

# Executive Function & Semantic Memory Impairments in Alzheimer's Disease

Investigating the Decline of Executive Function and Semantic Memory in  
Alzheimer's Disease through Computer-Supported Qualitative Analysis of  
Semantic Verbal Fluency and its Applications in Clinical Decision Support

Dissertation

zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

der Fakultät HW

Bereich Empirische Humanwissenschaften

der Universität des Saarlandes

vorgelegt von

Johannes Tröger, M.Sc.

aus Nürnberg

Saarbrücken 2021

<b>Dekan</b>	Prof. Dr. Jörn Sparfeldt
<b>Berichterstatterin 1</b>	Prof. Dr. Jutta Kray
<b>Berichterstatterin 2</b>	Prof. Dr. Tanja Michael
<b>Tag der Disputation:</b>	08.11.2021

**Author Note**

This thesis is based on manuscripts that have been published in several scientific journals and conferences from the field of neuropsychology and applied computer science in healthcare.

## **Acknowledgements**

First of all, I would like to thank my supervisor Prof. Dr. Jutta Kray for the great support she gave me during the whole thesis project. Thank you for accepting the challenge of supervising my project which might have been slightly outside the traditional psychological research comfort zone.

I would also like to thank Prof. Dr. Tanja Michael for serving as reviewer of my thesis.

I am grateful to The German Research Center for Artificial Intelligence (DFKI) for being such a great environment for innovative and multidisciplinary applied research. I would also thank all the great DFKI colleagues that have been of help during my research. In particular I would like to thank:

- Dr. Jan Alexandersson as my group leader and the person that onboarded me to DFKI. Thanks for entrusting me with the freedom to define and pursue my own personal research agenda that eventually got me here.
- Dr. Philipp Müller who was great help reviewing and improving this document. Always remembering that it was with you I conceptualized and succeeded my first own research project.
- Hali Lindsay, who joined me halfway but immediately had an impact. You have a great career lying ahead of you!
- Dr. Andrey Girenko who not only always had a Russian saying but also a very motivating one: "You work three years for the PhD but then the PhD works for you the rest of your life."

I also want to thank all the wonderful people that I did research with, coming from other academic or industrial research organizations. In particular I would like to thank:

- Nicklas Linz, who was part of this research from the very beginning and without whom the applied and multidisciplinary character of this work would have never been possible. Thanks for helping to make the difference!
- Dr. Alexandra König, who was pioneering computer-supported speech analysis in AD long before I entered the field. Belief it or not: if it wouldn't have been you and your research leading the way I would have not started all this work in the beginning.

- PD Dr. Jessica Peter, who served as stronghold for all the clinical neuropsychology aspects around this work. Thanks for the guidance and teaching me that extra scrutiny and precision when it comes to publishing.

Finally, I thank my family and friends for supporting me in all non-academic aspects of life. I am particularly grateful to my parents who raised me as a critical mind, a creative problem solver and someone striving for excellence. I want to thank Elodie who is the love of my life and my two kids Gustave and Suzanne who always reminded me during this time that science is not the only important thing in life.

**General abstract**

Alzheimer's Disease (AD) has a huge impact on an ever-aging society in highly developed industrialized countries such as the EU member states: according to the World Alzheimer's Association the number one risk factor for AD is age. AD patients suffer from neurodegenerative processes driving cognitive decline which eventually results in the loss of patients' ability of independent living. Episodic memory impairment is the most prominent cognitive symptom of AD in its clinical stage. In addition, also executive function and semantic memory impairments significantly affect activities of daily living and are discussed as important cognitive symptoms during prodromal as well as acute clinical stages of AD. Most of the research on semantic memory impairments in AD draws evidence from the Semantic Verbal Fluency (SVF) task which evidentially also places high demands on the executive function level. At the same time, the SVF is one of the most-applied routine assessments in clinical neuropsychology especially in the diagnosis of AD. Therefore, the SVF is a prime task to study semantic memory and executive function impairment side-by-side and draw conclusions about their parallel or successive impairments across the clinical trajectory of AD.

