. The annotation scheme of Russian CliPS 1.0 includes the following tiers: quasiphonetic, lexical, lemma, part of speech tags, grammatical properties, errors, laughter, segmentation into clauses and utterances. Also analysis of such measures as informativeness, local and global coherence, anaphora, and macrostructure is planned as a next stage of the corpus development.

Keywords: aphasia, brain damage, discourse, Russian

1. Introduction

We present a corpus of Pear film (Chafe, 1980) retellings made by brain-damaged individuals and neurologically healthy speakers of Russian language. The primary aim of the Russian CliPS (Clinical Pear Stories) project is investigation of discourse abilities of people with aphasia (PWA) and right hemisphere damage (RHD) in comparison with neurologically healthy speakers. In the recent years there has been development in the studies of discourse in aphasia and other neurological conditions (Armstrong, 2000; Linnik et al., 2015), however, the effect of lesions in language-dominant and non-language-dominant hemispheres on discourse production and comprehension is still discussed. Aphasia is an acquired language impairment resulting from brain damage to the language-dominant hemisphere (usually left; Dronkers and Baldo, 2010). Aphasia of different types can manifest in disturbances in both production and comprehension of language on different levels: phonetic, lexical and syntactic. However, the research shows the discrepancy between the language competence of PWA on micro- and macro-linguistic levels (Armstrong, 2000; Linnik et al., 2015; Wright, 2011). Though early studies show that discourse structure is not impaired in aphasia (Ulatowska et al., 1983a, 1983b), some of the recent research demonstrates that it is not necessarily true. While several studies report that such discourse properties as informativeness and coherence are significantly different in discourse of PWA and healthy speakers (Van Leer and Turkstra, 1999; Nicholas and Brookshire, 1993; Wright et al., 2010), other studies report the opposite results (Glosser and Deser, 1990; Marini et al., 2005).

The damage to the non-language-dominant hemisphere (usually right) is not directly linked to any problems on micro-linguistic level; however there is evidence that people with RHD experience difficulties in language comprehension and production at discourse level (Brookshire and Nicholas, 1984; Tompkins et al., 1997).

At the present moment the only large corpus of aphasic speech is AphasiaBank (MacWhinney et al., 2011). Discourse elicitation stimuli for AphasiaBank are several pictorial stimuli as well as a picturebook with a Cinderella story. Though texts from AphasiaBank are used for discourse research (for example Richardson et al., 2016), any additional annotation does not become part of the corpus.

The goal of the Russian CliPS project is to create a corpus that could be used as a research tool with its existent annotation, and on the other hand, would be constantly developed by new research and additional information.

2. Corpus compilation

2.1 Speakers

Brain-damaged individuals were recruited in the inpatient departments of Moscow rehabilitation centers. The individuals with aphasia had been admitted to the centers with reported language problems after stroke in the left hemisphere and were diagnosed with chronic aphasia (not less than 6 months post-stroke). Aphasia types were diagnosed using Luria's classification (Akhutina, 2015; Luria, 1972), and the corpus contains stories by people with efferent motor, dynamic, acoustic-mnestic and sensory aphasia.

Aphasia can be generally divided in two different types: with fluent and non-fluent speech output. The non-fluent aphasia types include efferent motor aphasia and dynamic aphasia. Russian CliPS corpus contains 10 stories by individuals with efferent motor aphasia and 9 stories by individuals with dynamic aphasia. Aphasia types with fluent speech output include sensory and acoustic-mnestic aphasia. Russian CliPS corpus contains 10 stories by people with sensory aphasia and 10 stories by people with acoustic-mnestic aphasia.

Group	Number of speakers	Gender	Mean age	Age range	SD
Acoustic-mnestic aphasia	10	5 female, 5 male	51.3	40-68	9.4
Dynamic aphasia	9	5 female, 4 male	51.8	41-68	8.1
Efferent motor aphasia	10	3 female, 7 male	48.6	30-57	8.0
Sensory aphasia	10	4 female, 6 male	59.3	33-81	8.1
RHD	5	2 female, 3 male	50	41-56	12.3
Brain damage (total)	44	19 female, 25 male	52.8	30-81	10.4
Healthy	22	11 female, 11 male	58	25-84	13.9

Table 1. Demographic information on Russian CliPS 1.0 speakers

Individuals with RHD all were at least 6 months post-stroke and were right-handed.

Speakers from the neurologically healthy group did not report any history of neurological disease or head traumas. All the participants were native speakers of Russian language. The information about all speakers is summarized in Table 1.

2.2 Material and procedure

The elicitation stimulus, the Pear film, was made at the University of California in Berkeley in 1975 specifically for elicitation and collection of narratives by people from various cultures and languages (Chafe, 1980). It is a color film, and though it is not silent, characters do not produce any language. The film has a unique plot that was written to not resemble any other film, or book, or tale. Some characters of the film are important for the plot, and some just appear for a short moment and do not participate in the story. The film motivates the retellers to provide some moral judgement or interpretation of the story.

For the Russian CliPS corpus all speakers were asked to watch the film and then retell it in detail to the person who had not seen it before (the listener could be present at the time of the retelling, or the experimenter told the speaker that a person would listen to the recording afterwards). Both the experimenter and the listener did not ask any specific questions about the story, but could encourage the speaker with the general questions such as "And what happened next?" or "Would you like to add anything else?" The retelling of the story was audio recorded, also 20 brain-damaged speakers and all healthy speakers gave permission to be recorded on video.

3. Corpus Annotation

The annotation of the corpus was performed in ELAN (Wittenburg et al., 2006). The annotation scheme 1.0 includes the following tiers: quasiphonetic, lexical, lemma, POS, grammatical properties, errors, laughter, segmentation into clauses and utterances.

The quasiphonetic tier (Transcript) is aligned with the media files, and contains orthographic transcript of the speech recorded. Most words in this tier appear in their regular spelling, however in the cases of a phonetic error or a specific pronunciation, the transcript reflects these declinations from regular language. For example, in Russian the word *сейчас 'seychas' – now* in oral speech can appear in its full form or as a reduced variant *щас 'tschas'*. In writing, however, only the full variant is acceptable in standard language. In this case the quasiphonetic transcript should follow pronunciation rather than standard language rules. Phonemic paraphasias (errors) that happen in speech of PWA are also reflected in this tier, for example *велосипел 'velosipel'* (the correct word is *велосипед 'velosiped' – bike*). At the quasiphonetic level all the pauses that are longer than 70ms are annotated, both absolute and filled pauses. If some segment of speech is not comprehensible, the note "incomprehensible" is used.

The quasiphonetic transcript makes it possible to capture some features of oral speech as well as phonemic paraphasias, but it would cause problems for analysis of lexical diversity and lexical density. The lexical transcript tier (Transcript_lex) contains the same information as the quasiphonetic tier, although all the spellings are brought to standard. Lexical transcript is used for calculating lexical richness, because different pronunciations of one lexeme are not counted as different words.

Lemma tier (Lemma) contains initial forms of the words, and in the English lemma tier (Lemma_eng) all the words are translated into English, which, in combination with information from grammatical tiers, makes the data from Russian CliPS available for non-Russian speaking researchers.

The part of speech tagging scheme and the annotation of grammatical categories is based on the manual of Russian National Corpus (http://www.ruscorpora.ru/en/corpora-morph.html).

Laughter is annotated on a separate tier (Laughter) and is aligned with the sound wave. This annotation enables the analysis of laughter as a marker of failure to produce a correct word or interpretation of an event in the stimulus film, as well as dissatisfaction with the whole narrative (Khudyakova and Bergelson, 2015).

Grammatical, semantic and phonetic errors are annotated in the special tier (Error). Phonetic errors include replacement of one sound with another, for example *сапка 'sapka'* (the correct word is *шапка 'shapka'* – hat), omission of a phoneme or inclusion of an extra one, for example *поропал 'poropal'* (the correct word is *пропал 'propal'* – got lost), and use of a word that is phonetically similar with the intended one, but is distant semantically, for example *грустные 'grustnye' – sad*, instead of *груши*

'grushi' – pears. Semantic errors include use of a member of the same semantic category as the target word, for example *apples* instead of *pears*, or *sheep* instead of *goat*. In several cases the distinction between the two types of errors is impossible, for example use of *сановник 'sanovnik' – dignitiary* instead of *садовник 'sadovnik' – gardener* can be interpreted both as a phonetic error (replacement of /d/ with a /n/) or as a semantic one (use of a wrong word from 'professions' category), and in this case both types of error are annotated. Grammatical errors include errors in agreement and number.

Segmentation into clauses is based on grammatical rather than prosodic (Kibrik and Podlesskaya, 2009) principle. Utterances include a main clause with all its subordinate clauses (Glosser and Deser, 1990). A ratio of clauses and utterances can be interpreted as a measure of grammatical complexity (Marini, 2012).

4. Corpus data

The current version of Russian CliPS contains 66 narratives. The total length of the recorded material is 4 hours 33 minutes. The mean length of each recording is 4 minutes 7 seconds (min – 38seconds, max – 18 minutes 27 seconds, SD = 175 seconds).

The quantitative information on the current version of the Russian CliPS corpus is shown in Table 2.

At the present moment the Russian CliPS corpus is not publicly available.

5. Coreference Annotation

We started with coreference annotation. As an annotation tool, we have chosen the platform that was designed for the annotation of RuCoref – Russian Coreference corpus (Toldova et al., 2014; http://ant0.maimbava.net/). This platform was developed for the purpose of anaphora resolution systems evaluation campaign. It is based on MySQL database engine and has a convenient Web-interface that allows parallel annotation by several annotators and on-line tracking of discrepancies between annotators. It supports embedding of annotations, marking zero anaphora, annotation features enhancing by users and establishing links between markables. At start, this platform had built-in annotation scheme worked-out for coreference annotation of written texts (primarily, news). The scheme has a set of features concerning NP structure type, referential status type and coreferent NPs relation type.

We have completed pilot annotation of 10 narratives by healthy speakers and 5 narratives by PWA (efferent motor). This pilot stage revealed steps needed for adaptation of the coreference annotation scheme designed for written texts for a genre of oral narrations.

Firstly, while in written texts discontinuous noun phrases are a special type of coreference relations (e.g. a referent is a group of two other referents before mentioned in a text), in oral texts disrupted NPs are a very frequent phenomenon. The disruption is due to pauses, discourse markers and filler words. The NP disruptions are even more frequent in narratives by people with non-fluent types of aphasia. Thus, we have to create additional functionality for our platform, namely, the annotation of two disrupted text pieces as one markable.

Another problem is that the standard relation between two NPs denoting the same entity in written texts is a

Group		Narrative length (ms)	Pauses (%)	Narrative length (words)	Narrative length (clauses)	Narrative length (utterances)
Acoustic-mnestic aphasia	**Mean**	231 196	43	281,1	52,6	45
	Range	85 229 - 473 025	25-55	76-480	18-84	16-69
	SD	106 700	10	122,2	19,7	16,6
Dynamic aphasia	**Mean**	406 023	60	220,4	39,8	38,8
	Range	138 096 – 810 867	29-71	135-371	27-59	26-59
	SD	196 132	13	91,1	9,4	9,9
Sensory aphasia	**Mean**	275 765	40	346,4	66,3	58,9
	Range	148 023 – 549 223	24-56	170-631	28-110	25-94
	SD	117 912	9	174,7	29,6	25,6
Efferent motor aphasia	**Mean**	377 137	45	228,8	49,9	43,8
	Range	167 879 – 1 107 112	26-72	58-436	14-91	14-64
	SD	275 043	14	119,7	24,4	17,8
RHD	**Mean**	195 922	49	279	63	55,7
	Range	122 845 – 427 025	39-65	185-477	32-120	29-105
	SD	147 132	11	133,5	39,2	33,8
Healthy speakers	**Mean**	152 437	33	269,5	53,7	42,2
	Range	47 389 – 296 805	17-51	88-405	16-80	9-71
	SD	62 524	9	113,7	21,7	18,4

Table 2. Quantitative data on Russian CliPS 1.0

coreferential relation (the relation of referent identity), though other relations such as apposition (c.f. *a 10-year boy*, *the one with a basket*, …) or predicative relation (c.f. *this boy is a boy who* …) are taken into consideration in the built-in scheme. It oral texts, some NPs denoting the same referent as a previous NP are just mere NP repetitions (c.f. *a boy, this boy, went*…), or self-correcting (*a gardener, a farmer, went*…). Thus, in order to distinguish the latter from apposition we need additional labels for NPs relational types (e.g. repetition, renaming). We also need additional rules in the annotation instruction for the differentiation of appositions vs. different types of repetitions.

The third issue worth mentioning concerns the naming problem in speech-impaired people. These are the cases of semantic paraphasias (c.f. *apples* instead of *pears*). Sometimes the speakers make self-correction during the narration. These cases are also should be captured by our scheme. The annotation scheme should also allow marking potential coreference relations between an NP and more than one potential referent.

Our pilot study has highlight some peculiarities of coreference chaining in oral discourse and in speech of different kinds of brain-damaged individuals and some special issues in annotation process for this type of discourse. Thus, this discourse level needs further investigation and the coreference annotation scheme needs further enhancing and adaptation.

6. Future Work

The Russian CliPS corpus at its present stage is annotated on micro-linguistic level. Much discourse annotation is still needed in order to evaluate the discourse abilities of brain-damaged speakers.

The next version of Russian CliPS will also have annotation of informativeness, global and local coherence, and macrostructure of discourse.

7. Bibliographical References

Akhutina, T. (2015). Luria's classification of aphasias and its theoretical basis. *Aphasiology*, 1–20. doi:10.1080/02687038.2015.1070950.

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology* 14, 875–892. doi:10.1080/02687030050127685.

Brookshire, R. H., and Nicholas, L. E. (1984). Comprehension of directly and indirectly stated main ideas and details in discourse by brain-damaged and non-brain-damaged listeners. *Brain and language* 21, 21–36.

Chafe, W. (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. , ed. W. Chafe Norwood, New Jersey: Ablex.

Dronkers, N. F., and Baldo, J. V. (2010). "Language: Aphasia," in *Encyclopedia of Neuroscience* (Elsevier Ltd), 343–348.

Glosser, G., and Deser, T. (1990). Patterns of Discourse Production among Neurological Patients with Fluent Language Disorders. *Brain and Language* 49, 67–88.

Khudyakova, M. V., and Bergelson, M. B. (2015). Interpretation of "Embarrassement" Laughter in Narratives by People with aphasia and Non-language-impaired speakers. in *Proceedings of the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech, 14-15 April 2015* (Enschede).

Kibrik, A. A., and Podlesskaya, V. I. eds. (2009). *Night Dream Stories: A corpus study of spoken Russian discourse*. Moscow: Languages of Slavonic Culture.

Van Leer, E., and Turkstra, L. (1999). The effect of elicitation task on discourse coherence and cohesion in adolescents with brain injury. *Journal of Communication Disorders* 32, 327–349. doi:10.1016/S0021-9924(99)00008-8.

Linnik, A., Bastiaanse, R., and Höhle, B. (2015). Discourse production in aphasia : a current review of theoretical and methodological challenges. 7038. doi:10.1080/02687038.2015.1113489.

Luria, A. R. (1972). Aphasia reconsidered. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior* 8, 34.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology* 25, 1286–1307. doi:10.1080/02687038.2011.589893.

Marini, A. (2012). Characteristics of narrative discourse processing after damage to the right hemisphere. *Seminars in Speech and Language* 33, 68–78. doi:10.1055/s-0031-1301164.

Marini, A., Carlomagno, S., Caltagirone, C., and Nocentini, U. (2005). The role played by the right hemisphere in the organization of complex textual structures. *Brain and Language* 93, 46–54. doi:10.1016/j.bandl.2004.08.002.

Nicholas, L. E., and Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of speech and hearing research* 36, 338–350.

Richardson, J. D., Dalton, S. G., Richardson, J. D., Grace, S., Main, D., Richardson, J. D., et al. (2016). Main concepts for three different discourse tasks in a large non-clinical sample. 7038. doi:10.1080/02687038.2015.1057891.

Toldova, S., Roytberg, A., Ladygina, A., Vasilyeva, M., Azerkovich, I., Kurzukov, M., et al. (2014). RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"* 13, 681–695.

Tompkins, C. A., Baumgaertner, A., Lehman, M. T., and Fossett, T. R. D. (1997). Suppression and discourse comprehension in right brain-damaged adults: A preliminary report. *Aphasiology* 11, 505–519.

Ulatowska, H. K., Doyel, A. W., Stern, R. F., Haynes, S.

M., and North, A. J. (1983a). Production of procedural discourse in aphasia. *Brain and language* 18, 315–341. doi:10.1016/0093-934X(83)90023-8.

Ulatowska, H. K., Freedman-Stern, R., Doyel, A. W., Macaluso-Haynes, S., and North, A. J. (1983b). Production of Narrative Discourse in Aphasia. *Brain and Language* 19, 317–334. doi:10.1016/0093-934X(83)90074-3.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. in *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.

Wright, H. H. (2011). Discourse in aphasia: An introduction to current research and future directions research. *Aphasiology* 25, 1283–1285.

Wright, H. H., Koutsoftas, A., Fergadiotis, G., and Capilouto, G. (2010). Coherence in Stories told by Adults with Aphasia. *Procedia - Social and Behavioral Sciences* 6, 111–112. doi:10.1016/j.sbspro.2010.08.056.

Mining Auditory Hallucinations from Unsolicited Twitter Posts

M. Belousov[1], M. Dinev[1], R. M. Morris[2], N. Berry[2,3], S. Bucci[2], G. Nenadic[1,3]
[1] School of Computer Science, University of Manchester
[2] School of Psychological Sciences, University of Manchester
[3] Health eResearch Centre (HeRC), The Farr Institute of Health Informatics Research
School of Computer Science, University of Manchester, Kilburn Building, Manchester, United Kingdom, M13 9PL
{mbelousov, mdinev, gnenadic}@cs.man.ac.uk

Abstract

Auditory hallucinations are common in people who experience psychosis and psychotic-like phenomena. This exploratory study aimed to establish the feasibility of harvesting and mining datasets from unsolicited Twitter posts to identify potential auditory hallucinations. To this end, several search queries were defined to collect posts from Twitter. A training sample was annotated by research psychologists for relatedness to auditory hallucinatory experiences and a text classifier was trained on that dataset to identify tweets related to auditory hallucinations. A number of features were used including sentiment polarity and mentions of specific semantic classes, such as fear expressions, communication tools and abusive language. We then used the classification model to generate a dataset with potential mentions of auditory hallucinatory experiences. A preliminary analysis of a dataset ($N = 4957$) revealed that posts linked to auditory hallucinations were associated with negative sentiments. In addition, such tweets had a higher proportionate distribution between the hours of 11pm and 5am in comparison to other tweets.

Keywords: machine learning, text mining, hallucinations, psychosis, psychotic-like experience, social media, twitter

1. Background

Social networking is pervasive, with an estimated 305 million monthly active users on Twitter only (Statista, 2015). This vast amount of user generated data presents a unique opportunity for researchers to access information regarding mental health that did not previously exist. Although Twitter posts are public and, therefore, may be influenced by audience awareness, this approach may reduce the observer or 'Hawthorne effect' (McCarney et al., 2007), negate methodological reactivity, and may have the advantage of accessing hard to reach groups who would not typically self-select for research. Moreover, data generated in naturalistic environments may decrease the effect of biases associated with recall (Stone and Shiffman, 2002) and unsolicited patient-generated data may confer the advantage of identifying novel (data-driven) associations of variables.