To effectively investigate semantic memory and executive function processes in the SVF, novel computational measures have been proposed that tap into data-driven semantic as well as temporal metrics scoring an SVF performance on the item-level. With a better and more differentiated understanding of AD-related executive function and semantic memory impairments in the SVF, the SVF can grow from a well-established screening into a more precise diagnostic tool for early AD. As the SVF is one of the most-applied easy-to-use and low-burden neurocognitive assessments in AD, such advancements have a direct impact on clinical practice as well. For the last decades huge efforts have been put on the discovery of disease-modifying compounds responding to specific AD biomarker-related cognitive decline characteristics. However, as most pharmaceutical trials failed, the focus has shifted towards population-wide early screening with cost-effective and scalable cognitive tests representing an effective mid-term strategy. Computer-supported SVF analysis responds to this demand.

This thesis pursues a two-fold objective: (1) improve our understanding of the progressive executive function and semantic memory impairments and their interplay in clinical AD as measured by the SVF and (2) harness those insights for applied early and specific AD screening.

To achieve both objectives, this thesis comprises work on subjects from different clinical stages of AD (Healthy Aging, amnesic Mild Cognitive Impairment—*aMCI*, and AD dementia) and in different languages (German & French). All results are based on SVF speech data generated either as a one-time assessment or a repeated within-participant testing. From these SVF speech samples, qualitative markers are extracted with different amount of computational support (ranging from manual processing of speech to fully automated evaluation).

The results indicate, that semantic memory is structurally affected from an early clinical—amnesic Mild Cognitive Impairment (*aMCI*)—stage on and is even more affected in the later acute dementia stage. The semantic memory impairment in AD is particularly worsened through the patients' inability to compensate by engaging executive functions. Hence, over the course of the disease, hampered executive functioning and therefore the inability to compensate for corrupt semantic memory structures might be the main driver of later-stage AD patients' notably poor cognitive performance. These insights generated on the SVF alone are only made possible through computer-supported qualitative analysis on an item-per-item level which leads the way towards potential applications in clinical decision support. The more fine-grained qualitative analysis of the SVF is clinically valuable for AD diagnosis and screening but very time-consuming if performed manually. This thesis shows though that automatic analysis pipelines can reliably and validly generate this diagnostic information from the SVF. Automatic transcription of speech plus automatic extraction of the novel qualitative SVF features result in clinical interpretation comparable to manual transcripts and improved diagnostic decision support simulated through machine learning classification experiments. This indicates that the computer-supported SVF could ultimately be used for cost-effective fully automated early clinical AD screening.

This thesis advances current AD research in a two-fold manner. First it improves the understanding of the decline of executive function and semantic memory in AD as measured through computational qualitative analysis of the SVF. Secondly, this thesis embeds these theoretical advances into practical clinical decision support concepts that help screen population-wide and cost-effective for early-stage AD.

*Keywords:* AD, MCI, *aMCI*, Semantic Memory, Executive Function, Automatic Qualitative Analysis, Semantic Verbal Fluency, Speech Analysis, Clinical Decision Support