Psychotic disorders are characterised by delusions, hallucinations, disorganised thinking, disorganised or abnormal motor behaviour and negative symptoms (American Psychiatric Association, 2013). The symptoms of psychosis present a unique challenge for text mining processes as the population frequently lack insight into their disorder (Lincoln et al., 2007) and by definition lack insight into one of the hallmark symptoms (delusions; a firmly held yet erroneous belief). Recent reviews and meta-analyses (Clarke et al., 2012; Van Os et al., 2009) suggest that there may be numerous risk factors for psychosis and psychotic-like experience. Psychotic-like experiences are similar to psychotic symptoms, however, they are often; attenuated and/or transient; and do not result in a clinical need (Van Os et al., 2009). Nonetheless, they can provide valuable information about factors influencing psychotic disorder (Van Os et al., 2009). In the absence of one putative cause of psychosis, for example the 'schizogene' (Meehl, 1990), it is necessary to identify factors that influence the incidence of psychotic-like experience, psychotic episodes, or psychotic disorder, which will have implication for practice, policy and resource allocation.

Individuals who experience mental health problems, including psychosis, have reportedly high rates of social media usage (Gowen et al., 2012; Birnbaum et al., 2015). However, previous reports of social media usage in this population are limited to small scale studies ($N = 207$ and $N = 80$ respectively), which lack the rigor of larger population-based studies, in addition, these studies recruited young people (range 12-24 years old) experiencing mental health problems. Within the field of psychology, Twitter has been used to collect information via specific hashtags (Joseph et al., 2015; Reavley and Pilkington, 2014; Shepherd et al., 2015). Previous research has also utilised text mining approaches on Twitter to collect unsolicited Twitter posts regarding mental health. For example, researchers have automatically mined Twitter posts containing self-reported diagnoses of mental health problems and were able to distinguish differences in language use between disorders (Coppersmith et al., 2014; Coppersmith et al., 2015).

This exploratory study seeks to establish whether it is feasible to generate datasets from unsolicited Twitter posts regarding psychotic(-like) phenomena[1]. We are specifically interested in auditory hallucinations, i.e. the interpretation of stimuli without the appropriate sensory input. This study also seeks to identify the semantic classes and sentiment in the associated tweets and explore associations between the aforementioned variables and the tweet 'meta-data' (e.g. tweet time).

[1] In the current investigation it is impossible to ascertain if the phenomena being reported is a symptom of psychosis (i.e. a psychotic experience), or not (i.e. a psychotic-like experience). Due to this ambiguity the term psychotic(-like) is used.

2. Methods

This project utilised an iterative workflow, which is described below (see Figure 1).

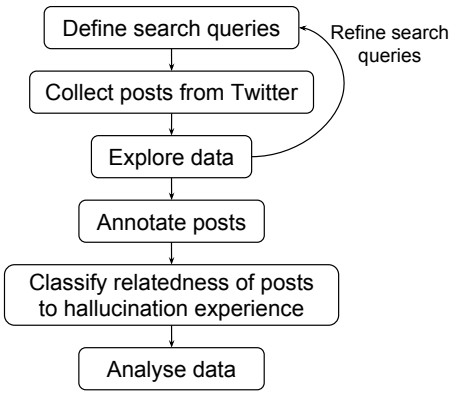


Figure 1: Study workflow

2.1. Data Collection

To collect the posts about potential hallucinatory experiences we have used the Twitter Search API. Using the query operators provided, we have defined seven different search queries (Table 1) based on keywords from the literature (Nayani and David, 1996) and informed by researchers with experience of delivering clinical assessments. The data crawling process was triggered multiple times per day to get as complete result set as possible, between November 2014 and March 2015.

Search query
hallucinating hearing
(“hear things” OR “hearing things”) “in my head”
hearing scary things “in my head”
(hear OR hearing) (“other people” OR “other ppls” OR “other ppl”) thoughts
(voice OR voices) (commenting OR criticising) (scary OR frightening OR “everything I do”)
(hear OR hearing) (voice OR voices) (god OR angel OR allah OR soul OR spirit OR “holy spirit” OR djinn OR jinn)
(hear OR hearing) (voice OR voices) (scary OR devil OR demon OR daemon OR evil OR “evil spirit”)

Table 1: Search queries

Local time identification: Twitter stores information regarding the time of posting in a UTC format, rather than the local time in which the post was generated. Whilst maintaining the anonymity of the user, we implemented an algorithm to calculate the local time of each tweet, based on the Twitter data attributes. Our proposed method estimates the local user's timezone in a hierarchical order, first accessing the geolocation attributes of the specific tweet. If this information is not available, we utilise UTC offset as defined in the user's profile.

2.2. Duplicate Detection

In Twitter, more than 85% of posts are news-related (Kwak et al., 2010). Therefore, even though data collection was conducted using specific queries, the results still contained repetitive posts that describe widely held beliefs (e.g. “Cannabis causes psychosis and people to hear voices”), news or advertisements (i.e. spam). A duplicate detection phase is introduced to identify the posts that convey the same or highly lexically similar content from the dataset as specified below.

In a recent study, five different levels of near-duplicates was defined (Tao et al., 2013), from which we derived only the first two with slight modifications and also introduced one additional similarity layer.

Exact copy: Two tweets are case-insensitively (i.e. ignoring the letter case) identical.

Nearly exact copy: A pair of tweets are case-insensitively identical except for Twitter-related entities (i.e. *#hashtags*, *URLs* and *@mentions*).

Lexically similar copy: In consideration of all properties from two previous levels, the similarity ratio between two candidates is measured and compared against a defined threshold. This level of duplicates can be helpful in cases when the defined search queries match popular song lyrics or quotes (e.g. *“Sometimes thoughts, fears, or other people get too loud and we can’t hear ourselves”*). We have observed that frequently people quote them *indirectly* with syntactic variations, spelling variations and typos. To measure the difference between two texts, we used *Levenshtein distance* and then the similarity ratio is calculated as $1 - \frac{\text{levenshtein distance}}{\text{total number of characters}}$.

Based on observation of various experiments, the threshold value of 0.85 was identified as the most appropriate and everything above this value was marked as duplicates and excluded from the dataset.

2.3. Initial Data Exploration

It was important to find features that would refine the search queries and thereby improve the results. To investigate the different attributes and derive a suitable collection of tweets, we manually reviewed and profiled the data through distributional statistics such as word frequency, term frequency inverse document frequency (TF-IDF) and *n*-grams. A set of standard stop words was excluded from this initial exploration, but was included in all subsequent analyses.

2.4. Data Annotation

Supervised machine learning systems typically require a large amount of labelled data which can be used to train a model. The process of data annotation may include assigning specific class tags to a whole tweet (e.g. associated sentiment such as positive, negative or neutral), label a whole

word or phrase, and even specify relations. Providing such annotations, especially in a mental health context, requires knowledge of the domain and is resource (time) intensive. Our investigations revealed no annotated datasets available at the time of our research, so we have conducted our own annotation procedure.

2.4.1. Generating a Dataset for Annotators

Finding *relevant examples* that mention auditory hallucinatory experiences is a time-consuming task for annotators due to the high proportion of unrelated examples. Therefore, we decided to filter the dataset using various queries (i.e. search terms associated with possible hallucination experiences) that aimed to further narrow the results. Then, we provided to the annotators a random combination of the filtered and unfiltered tweets. It is necessary to include a combination of filtered and unfiltered tweets to ensure informative data is not missed due to a selection bias.

2.4.2. Annotation Environment

In order to optimise the annotation process, research psychologists need a user-friendly and efficient tool to label tweets efficiently. Although there are several web-based text annotation tools available such as *BRAT (brat rapid annotation tool)* (Stenetorp et al., 2012) and *GATE Teamware* (Bontcheva et al., 2013), we have designed and developed our own bespoke annotation application, that was aimed to minimise the time spent on the annotation of each example (see Figure 2).

In order to specify whether the post is related to hallucinatory experience or not, the annotators assign an appropriate class to the whole tweet. Also the annotators were asked to explain their decision by highlighting corresponding words and phrases that inform their classification. This information was also used to identify potentially useful characteristics of different classification categories.

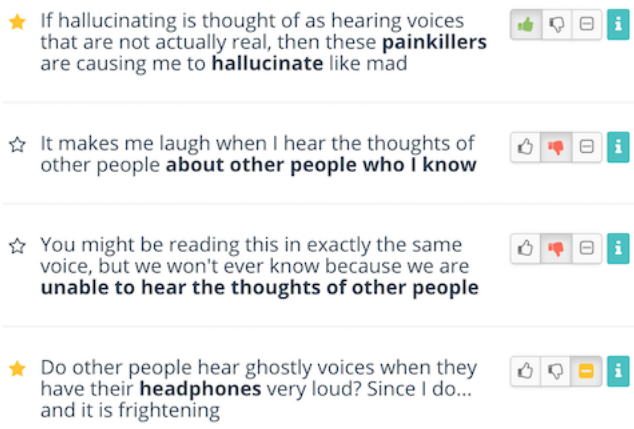


Figure 2: Human annotation of hallucinatory experiences to illustrate the annotation tool in use.

The application was utilised by two research psychologists (RMM, NB) who work within the field of psychosis. Annotations were conducted independently to reduce bias and feedback was provided which resulted in an additional classification category ("unsure") that was used to skip the instance when information was insufficient. The annotators worked according to pre-defined guidelines, which specified that: Twitter posts must be in regards to a personal experience; Twitter posts must have an explicit mention of an auditory hallucinatory experience which is not accounted for by other factors in the text (e.g. *"hearing voices, the radio is too loud"*); Twitter posts are coded conservatively (i.e. ambiguous posts are assigned the unsure/unrelated category).

To ensure consistency between the annotators, we measured the inter-annotator agreement, through the Cohen's kappa. The observed agreement was 0.849 on 41 examples (10% of the final annotated set).

We observed the unequal distribution of classes. Final annotated dataset contained 401 examples: 94 related and 307 unrelated examples (skew ratio is 3.27).

2.4.3. Defining Semantic Classes

During the analysis of annotated data, we have identified several groups of words with shared semantic properties, such as various communication tools and family members. Sixteen different semantic classes were then defined (Table 2) and terms associated with the semantic classes were extracted as a feature from the data. We note that semantic classes did not form the basis of any rule based decision (i.e. explicit mentions of semantic classes were not taken as unequivocally indicative of auditory hallucinations).

Semantic class	Common examples	Total
Abusive Language	*f*cking, sh*t, hell*	503
Relative	*baby, mom, son, ex, father, friend, grandma*	130
Religious Term	*Jesus, prayer, psalm, Bible, pastor, church*	21
Audio Recording	*recording, voicemail, voice message*	20
Audio & Visual Media, Application	*song, music, YouTube, Siri*	19
Audio Device	*radio, TV, speaker, headphones*	12
Fear Expression	*scary, scared, creepy, afraid, nervous*	12
Drug	*cannabis, weed, LSD, pain killers, ecstasy*	11
Stigmatising Language	*crazy, insane*	11
Emotion Support	*helpline, lifeline, IFOTES, Samaritans*	9
Negative Supernatural	*devil, demon*	7
Own Voice	*my voice, our own voice*	6
Cause	*fever, sleep paralysis*	5
Communication Tool	*phone, smartphone*	5
Possible Hallucination	*in my head, seeing things*	4
Supernatural	*spirit*	3

Table 2: Semantic classes, their common examples and total number of members

2.5. Text Pre-processing and Normalisation

We used a set of pre-processing steps for social media which aimed to reduce amount of possible noise in a text and improve accuracy of natural language processing (NLP) algorithms such as named entity recognition.
As an initial step, all Twitter-related entities (i.e. *#hashtags*, *URLs*, *@mentions*), unicode emoticons and numerals were removed.

2.5.1. Identification of Nonstandard Language

In order to identify nonstandard words (such as spelling variations and slang), we have utilised several NLP techniques. After splitting the text into tokens, we determine a part-of-speech (POS) tag for each word which will be used in further phases. We found *TweetNLP tagger* (Owoputi et al., 2013) the most suitable for our cases, since it was initially trained on Twitter data, has a built-in tokeniser and handles additional POS tags, emoticons and proper noun recognition (Owoputi et al., 2012). Spell checking based on *MySpell* is applied to find words that are not in the dictionary.
We have distinguished different types of *out-of-vocabulary* (OOV) words and proposed corresponding processing approaches to each of them. The occurrences and specific characteristics of each type is recorded and used later in the feature extraction stage.

Abbreviations, slang and interjections: To transform slang and abbreviations (e.g, "idk" $\mapsto$ "I don't know"), we have collected our own vocabulary based on data from two public dictionaries[2,3]. We have used and extended *Dictionary of interjections*[4] to expand interjections like "oops" $\mapsto$ "I didn't mean to do that" and "wow" $\mapsto$ "amazing".

Named entities: We have observed that Twitter posts often do not use case-sensitive words, and that even named entities are not capitalised properly. Therefore, the mentions of *potential* proper nouns were transformed to the title case to improve named entity recognition. Any out-of-vocabulary word that was tagged as a *proper noun*, *proper noun + possessive* or *determiner* (Gimpel et al., 2013) was classified as a potential proper noun and transformed appropriately in addition to known named entities.

Semantic classes: Nonstandard words are also matched against the dictionary of semantic classes defined after the annotation phase. For example, *Relative* semantic class also contained different slang words used to describe family members and friends (e.g. "*ex*", "*grandma*").

Misspelled words: To automatically correct any remaining misspellings, we utilised a spell correction algorithm that suggests replacement candidates by calculating minimum edit distance between out-of-vocabulary and words from the English dictionary provided as part of *MySpell*.

Unknown slang: Words that failed to fit into at least one of the categories above were marked as "unknown slang" and removed from the text to reduce the ambiguity.

[2] http://www.gaarde.org/acronyms/

[3] http://www.noslang.com/dictionary/

[4] http://www.vidarholen.net/contents/interjections/

2.6. Feature Extraction

During this stage we apply various techniques to extract specific features from the Twitter posts, that can be used in further analysis and classification.
We have engineered nine groups of features, including frequencies of mentions of individual semantic classes (for each class from Table 2), POS tags (as the number of mentions of each POS), popularity of the post (number of likes and retweets, as returned by Twitter), number of Twitter entities (i.e. URLs, hashtags, mentions and financial symbols), mentions of specific named entities (persons, locations and organisations) and lexical distribution (i.e. number of sentences, words, upper-case and lower-case characters). Additionally, we also used the following feature groups:

Use of nonstandard language: All transformations performed during the *text normalisation* (described in the previous section) tend to be an informative characteristic of the narrative itself. Therefore, information like mentions of different semantic classes, number of spelling mistakes and overuse of unknown slang is extracted for further analysis.

Sentiment polarity: In subjective posts, where people describe personal opinions or feelings, it is useful to know which emotion (positive, negative or neutral) was expressed. Therefore, we have applied an unsupervised sentiment classifier which was originally utilised for improving one-class collaborative filtering of user comments over TED talks (Pappas and Popescu-Belis, 2013).

Key phrases: There is an inherent value in determining the nature and content of auditory hallucinations (Haddock et al., 1999). Thus, in addition to analysing information contained within the whole of the Twitter post, further analyses were performed on the portion of the post referring to a (potential) auditory hallucination. To identify such phrases in the text, we implemented an algorithm which analyses the parse (syntactic) tree of a given sentence generated by *Stanford Parser* (Chen and Manning, 2014), and seeks the specified target node (verb "hear" in any tense). We then extracted all relevant words descendant from the target node to construct a *key phrase*. We then added the following features related to a given key phrase: its sentiment polarity, mentions of semantic classes within the key phrase, and POS tags from it.

2.7. Classification

In order to produce a classifier that automatically predicts whether a given tweet is related to auditory hallucinatory experiences or not, we have defined a set of 100 **semantic features** as specified above. We have experimented with different types of classification algorithms, such as Naive Bayes, Supporting Vector Machines and AdaBoost. As a baseline for comparisons, we have used simple **lexical features**, which were generated using TF-IDF of stemmed words (without removing stop words). This feature set contains 1370 features.
Having an imbalanced dataset, a high accuracy score (or low error rate) does not necessarily mean a good classification performance.

Therefore, F_β-measure was selected as a metric which reflects performance by combining both precision and recall. Moreover, β parameter is used to define the importance of these underlying metrics, so that $\beta < 1$ puts more weight on the precision and $\beta > 1$ emphasises the recall (Rijsbergen, 1979).
In our classification scenarios, as in many other healthcare applications, it seems reasonable to keep the number of *false negatives* low. Therefore it was decided to put more weight on a recall and use **F2-measure** as an evaluation metric of classification models.

2.8. Data Visualisation

A visualisation application was constructed (Figure 3) in order to present aggregated statistics, such as part of the day distribution, sentiment polarity distribution, different types of named entities and distribution of semantic classes.

Figure 3: Fragment of data visualisation application which presents the distribution of semantic classes and their frequent members with a word cloud.

2.9. Ethical Considerations

A vital component of the development of the methodology for this project was identifying and addressing the various ethical issues associated with social media research. To this end, the project was developed in accordance with the guidelines stipulated by the Association of Internet Researchers (AOIR, 2012), the British Psychology Society (BPS, 2013) and INVOLVE (INVOLVE, 2014). Specifically, these guidelines were consulted to maintain the anonymity and confidentiality of Twitter users' and their data through the implementation of strategies, such as dataset anonymisation, password protected data storage, paraphrasing of tweets in publications and avoidance of individual profiling. The project was ethically approved by the School of Computer Science Research Ethics Committee, University of Manchester in October 2014.
It has been argued that data on Twitter is publicly available and, therefore, acceptable to be utilised for research purposes (Shepherd et al., 2015). However, other researchers have noted that even though social media postings are public, it does not mean that it is always ethically acceptable to use the data for any research purpose (Conway and O'Connor, 2016). Therefore, several methodological considerations were implemented to ensure that personal data mined from Twitter was treated both ethically and respectfully.
First, data was only mined from profiles set to public (not private). To maintain data confidentiality, Twitter handles (usernames) and user mentions were automatically removed from the data before any processing or analysis, including annotation. The data was stored on an encrypted drive. In addition, each member of the research team was assigned authentication credentials to access the tweets mined from Twitter for analysis purposes. Unique user IDs and passwords for each researcher ensured the confidentiality and privacy of the dataset and prevented unauthorised individuals gaining access.
To preserve anonymity, direct quotations of tweets were not included in any manuscript or presentation. Instead, tweets were paraphrased to ensure that individual users could not be identified if the quote was entered into the Twitter search bar or a search engine (i.e. Google and Microsoft Bing). Paraphrasing any tweets for publication ensured that the anonymity of the Twitter users was protected and they were not identifiable from any of the quotations included (Rivers and Lewis, 2014).

3. Results

To evaluate the performance of the classifiers over the two different sets of features (lexical and semantic), we have performed ten experiments of stratified 10-fold cross validation and measured F2-score and area under the receiver operating characteristic (ROC) curve (AUC) as shown in Table 3.

Classifier	Semantic features		Lexical features	
	F2-score	AUC	F2-score	AUC
NB	0.831	0.889	0.486	0.711
AdaBoost	0.772	0.858	0.711	0.822
SVM	0.743	0.836	0.751	0.848

Table 3: Classification results for the Naive Bayes (NB), AdaBoost and Support Vector Machines (SVM) classifiers on two different sets of features.