## **Zusammenfassung**

Die Alzheimer-Krankheit (AD) stellt eine enorme Herausforderung für die immer älter werdende Gesellschaft in hochentwickelten Industrieländern wie den EU-Mitgliedsstaaten dar. Nach Angaben der World Alzheimer's Association ist der größte Risikofaktor für AD das Alter. Alzheimer-Patienten leiden unter neurodegenerativen Prozessen, die kognitiven Abbau verursachen und schließlich dazu führen, dass Patienten nicht länger selbstbestimmt leben können. Die Beeinträchtigung des episodischen Gedächtnisses ist das prominenteste kognitive Symptom von AD im klinischen Stadium. Darüber hinaus führen auch Störungen der Exekutivfunktionen sowie der semantischen Gedächtnisleistung zu erheblichen Einschränkungen bei Aktivitäten des täglichen Lebens und werden als wichtige kognitive Symptome sowohl im Prodromal- als auch im akuten klinischen Stadium von AD diskutiert. Der Großteil der Forschung zu semantischen Gedächtnisbeeinträchtigungen bei AD stützt sich auf Ergebnisse aus dem Semantic Verbal Fluency Tests (SVF), der auch die Exekutivfunktionen stark fordert. In der Praxis ist die SVF eines der am häufigsten eingesetzten Routine-Assessments in der klinischen Neuropsychologie, insbesondere bei der Diagnose von AD. Daher ist die SVF eine erstklassige Aufgabe, um die Beeinträchtigung des semantischen Gedächtnisses und der exekutiven Funktionen Seite an Seite zu untersuchen und Rückschlüsse auf ihre parallelen oder sukzessiven Beeinträchtigungen im klinischen Verlauf von AD zu ziehen.

Um semantische Gedächtnis- und Exekutivfunktionsprozesse in der SVF effektiv zu untersuchen, wurden jüngst neuartige computergestützte Verfahren vorgeschlagen, die sowohl datengetriebene semantische als auch temporäre Maße nutzen, die eine SVF-Leistung auf Item-Ebene bewerten. Mit einem besseren und differenzierteren Verständnis von AD-bedingten Beeinträchtigungen der Exekutivfunktionen und des semantischen Gedächtnisses in der SVF kann sich die SVF von einem gut etablierten Screening zu einem präziseren Diagnoseinstrument für frühe AD entwickeln. Da die SVF eines der am häufigsten angewandten, einfach zu handhabenden und wenig belastenden neurokognitiven Assessments bei AD ist, haben solche Fortschritte auch einen direkten Einfluss auf die klinische Praxis. In den letzten Jahrzehnten wurden enorme Anstrengungen unternommen, um krankheitsmodifizierende Substanzen zu finden, die auf spezifische, mit AD-Biomarkern verbundene Merkmale des kognitiven Abbaus reagieren. Da jedoch die meisten

pharmazeutischen Studien in jüngster Vergangenheit fehlgeschlagen sind, wird heute als mittelfristige Strategie bevölkerungsweite Früherkennung mit kostengünstigen und skalierbaren kognitiven Tests gefordert. Die computergestützte SVF-Analyse ist eine Antwort auf diese Forderung.

Diese Arbeit verfolgt deshalb zwei Ziele: (1) Verbesserung des Verständnisses der fortschreitenden Beeinträchtigungen der Exekutivfunktionen und des semantischen Gedächtnisses und ihres Zusammenspiels bei klinischer AD, gemessen durch die SVF, und (2) Nutzung dieser Erkenntnisse für angewandte AD-Früherkennung.

Um beide Ziele zu erreichen, umfasst diese Thesis Forschung mit Probanden aus verschiedenen klinischen AD Stadien (gesundes Altern, amnestisches Mild Cognitive Impairment-aMCI, und AD-Demenz) und in verschiedenen Sprachen (Deutsch & Französisch). Alle Ergebnisse basieren auf SVF Sprachdaten, erhoben im Querschnittsdesign oder als wiederholte Testung in einem Längsschnittsdesign. Aus diesen SVF-Sprachproben werden mit unterschiedlicher rechnerischer Unterstützung qualitative Marker extrahiert (von manueller Verarbeitung der Sprache bis hin zu vollautomatischer Auswertung).