Although the SVM classifier performed better on lexical features, the difference was statistically insignificant (p-value $= 0.375$). In all other cases, better performance was demonstrated using the proposed semantic features. The Naive Bayes classifier with our suggested feature set achieved the highest classification score.

3.1. Contribution of Features

To investigate contribution of each group of features on the NB classification scores we have performed several leave-one-out classifications.
It can be observed that the majority of feature groups tend to improve classification performance and both F2-score and AUC scores decreased after the exclusions. Based on this results, we can conclude that mentions of semantic classes, sentiment polarity and features extracted from key phrases significantly contribute to the classification score.
When the lexical distributional features and named entities were excluded, the scores slightly increased, which means that these groups of features decreased the efficacy of the classifier, although the difference was very small and not statistically significant.

Feature group	F2-score	AUC
Mentions of semantic classes	* 0.769	0.848
Key phrases	* 0.788	0.866
Part-of-speech tags	0.817	0.882
Sentiment	* 0.818	0.881
Popularity of the post	0.828	0.887
Use of nonstandard language	0.831	0.889
Number of Twitter entities	0.832	0.889
Named entities	0.832	0.890
Lexical distribution	0.833	0.889
All	0.831	0.889

Table 4: Leave-one-out classification scores showing how NB classifier performance was affected as one group of features was excluded from the set. Statistically significant differences are marked with asterisk.

3.2. Error Analysis

We have conducted an error analysis of all 29 misclassified examples identified during the ten experiments of stratified 10-fold cross validation. Ten misclassified examples (34%) contained mention(-s) of semantic class(-es) that were usually observed in the *hallucination-related* classification category (such as *fear expressions* or *abusive language*).

Five classification errors (17%) were subjective posts containing a mention of someone else who hears something (e.g. *"I am listening to an audio about how a person with schizophrenia hears voices, it is really scary."*).

Three examples (10%) that have been incorrectly classified as a hallucination-related contained negations (e.g. *"I do not hear voices, I am not paranoid"*, *"I'm hallucinating I'm hearing hawks! Oh hang on, it is just the television"*).

Around 10% of tweets were too brief, so the information extracted from that post was not enough to automatically identify informative features.

Also, we have observed that extending our semantic classes, especially defining new phrases that indicates possible hallucinations (e.g. *"telling me to hurt myself"*) could improve classification performance.

3.3. Preliminary Data Analysis

The best-performing classifier was applied to automatically predict classes for a larger set of unannotated data (4556 examples). About 10% of these tweets (452 examples) were predicted to be related to an auditory hallucinatory experience. In order to perform preliminary data analysis, we combined them with 401 already manually annotated examples which includes 94 related and 307 unrelated posts. Our final dataset for analysis contained 4957 examples (546 related and 4411 unrelated posts).

3.3.1. Sentiment Polarity and Hallucinatory Experience

Sentiment polarity (either positive or negative) was identified in 2830 tweets: 1883 (67%) were predicted to have a positive sentiment (Table 5). There was a significant association between the prediction of the tweet as reporting a hallucination and the sentiment classified within that tweet ($\chi^2(1) = 3337.09$, $p < .001$). The odds ratio of the tweet having a negative sentiment were 11.22 times higher if the tweet was categorised as related to hallucination than if not. This indicates that tweets predicted to be related to hallucinations had more negative sentiment than those deemed not related to hallucinations.

	Unrelated to hallucination count (% of column)	Related to hallucination count (% of column)
Negative sentiment	708 (27.92)	239 (81.29)
Positive sentiment	1828 (72.08)	55 (18.71)
Total	2536	294

Table 5: Cross tabulation of polarity of sentiment by predicted relatedness of tweet to hallucination.

3.3.2. Time of Tweet and Hallucinatory Experience

We sought to explore the relationship between the (local) time the tweet was generated and whether the tweet was predicted to be related to report of a hallucination. Figure 4 plots the Gaussian kernel density estimate across time. It indicates that between the hours of 11pm and 5am there is a greater proportional density of tweets predicted to relate to a hallucination. This data could plausibly indicate that there is an important relationship between the time of day and psychotic(-like) experience.

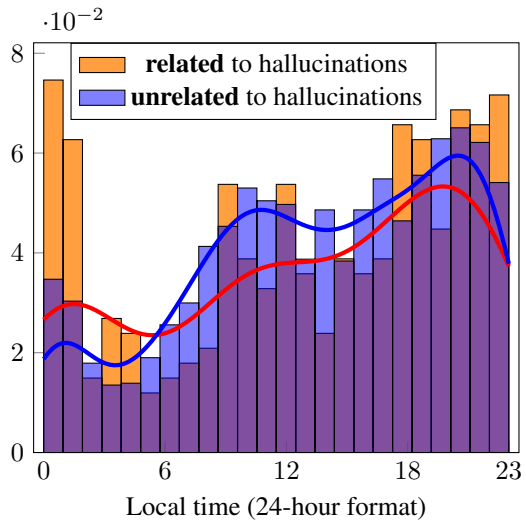


Figure 4: Gaussian kernel density estimate across time

3.3.3. Semantic Classes as Predictors of Hallucinatory Experience

To establish if the most pertinent semantic classes of interest were predictive of whether or not a tweet was assigned to be related a logistic regression was performed. The model included the semantic classes listed in Table 6, in addition to *Drugs* and *Emotional Support*. Due to the *Drugs* and *Emotional Support* classes predicting the outcome of the model "perfectly", they were omitted from the

model. The results of the logistic regression (see Table 6) indicate that *Abusive language*, *Expression of Fear*, *Stigmatising Language*, and *Negative Supernatural* phenomena were all significant positive predictors of the relatedness of the tweet. On the other hand, the mention of *relatives* was a significant negative predictor and *positive/neutral supernatural* phenomena was insignificant.

Semantic class	Odds ratio	P	95% Confidence interval	
Abusive Language	2.33	$< .001$	1.87	2.90
Fear Expression	8.99	$< .001$	7.03	11.50
Negative Supernatural	2.61	$< .001$	2.03	3.35
Relative	0.40	$< .001$	0.24	0.65
Stigmatising Language	7.85	$< .001$	4.52	13.65
Supernatural	0.52	.10	0.24	1.13

Table 6: Semantic class as predictor of relatedness to hallucinatory experiences.

4. Conclusions

This exploratory study aimed to establish the feasibility of harvesting psychotic(-like) experiences from unsolicited tweets. We have developed a text mining methodology to collect, process and classify tweets as being related to auditory hallucinations. The study also sought to explore the relationship between self-reported psychotic(-like) experiences and meta-data, in addition to identifying the semantic classes and sentiment polarity in Twitter posts. We were able to identify that negative sentiments were significantly associated with tweets that indicated the occurrence of auditory hallucinations, which supports the notion that auditory hallucinations can sometimes be a particularly negative and distressing experience. Moreover, tweets associated with auditory hallucinations were found to have a higher proportionate distribution between the hours of 11pm and 5am, which may be indicative of an effect of time of day on the expression of the phenomena. The number of tweets obtained during data collection and the associated analysis of these tweets suggest the methodology employed was feasible to generate datasets from Twitter regarding the occurrence of psychotic(-like) experiences.

When considering the implementation of this methodology, it is important to discuss the associated strengths and limitations. First, data collected on Twitter may not be directly applicable to the entire population (i.e. individuals experiencing psychosis-related phenomena) due to the exclusion of individuals who do not actively utilise Twitter. It is also feasible that indications of an auditory hallucination may not have been identified due to the limited amount of information that can be included in a 140-character (maximum) Twitter post. In addition, search terms only included words from English, which will have excluded tweets from non-English speaking users. Finally, the time zones in which tweets were posted were unavailable for some of the data collected. Therefore, it was only feasible to investigate the association between time of tweet posting and symptom occurrence for 62% of the tweets collected.

A considerable strength of this research was the interdisciplinary approach taken during the development of the study, data collection and analysis of the findings. The early inclusion of researchers in the field of psychology within the research team ensured that search terms were accurate for the phenomena of interest and the research questions were clinically relevant. The high level of agreement between the researchers for the annotation of tweets was also a significant strength of the study, indicating that classification of tweets was accurate. Finally, the large proportion of related tweets obtained that contained negative sentiments in comparison to positive sentiments which is in line with current opinions that the experience of psychosis-related symptoms is often subjectively unpleasant for the individual involved (Hustig and Hafner, 1990). This suggests that the methodology was successful in identification of psychotic(-like) phenomena.

The initial findings from this exploratory study have significant clinical and practical implications. The data indicates that the methods reported were accurate in identifying psychotic(-like) phenomena. Therefore, it may be possible for researchers to create a large database of tweets containing reference to psychotic(-like) experiences for the identification and analysis of further associated factors. Further work needs to investigate whether posts indicating negative sentiments of auditory hallucinations are comparable across different locations. Future research should also identify potential factors in tweets reporting auditory hallucinations that may predict whether or not a negative sentiment is expressed. The finding that tweets that contained auditory hallucinations had a higher proportionate distribution during nighttime, suggests that time of day may influence the expression of the phenomena. Therefore, future research is planned to investigate expressions of sleep in Twitter users' who report a diagnosis of a psychosis-related disorder.

5. Bibliographical References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

AOIR. (2012). Ethical decision-making and internet research 2.0: recommendations from the aoir ethics working committee. *Retrieved from: http://www.aoir.org/reports/ethics2.pdf*.

Birnbaum, M. L., Rizvi, A. F., Correll, C. U., and Kane, J. M. (2015). Role of social media and the internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. *Early intervention in psychiatry*.

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. (2013). Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.

BPS, B. P. S. (2013). Ethics guidelines for internet-mediated research. *Retrieved from:*

http://www.bps.org.uk/system/files/Public%20files/inf206-guidelines-for-internet-mediated-research.pdf.

Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

Clarke, M. C., Kelleher, I., Clancy, M., and Cannon, M. (2012). Predicting risk and the emergence of schizophrenia. *Psychiatr. Clin. North Am*, 35(3):585–612.

Conway, M. and O'Connor, D. (2016). Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*.

Coppersmith, G., Dredze, M., and Harman, C. (2014). Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. (2015). From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. *NAACL HLT 2015*, page 1.

Gimpel, K., Schneider, N., and O'Connor. (2013). Annotation guidelines for Twitter part-of-speech tagging version 0.3. March.

Gowen, K., Deschaine, M., Gruttadara, D., and Markey, D. (2012). Young adults with mental health conditions and social networking websites: Seeking tools to build community. *Psychiatric Rehabilitation Journal*, 35(3):245.

Haddock, G., McCarron, J., Tarrier, N., and Faragher, E. (1999). Scales to measure dimensions of hallucinations and delusions: the psychotic symptom rating scales (psyrats). *Psychological medicine*, 29(04):879–889.

Hustig, H. H. and Hafner, R. J. (1990). Persistent auditory hallucinations and their relationship to delusions and mood. *The Journal of nervous and mental disease*, 178(4):264–267.

INVOLVE. (2014). Guidance on the use of social media to actively involve people in research. *Retrieved from: http://www.invo.org.uk/wp-content/uploads/2014/11/9982-Social-Media-Guide-WEB.pdf*.

Joseph, A. J., Tandon, N., Yang, L. H., Duckworth, K., Torous, J., Seidman, L. J., and Keshavan, M. S. (2015). # schizophrenia: Use and misuse on twitter. *Schizophrenia research*, 165(2):111–115.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Lincoln, T. M., Lüllmann, E., and Rief, W. (2007). Correlates and long-term consequences of poor insight in patients with schizophrenia. a systematic review. *Schizophrenia bulletin*, 33(6):1324–1342.

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., and Fisher, P. (2007). The Hawthorne Effect: a randomised, controlled trial. *BMC medical research methodology*, 7(1):1.

Meehl, P. E. (1990). Toward an integrated theory of schizotaxia, schizotypy, and schizophrenia. *Journal of Personality Disorders*, 4(1):1.

Nayani, T. H. and David, A. S. (1996). The auditory hallucination: a phenomenological survey. *Psychological medicine*, 26(01):177–189.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., and Schneider, N. (2012). Part-of-speech tagging for Twitter: Word clusters and other advances. *Carnegie Mellon University*.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Pappas, N. and Popescu-Belis, A. (2013). Sentiment analysis of user comments for one-class collaborative filtering over TED talks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 773–776. ACM.

Reavley, N. J. and Pilkington, P. D. (2014). Use of twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. *PeerJ*, 2:e647.

Rijsbergen, K. V. (1979). *Information Retrieval*. Butterworths, London, 2nd edition.

Rivers, C. M. and Lewis, B. L. (2014). Ethical research standards in a world of big data. *F1000Research*, 3.

Shepherd, A., Sanders, C., Doyle, M., and Shaw, J. (2015). Using social media for support and feedback by mental health service users: thematic analysis of a Twitter conversation. *BMC psychiatry*, 15(1):1.

Statista. (2015). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2015 (in millions). *Retrieved from: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/*.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the ACL*, pages 102–107. ACL.

Stone, A. A. and Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, 24(3):236–243.

Tao, K., Abel, F., Hauff, C., Houben, G.-J., and Gadiraju, U. (2013). Groundhog day: near-duplicate detection on Twitter. In *Proceedings of the 22nd international conference on WWW*, pages 1273–1284. International WWW Conferences Steering Committee.

Van Os, J., Linscott, R. J., Myin-Germeys, I., Delespaul, P., and Krabbendam, L. (2009). A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness–persistence–impairment model of psychotic disorder. *Psychological medicine*, 39(02):179–195.

Combining data mining and text mining for detection of early stage dementia: the SAMS framework

Christopher Bull*, Dommy Asfiandy†, Ann Gledson†, Joseph Mellor†, Samuel Couth‡, Gemma Stringer‡, Paul Rayson*, Alistair Sutcliffe*, John Keane†, Xiaojun Zeng†, Alistair Burns‡, Iracema Leroi‡, Clive Ballard§, Pete Sawyer*

*School of Computing and Communications, Lancaster University, UK
†School of Computer Science, University of Manchester, UK
‡Institute of Brain, Behaviour and Mental Health, University of Manchester, UK
§Wolfson Centre for Age-Related Diseases, King's College London, UK

Abstract

In this paper, we describe the open-source SAMS framework whose novelty lies in bringing together both data collection (keystrokes, mouse movements, application pathways) and text collection (email, documents, diaries) and analysis methodologies. The aim of SAMS is to provide a non-invasive method for large scale collection, secure storage, retrieval and analysis of an individual's computer usage for the detection of cognitive decline, and to infer whether this decline is consistent with the early stages of dementia. The framework will allow evaluation and study by medical professionals in which data and textual features can be linked to deficits in cognitive domains that are characteristic of dementia. Having described requirements gathering and ethical concerns in previous papers, here we focus on the implementation of the data and text collection components.

Keywords: dementia, corpus linguistics, natural language processing, data mining

1. Introduction

Dementia is a condition that currently affects around one in six people at the age of 80. Increasing life expectancy means that the number of people who develop dementia will increase. Taking the UK as an example, the number of people living with the condition is predicted to increase from the current figure of 850,000 to over two million by 2051 (Knapp et al., 2007).

Although most forms of dementia such as Alzheimer's Disease are currently irreversible and some are ultimately fatal, obtaining an early diagnosis can help maintain quality of life by treating debilitating side effects, such as depression. Moreover, when improved therapies do eventually become available, it is likely that they will have to be administered before the damage to the brain becomes so severe as to render the therapy ineffective. Currently, diagnosis of dementia or of its harbinger, Mild Cognitive Impairment (MCI), is usually performed using paper-based cognitive tests such as the Montreal Cognitive Assessment (MoCA (Nasreddine et al., 2005)). These are designed to be administered in a clinical setting such as a memory clinic but this can be stressful for the subject and yield poor ecological validity. Worse, many subjects do not refer themselves for a health check until the disease is well advanced. There is therefore a strong interest in developing new techniques for detecting cognitive decline that do not suffer from these disadvantages.

Our work seeks to check for deficits in the same cognitive domains (memory, executive function, motor control and so on) that are tested by the paper tests, using everyday computer tasks as proxies for tasks in the tests (Jimison et al., 2006). The work is based on the simple idea that if someone is finding it increasingly hard to use their computer, then it might be because of change in cognitive function. Many older adults use a computer for (e.g.) home banking, shopping, and keeping in touch with family, so there is an opportunity to exploit the penetration of technology into seniors' homes by developing a non-invasive software tool that helps develop awareness of the users' cognitive health. In our work so far on the SAMS ("Software Architecture for Mental health Self management") project[1], we have focussed on practical problems of how to collect requirements for our monitoring software in order to achieve better acceptance by its end users, as well as the important related ethical concerns for the project (Sutcliffe et al., 2014; Sawyer et al., 2015; Stringer et al., 2015). In this paper, we describe the next stage of the development process. We provide an overview of the SAMS framework for data and text collection created in accordance with these requirements and cross cutting concerns. We also describe a preliminary analysis of initial data mining results.

2. Related Work

To date, little work appears to have been done in the data mining community on analysing sequential patient/user activities to detect the clinical indicators of disease. Seelye et al. (2015) use multiple regression and correlation on mouse movement data from 42 healthy and 20 participants with mild cognitive impairments (MCI) in order to observe that computer mouse movements are a potential indicator of MCI. Much research on mining healthcare data to detect links between health conditions uses association rule mining. This is the technique used by Shin et al. (2010) to mine diagnostic data for patients with essential hypertension and results demonstrate an association between essential hypertension, non-insulin dependent diabetes mellitus, and cerebral infarction. Ohsaki and Sato (2002) use pattern-based time-series data mining for "real medical data that are sequential, numerical and ill-defined", resulting in

[1] http://ucrel.lancaster.ac.uk/sams/

pattern combination rules for testing on chronic hepatitis medical test results. Outside the healthcare domain, sequential pattern mining is used on sequential data representing user activities; for example Pachidi et al. (2014) use this technique to analyse user clickstreams, the clicks made by computer users in order to analyse their use of software.

Turning to related work in the text mining or natural language processing (NLP) area, there is a growing body of research and interest in health-related research in recent years. In addition to this year's RaPID-2016 workshop hosted at LREC, there have been three "Computational Linguistics and Clinical Psychology" workshops held annually at ACL or NAACL since 2014[2], six "International Workshops on Health Text Mining and Information Analysis" held at various locations since 2008[3], and a NIPS 2015 Workshop on Machine Learning in Healthcare[4].

A number of papers have focussed on the notion of "idea density", approximated as the number of verbs, adjectives, adverbs, prepositions, and conjunctions divided by the total number of words, and its decline in old age and Alzheimer's disease (Snowdon et al., 1996; Kemper et al., 2001; Brown et al., 2008). This research used data from what became known as the longitudinal "Nun Study": a collection of autobiographies from the School Sisters of Notre Dame, written when they became nuns (18–32 years old), and cognitive tests much later in life (75–95 years old), and the CPIDR (Computerized Propositional Idea Density Rater) software which implements the metric. Real clinical study data that has been released for replication studies is hard to come by, no doubt due to medical ethics restrictions, so NLP researchers have tended to look elsewhere in their work. Garrard et al. (2005) considers three publications by British writer Iris Murdoch who continued to write novels even after she developed Alzheimer's disease. Garrard et al. (2005) analysed Murdoch's first published work (1954), her last (1995) along with another from 1978, in order to investigate language change using various measures include lexical diversity. Le et al. (2011) and Hirst and Feng (2012) extend this by including a large number of measures and more comparative data: 20 novels for Iris Murdoch, 15 novels for Agatha Christie, and 15 novels for PD James (as control). Using an SVM classifier, they deduce that Agatha Christie also probably developed dementia towards the end of her life.

The Western Collaborative Group Study (WCGS) proves to be a rich source of data for Jarrold et al. (2010), as it provides transcriptions of 15-minute interviews from a 40+ year wide ranging longitudinal study. They use a combination of part-of-speech tagging software, Linguistic Inquiry Word Count (LIWC) (Tausczik and Pennebaker, 2010) and CPIDR to contribute to key measures in a predictive model for Alzheimer's Disease, cognitive impairment and clinical depression. Similar methods were used by Jarrold et al. (2014) on samples from the Western Aphasia Battery to determine dementia subtypes. Finally, a more promising publicly available dataset is the DementiaBank. This is used by Orimaye et al. (2015), who apply machine learning to a combination of skip-gram features, and by Fraser et al. (2015) who employ a much larger (370) set of features to train a machine learning classifier to distinguish participants with Alzheimer's from healthy controls. The DementiaBank[5] clinical dataset consists of interview transcripts of MCI and control participants describing the Cookie-Theft picture component of the Boston Diagnostic Aphasia Examination. Compared to all these elicited interview datasets, the type of text that we are collecting via the SAMS non-invasive approach is significantly different. In contrast to other studies, SAMS will analyse text captured from everyday activities, i.e. email and dairies.

3. SAMS Framework: data/text collection

The SAMS framework is designed to record low-level events (i.e. mouse and keyboard), as well as higher-level contextual information about the Operating System and applications (e.g. drag/drop events, window resizing).

The framework faced a number of challenges, but one of the primary challenges derived from our aim to deploy it on real users' home computers and to collect data as they used their computers to do everyday things. Resources did not permit us to develop a SAMS product line configurable for every type, brand and version of desktop computer, operating system, web browser and desktop application. Guided by information about home computer usage and configurations that we elicited from a superset of the older adults we recruited as SAMS study participants, we took the pragmatic decision to develop SAMS to work on Windows 7, 8, and 10, the MS Office 2007 and later suites of desktop application software, and the Internet Explorer 11 and Chrome web browsers.

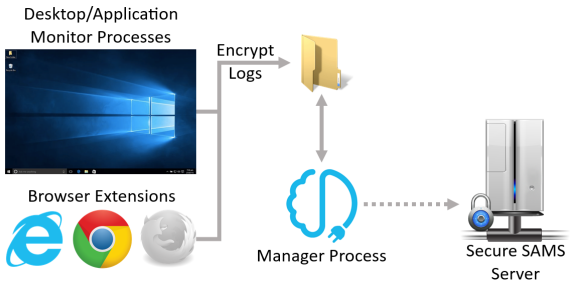


Figure 1: Abstract architecture of SAMS framework

The part of the SAMS framework dedicated to data and text collection is split into several components: the primary desktop logger, web browser logging extensions, and a manager component, see Figure 1. The desktop and web browser loggers are both responsible for data collection and text collection. The browser extensions are required in addition to the desktop logger due to web content in a browser being inaccessible to the desktop logger; a browser extension can have direct access to webpages content. The logs generated by these components are immediately secured using asymmetric encryption. The manager component

[2] http://clpsych.org/

[3] http://louhi2015.limsi.fr

[4] https://sites.google.com/site/nipsmlhc15/

[5] https://talkbank.org/DementiaBank/

is responsible for all user interface elements, for starting, stopping, and pausing the loggers, uploading the encrypted logs to the SAMS server, and updating the SAMS software.

3.1. Desktop Logger Component

The desktop logger records user activities at three levels, as shown in Table 1; level 1: keyboard and mouse, level 2: operating system (e.g. desktop activities), and level 3: application level. All windows events deemed potentially useful for detecting the clinical indicators of dementia are recorded, with the view to further analysis to determine those that are most pertinent. The events that are logged are detailed throughout this section. Activities are captured as a list of time-stamped events using a variety of technologies. Mouse/keyboard level detection utilises an imported .NET library[6]. At the operating system level, native C# .NET libraries[7] [8] are used to detect file system events (files changed, created and renamed) and changes to the clipboard. Microsoft UI Automation events[9] are used to record events such as opening/closing/minimizing/maximizing windows, changes in focus, menus opened/closed and elements selected by the user. At the application level, the Office Primary Interop Assemblies[10] and the Internet Explorer automation object [11] are used to detect events from Microsoft Word, Outlook and Internet Explorer, the three applications considered most relevant for monitoring activities of older adult users.

Further 'high level' events have been developed for the SAMS framework, derived from the low level data events described above. A mouse monitor has been created to read original mouse events, too abundant to be efficiently recorded and too low-level to be of use for later analysis, and aggregates these into mouse drags and mouse 'phases' (time periods between clicks or half second intervals), obtaining more useful information such as time, distance, and screen areas crossed. Similarly, key up and down events are paired and the code and duration are recorded. At the operating system level, mouse drag events are classified where possible into 'move', 'move into', 'resize' and 'scroll' events based on what is known about simultaneous low-level events (for example icon/window position or size changes, scroll and file system events). In addition, UI Automation[12] is used to maintain a map of the desktop including all window and icon positions. This map is used to derive higher level mouse move events, capturing moves into and out of icon or windows and to augment mouse event data with information such as the underlying icon/window name, position and display level.

3.2. Web Browser Extensions

The browser extensions provide the SAMS framework access to webpage content which otherwise is a blackbox to the desktop logger. When a web browser has focus we elect to halt keypress logging in the desktop loggers, allowing the browser extensions to take over that responsibility. The desktop loggers continue to log all other events. This helps avoid the collection of sensitive information such as passwords, as the browser extensions can easily distinguish between password and normal text fields.

The browser extensions work by injecting Javascript (JS) into all webpages. Websites that are loaded with https (secure), and not http, are not monitored; the assumption here is that https webpages are considered private and will likely contain sensitive information or have it entered into them by the participant (e.g. bank details on shopping websites). The injected JS parses the websites DOM, adding numerous event listeners to a wide variety of text and non-text elements detailed in Table 2. The events indicate user interactions and collect text. They can then be analysed later to determine behaviours. When these events fire they are logged to an encrypted file on the user's computer. The Manager component periodically picks up these files, as well as all other SAMS logs, and sends them to the SAMS server. Dynamic webpages, those that create DOM elements after the page has loaded, have a 'Mutation Observer'[13] listen for when new elements are attached to the DOM and adds the event listeners at runtime.

The text-related events that are collected within the browser extensions record a higher fidelity of meta information as well. Upon each participants' interaction with a text element, see 'Text Elements' in Table 2, the selection range (the index of highlighted text) is also recorded. This allows for key presses to easily be reconstructed as full bodies of text later, rather than just individual characters in the log files, and also provides an additional future analysis vector: analysing text editing processes. For example, log entries will indicate if a participant highlights some text and then replaces it.

The browser extensions developed for the SAMS framework focus on the Internet Explorer and Chrome web browsers. In addition to the initial information elicited from the the superset of SAMS participants, Internet Explorer was chosen because it comes pre-installed on Windows computers, and therefore likely to be used by people who favour default setups, and Chrome because it is the most popular browser in 2014/15[14].

All of the main web browsers used on Windows computers

[6] Application and Global Mouse and Keyboard Hooks .Net Library in C#: http://globalmousekeyhook.codeplex.com/

[7] FileSystemWatcher Class: https://msdn.microsoft.com/en-us/library/system.io.filesystemwatcher(v=vs.110).aspx

[8] Clipboard (.NET): https://msdn.microsoft.com/en-us/library/windows/desktop/ms648709(v=vs.85).aspx

[9] Microsoft UI Automation events: https://msdn.microsoft.com/en-us/library/windows/desktop/ee671221(v=vs.85).aspx

[10] Office Primary Interop Assemblies: https://msdn.microsoft.com/en-us/library/15s06t57.aspx

[11] InternetExplorer object: https://msdn.microsoft.com/en-us/library/aa752084(v=vs.85).aspx

[12] Microsoft UI Automation: https://msdn.microsoft.com/en-us/library/windows/desktop/ee684009(v=vs.85).aspx

[13] JS Mutation Observer: https://www.w3.org/TR/dom/#mutation-observers

[14] Web browser statistics: http://www.w3schools.com/browsers/browsers_stats.asp

Table 1: Desktop Logger's captured events.

Level	Sub-level	Description
Level 1	Keyboard	KEYBOARD UP
	Mouse	DESKTOP MOUSE WHEEL MOVE
		MOUSE DOUBLE CLICKED
		MOUSE DRAG PHASE
		MOUSE PHASE COMPLETED (time and mouse movement between clicks)
		MOUSE UP
		MOUSE MOVES IN/OUT OF DESKTOP WINDOWS OR ICONS
Level 2	Clipboard	CLIPBOARD UPDATED
	Drag	DESKTOP DRAG (start/end times and positions etc.)
	File system events	FILE CHANGED
		FILE CREATED / FSW FILE DELETED
		FILE RENAMED
	User interface system events	ELEMENT ADDED/REMOVED FROM/TO A SELECTION
		ELEMENT SELECTED BY USER
		FOCUS CHANGED
		MENU OPENED
		USER INTERFACE OBJECT INVOKED
		WINDOW OPENED/CLOSED
		WINDOW MAXIMIZED/MINIMIZED/TO NORMAL
Level 3	Internet Explorer	OPEN/CLOSE IE WINDOW OR TAB
	Outlook	CHANGE EMAILS SELECTED
		MOVE EMAIL MESSAGE
		START/QUIT OUTLOOK
		READ/REPLY EMAIL
		SEND EMAIL
		SWITCH FOLDERS
	Word	CHANGE TEXT SELECTED
		OPEN/CLOSE/SAVE/SWITCH DOC

(IE, Chrome, Firefox) were found to be capable of allowing their extensions to write files to the user's computer, and therefore enable logging alongside a desktop application counterpart. Microsoft Edge is not included in the SAMS framework because, at the time of writing, extension support for that browser is not yet available.