Die Ergebnisse zeigen, dass das semantische Gedächtnis bereits im frühen aMCI Stadium strukturell beeinträchtigt ist und im späteren akuten Demenzstadium noch stärker betroffen ist. Die strukturelle Beeinträchtigung des semantischen Gedächtnisses bei Alzheimer wird insbesondere dadurch verschlimmert, dass die Patienten nicht in der Lage sind, dies durch den Einsatz exekutiver Funktionen zu kompensieren. Daher könnten im Verlauf der Erkrankung eingeschränkte Exekutivfunktionen und damit die Unfähigkeit, degenerierte semantische Gedächtnisstrukturen zu kompensieren, die Hauptursache für die auffallend schlechten kognitiven Leistungen von AD-Patienten im Akutstadium sein. Diese Erkenntnisse basierend auf der SVF alleine werden erst durch die computergestützte qualitative Analyse auf Item-per-Item-Ebene möglich und weisen den Weg zu möglichen Anwendungen in der klinischen Entscheidungsunterstützung. Die feinkörnigere qualitative Analyse der SVF ist klinisch wertvoll für die AD-Diagnose und das Screening, aber sehr zeitaufwändig, wenn sie manuell durchgeführt wird. Diese Arbeit zeigt jedoch, dass automatische Analysepipelines diese diagnostischen Informationen zuverlässig und valide aus der SVF generieren können. Die automatische Transkription von Sprache plus die automatische Extraktion der neuartigen qualitativen SVF-Merkmale führen zu einer klinischen

Interpretation, die mit manuellen Analysen vergleichbar ist. Diese Verarbeitung führt auch zu einer verbesserten diagnostischen Entscheidungsunterstützung, die durch Klassifikationsexperimente mit maschinellem Lernen simuliert wurde. Dies deutet darauf hin, dass die computergestützte SVF letztendlich für ein kostengünstiges vollautomatisches klinisches AD-Frühscreening eingesetzt werden könnte.

Diese Arbeit bringt die aktuelle AD-Forschung auf zweifache Weise voran. Erstens verbessert sie unser Verständnis der kognitiven Einschränkungen im Bereich der Exekutivfunktionen und des semantischen Gedächtnisses bei AD, gemessen durch die computergestützte qualitative Analyse der SVF. Zweitens bündelt diese Arbeit diese theoretischen Fortschritte in ein praktisches Konzept zur klinischen Entscheidungsunterstützung ein, das zukünftig ein bevölkerungsweites und kosteneffektives Screening für AD im Frühstadium ermöglichen könnte.

**Outline**

**2 BACKGROUND ..... 8**

2.1 AD AS DIAGNOSTIC ENTITY AND ITS STAGES ..... 8

2.2 AD DEMENTIA'S SPECTRUM OF COGNITIVE FUNCTION DECLINE ..... 13

    2.2.1 *Semantic Memory*..... 13

    2.2.2 *Executive Function* ..... 14

2.3 QUALITATIVE ANALYSIS OF THE SEMANTIC VERBAL FLUENCY TASK ..... 16

    2.3.1 *Semantic Qualitative Analysis*..... 18

    2.3.2 *Temporal Qualitative Analysis*..... 21

    2.3.3 *Qualitative Analysis of SVF in AD dementia*..... 23

**3 CHARACTERIZING THE IMPAIRMENT OF EXECUTIVE FUNCTION AND SEMANTIC MEMORY .....25**

3.1 EXPLOITATION VS. EXPLORATION—COMPUTATIONAL TEMPORAL AND SEMANTIC ANALYSIS EXPLAINS SEMANTIC VERBAL FLUENCY IMPAIRMENT IN ALZHEIMER'S DISEASE ..... 27

    3.1.1 *Introduction* ..... 28

    3.1.2 *Methods*..... 31

    3.1.3 *Results*..... 39

    3.1.4 *Discussion* ..... 43

    3.1.5 *Limitations* ..... 46

    3.1.6 *Conclusion*..... 47

3.2 PATIENTS WITH AMNESTIC MCI FAIL TO ADAPT EXECUTIVE CONTROL WHEN REPEATEDLY TESTED WITH SEMANTIC VERBAL FLUENCY TASKS ..... 48