4. Preliminary Results

A controlled experiment has been completed comparing a healthy control group with a MCI/mild dementia group using a set task composed of GUI-Windows operations, Email-Outlook use, Word processing and Internet searching. Both groups experienced the same conditions in the experiment, and in the longitudinal study recordings are not intrusive and users will not be distracted by the monitoring software. Full consent for the study was given by all participants, following the ethics standards of Manchester University. In these preliminary results, we focus on the data mining aspects only, and will report text mining results in future papers.

The logger outputs time stamped records at the msec level for each user and system generated events at two levels: general from Microsoft UIA tool and SAMS augmented detail of event identities. Event identities are recorded faithfully from all Microsoft browsers but the fidelity of identity varied between web sites with other Internet Browsers.

Preliminary analysis of logs produced by the SAMS tool have shown that even simple frequency analysis of general event types display encouraging trends. For instance, the frequency of individual low-level events associated with mouse movement and keyboard presses have been observed to be different in distribution between healthy and MCI groups.

The difference in distribution amongst the groups for some of these general event-types was found to be significant according to a Mann Whitney U test. Some results can be seen in Table 3. The fact that such differences exist, especially in mouse-movement data, is supported by the work of Seelye et al. (2015).

We are now engaged in a longitudinal study, with 32 installations of the SAMS software running unobtrusively on participants' home computers/laptops. Participants have been recruited that conform to a set of selection criteria based on factors such as their age and home computer ownership and use. Our aim is to discover whether the SAMS software can detect cognitive change within individuals during

Table 2: Web events collected.

	HTML Elements	JS Events
All Elements	(all text and non-text elements, listed below, have this superset of event listeners attached)	click, dblclick, mouseover, contextmenu[a], focusin, focusout
Text Elements	<input type="text">, <input type="search">, <textarea>, <* contenteditable>, <* g_editable>	keydown, keyup, keypress, mouseup, cut, copy, paste, dragstart, dragend
Non-Text Elements	<a>, <button>, <* role="button">, <input>[b], <select>, <img>	mousedown[c], keydown[d], keyup[d], keypress[d]

[a] Could indicate a right-click spelling correction.
[b] Includes password fields, avoiding password collection.
[c] Log event before (e.g) button causes page navigation.
[d] Only for collecting 'Enter' or 'Tab' key event.

the course of the study, informed by what we discover from analysis of the controlled experiment. Ground truth is established by clinical cognitive assessments of each participant at the start, mid-point and end of the study period.
Our current analysis strategy is to apply data mining cluster and pattern analysis algorithms to investigate changes within individuals over time and inter-individual variations with known norms for age/gender cohorts of our senior participants (range 65–78 years). Given these reassuring findings, future work includes sequence analysis such as learning Markov models or using SPADE-like algorithms, which have been applied to finding temporal patterns in web-log data (Demiriz, 2002), to discover richer interaction of low-level events over time capable of identifying signs of MCI. Sequence mining will be used to identify atypical user behaviour and errors which might indicate cognitive problems linked to MCI and early dementia. Integration of evidence from data mining activity patterns, sequences of computer operation, and text analysis metrics will be investigated using Bayesian nets to implement a 'diagnostic' model that traces measures derived from data and text mining to cognitive indicators which are associated with MCI. The challenge we face is finding a weak signal indicative of disease in noisy data where variations might be caused by interruptions, changes in user mood, or many environment factors.

5. Conclusion and Future Work

We have developed a novel system architecture that not only logs keyboard, mouse, and contextual environment/application data but also interprets these events as user behaviours. This, combined with text capture from email and diary entries, is input into data and text mining tools, so we can analyse early signs of dementia by combining evidence from many measures across time. The SAMS project is now entering the analysis phase and we await the end of our longitudinal study. We have selected a set of potential text mining features from the related work described in Section 2. These are being implemented around the already existing Wmatrix tag wizard pipeline for part-of-speech and semantic tagging (Rayson, 2008), along with variant detection using VARD (Baron and Rayson, 2008), and the extraction of type and token frequency data at three levels: lexical, grammatical and semantic tags. The SAMS software framework will be available open source from Github[15] and the project website. In future projects, we intend to apply the SAMS architecture for health monitoring in a wide range of domains including mental health as well as dementia.

6. Acknowledgements

The work described in this paper is funded by the Engineering and Physical Sciences Research Council (EPSRC) in the UK, within the Software Architecture for Mental Health Self-Management (SAMS) project, references EP/K015796/1, EP/K015761/1 and EP/K015826/1.

7. Bibliographical References

Baron, A. and Rayson, P. (2008). VARD2: a tool for dealing with spelling variation in historical corpora. Postgraduate Conference in Corpus Linguistics, Aston University, Birmingham, UK.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., and Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2):540–545.

Demiriz, A. (2002). webSPADE: a parallel sequence mining algorithm to analyze web log data. In *Proceedings of the International Conference on Data Mining (ICDM '02)*, pages 755–758. IEEE.

Fraser, K. C., Meltzer, J., and Rudzicz, F. (2015). Linguistic features identify Alzheimer's Disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Garrard, P., Maloney, L. M., Hodges, J. R., and Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128(2):250–260.

Hirst, G. and Feng, V. W. (2012). Changes in style in authors with Alzheimer's disease. *English Studies*, 93(3):357–370.

Jarrold, W. L., Peintner, B., Yeh, E., Krasnow, R., Javitz, H. S., and Swan, G. E., (2010). *Brain Informatics: International Conference, BI 2010, Toronto, ON, Canada, August 28-30, 2010. Proceedings*, chapter Language Analytics for Assessing Brain Health: Cognitive Impairment, Depression and Pre-symptomatic Alzheimer's Disease, pages 299–307. Springer Berlin Heidelberg, Berlin, Heidelberg.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., and Ogar, J. (2014). Aided diagnosis of dementia type through computer-based

[15] https://github.com/UCREL

Table 3: Frequencies of event types per user. The p-value is for a Mann Whitney U test between HC and MCI groups. We can see that the HC group press the keyboard more often. The MCI group double-click much more frequently.

Event type	HC event counts	MCI event counts	p-value
KEYBOARD_UP	546, 619, 458, 926, 445, 508, 406, 683, 849, 244, 482, 280, 718, 628, 350, 441, 599, 460, 543, 439	253, 214, 595, 402, 452, 554, 364, 206, 410, 289, 229	0.0036
MOUSE_DBLCLICKED	0, 0, 12, 0, 8, 0, 0, 10, 0, 0, 8, 0, 0, 0, 0, 0, 0, 0, 0, 4	33, 18, 59, 15, 7, 0, 75, 12, 27, 0, 18	0.0002

analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Jimison, H., Jessey, N., McKanna, J., Zitzelberger, T., and Kaye, J. (2006). Monitoring computer interactions to detect early cognitive impairment in elders. In *1st Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare, 2006. D2H2.*, pages 75–78. IEEE.

Kemper, S., Greiner, L. H., Marquis, J. G., Prenovost, K., and Mitzner, T. L. (2001). Language decline across the life span: Findings from the Nun Study. *Psychology and Aging*, 16(2):227–239.

Knapp, M., Prince, M., Albanese, E., Banerjee, S., Dhanasiri, S., Fernandez, J., Ferri, C., Snell, T., and Stewart, R. (2007). Dementia UK: report to the Alzheimer's Society. *King's College London and London School of Economics and Political Science*.

Le, X., Lancashire, I., Hirst, G., and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and Linguistic Computing*, 26(4):435–461.

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Ohsaki, M. and Sato, Y. (2002). A rule discovery support system for sequential medical data, in the case study of a chronic hepatitis dataset. In *Proceedings of the International Workshop on Active Mining (AM '02) in International Conference on Data Mining (ICDM '02)*, pages 97–102. IEEE.

Orimaye, S. O., Tai, K. Y., Wong, J. S., and Wong, C. P. (2015). Learning linguistic biomarkers for predicting mild cognitive impairment using compound skip-grams. In *Proceedings of the 2015 NIPS Workshop on Machine Learning in Healthcare (MLHC), Montreal, Canada*.

Pachidi, S., Spruit, M., and Van De Weerd, I. (2014). Understanding users' behavior with software operation data mining. *Computers in Human Behavior*, 30:583–594.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549.

Sawyer, P., Sutcliffe, A., Rayson, P., and Bull, C. (2015). Dementia and social sustainability: challenges for software engineering. In *37th International Conference on Software Engineering (ICSE '15)*, Florence, Italy. IEEE.

Seelye, A., Hagler, S., Mattek, N., Howieson, D. B., Wild, K., Dodge, H. H., and Kaye, J. A. (2015). Computer mouse movement patterns: A potential marker of mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(4):472–480.

Shin, A. M., Lee, I. H., Lee, G. H., Park, H. J., Park, H. S., Yoon, K. I., Lee, J. J., and Kim, Y. N. (2010). Diagnostic Analysis of Patients with Essential Hypertension Using Association Rule Mining. *Healthcare Informatics Research*, 16(2):77–81.

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

Stringer, G., Sawyer, P., Sutcliffe, A., and Leroi, I. (2015). From Click to Cognition. In Davide Bruno, editor, *The Preservation of Memory*, chapter 5, pages 93–103. Psychology Press.

Sutcliffe, A., Rayson, P., Bull, C., and Sawyer, P. (2014). Discovering Affect-Laden Requirements to Achieve System Acceptance. In *Proceedings of the 22nd IEEE International Requirements Engineering Conference (RE'14)*, pages 173–182, Karlskrona, Sweden. IEEE.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.