    3.2.1 *Introduction* ..... 49

    3.2.2 *Methods*..... 51

    3.2.3 *Results*..... 55

    3.2.4 *Discussion* ..... 57

    3.2.5 *Limitations* ..... 60

3.3 CHAPTER CONCLUSION ..... 61

**4 IMPLICATIONS FOR CLINICAL DECISION MAKING .....63**

4.1 FULLY AUTOMATIC SPEECH-BASED ANALYSIS OF THE SEMANTIC VERBAL FLUENCY TASK..... 65

EXECUTIVE FUNCTION & SEMANTIC MEMORY IMPAIRMENTS IN ALZHEIMER'S DISEASE

4.1.1 *Introduction* ..... 65

4.1.2 *Methods*..... 69

4.1.3 *Results*..... 72

4.1.4 *Discussion* ..... 75

4.1.5 *Conclusion*..... 77

4.2 TELEPHONE-BASED DEMENTIA SCREENING I: AUTOMATED SEMANTIC VERBAL FLUENCY ASSESSMENT ..... 79

4.2.1 *Introduction* ..... 80

4.2.2 *Related Work* ..... 82

4.2.3 *Methods*..... 87

4.2.4 *Results*..... 91

4.2.5 *Discussion* ..... 92

4.2.6 *Conclusion*..... 93

4.3 CHAPTER CONCLUSION ..... 94

**5 OVERARCHING DISCUSSION AND CONCLUSION .....96**

5.1 NEUROCOGNITIVE AD PROFILES FROM SVF—EMBEDDED INTO NEUROSCIENCE ..... 97

5.1.1 *Distinguishing AD-Related Executive Function & Semantic Memory Impairment*..... 97

5.1.2 *Computer-Supported Qualitative SVF Markers and Classic AD Neuroscience* ..... 99

5.2 THE SVF AS A COST-EFFECTIVE AND SCALABLE ASSESSMENT FOR AD ..... 102

5.2.1 *Automatic Qualitative Analysis of the SVF for Cost-effective AD Screening* ..... 103

5.2.2 *Feasibility of an SVF-Based Scalable AD Screening Approach* ..... 109

5.3 OVERALL CONCLUSION ..... 112

**6 OUTLOOK .....114**

**BIBLIOGRAPHY .....120**

Investigating the Decline of Executive Function and Semantic Memory in AD  
through Computer-Supported Qualitative Analysis of Semantic Verbal Fluency  
and its Applications in Clinical Decision Support

Alzheimer's Disease (AD) has a huge impact on our ageing society worldwide and especially in highly developed industrialized countries such as the EU member states. The socio-economic impact on our global society is estimated to grow up to more than \$1 trillion US dollars in 2030 (Wimo et al., 2017), including direct costs such as formal medical care as well as indirect costs such as costs of informal care or even intangible costs through reduced quality of life in patients as well as care givers (El-Hayek et al., 2019). According to the World Alzheimer's Association the number one risk factor for AD is age (Abbott, 2011). Especially for an ever-ageing population in Europe and developed countries, AD is set to be the biggest 'killer' of the 21<sup>st</sup> century (NHS long term plan; Alderwick & Dixon, 2019).

AD patients suffer from neurodegenerative processes driving cognitive decline which eventually results in the loss of patients' ability of independent living (for more comprehensive information on the course of the disease and etiology see <https://www.alz.org/>). It has been shown that characteristics of the cognitive decline depend on the stage of the disease, individual differences as well as co-morbidities. While it is proven that episodic memory deficits are the hallmark cognitive symptom of AD (Collie & Maruff, 2000), impairments of other cognitive functions especially executive function and semantic (long-term) memory and their AD clinical stage-related characteristics are subject to ongoing research (Guarino et al., 2019; Verma & Howard, 2012). Importantly, both memory and executive function have been shown to strongly relate to a patient's declining abilities in daily living (Tomaszewski Farias et al., 2009) which underpins the overall meaning of those cognitive functions for AD patients' health. Recent findings have shown, that semantic memory impairments can be found as early as episodic memory changes (Verma & Howard, 2012) at an early clinical and even pre-clinical stage of AD. The pattern of semantic memory impairment in early clinical stages of AD parallels the pattern in later clinical stages (dementia). However, the evidence for executive functioning impairment over the course of AD draws a less conclusive picture with reviews pointing more towards the later clinical stages of AD dementia (Guarino et al., 2019) and other reviews finding executive functioning to also be systematically impaired at a prodromal early clinical stage (Crowell, Luis, Vanderploeg, Schinka, & Mullan, 2002). Overall there is ongoing

research how the AD-related cognitive decline in memory—including semantic memory—is matched or even accelerated by executive function impairment (Buckner, 2004).

The great majority of evidence for AD-related semantic memory impairment stems from one task: the Semantic Verbal Fluency (SVF) (Verma & Howard, 2012). However, from a clinical assessment perspective, it has been argued that the well-established SVF impairment in AD—and dementia in general—draws upon the impairment of both, executive function as well as semantic memory (Amunts, Camilleri, Eickhoff, Heim, & Weis, 2020; Shao, Janse, Visser, & Meyer, 2014). This might be why the SVF is traditionally one of the most sensitive tasks to efficiently detect dementia, but at the same time one of the most difficult ones to use for the differential diagnosis of impaired cognitive functions (Shao et al., 2014). This highlights that the SVF is a prime task to study at the same time semantic memory and executive function impairment side-by-side across different clinical AD stages and draw conclusions about their parallel or successive impairments from that. To achieve this however, different neurocognitive processes in the SVF have to be modelled more carefully on a semantic as well as temporal level (Rohrer, Wixted, Salmon, & Butters, 1995; A. K. Troyer, Moscovitch, & Winocur, 1997). Current computational approaches tap into the same research gap providing objective yet reliable as well as cost-effective methods to model semantic production strategies in the SVF (Linz, Tröger, Alexandersson, & König, 2017; Linz, Tröger, Alexandersson, Wolters, et al., 2017). Harnessing latest computational qualitative analysis schemes for the SVF, the SVF becomes more informative for semantic memory as well as executive function impairments in AD and how they are related amongst each other. Once there is a better understanding about executive function and semantic memory impairment in the SVF and across clinical AD stages, the SVF, today a well-established early screening, can grow into a more precise AD-related yet ever-sensitive diagnostic tool.

Over the last decade, pharmacological research has put huge efforts on the development of targeted compounds responding to specific AD biomarker-related cognitive decline characteristics. But since many clinical trials have failed to find a cure, a conceptual shift has occurred considering Alzheimer's disease (AD) as a continuum for which early intervention may offer the best chance of therapeutic success (Dubois et al., 2016). Recent research has shown that prevention at prodromal stages (Mild Cognitive Impairment, MCI) show promising results and are more likely to be effective (Sindi, Mangialasche, & Kivipelto,

2015). Efficiently identifying people that are at this prodromal but clinical stage of AD remains a renowned challenge for our healthcare system (Dubois et al., 2016); this is partly due to the fact that patients don't show strong clinical symptoms and therefore don't consult a specialist. Therefore, applied research should focus on innovative concepts for detection of cognitive AD-related symptoms that could be used as population wide early yet specific screening tools. Within this scope, low-effort and scalable computerized cognitive tests represent the most effective mid-term strategy (Snyder et al., 2014).

Answering the above-mentioned challenges, this thesis makes a successive multi-step contribution to the theoretical understanding of AD's cognitive function impairments as well as how this improved understanding and modeling of neurocognitive impairments in the SVF can be used to improve diagnostic screening procedures used in clinical practice. In a first step, this thesis contributes to the current discussion about cognitive function impairments across the disease trajectory from a clinical prodromal AD stage (amnesic Mild Cognitive Impairment) towards dementia by better characterizing the decline of executive function and semantic memory in AD through computer-supported qualitative analysis of the SVF. Next, this thesis shows how the neurocognitive functions-related additional information generated from the SVF can help to better identify AD patients at a prodromal stage and also use this for more scalable screening solutions. This not only generates new insights into the impairment of semantic memory and executive function across the clinical stages of AD via the SVF but at the same generate results that have a more direct impact on clinical practice by advancing cost-effective and scalable AD screening.

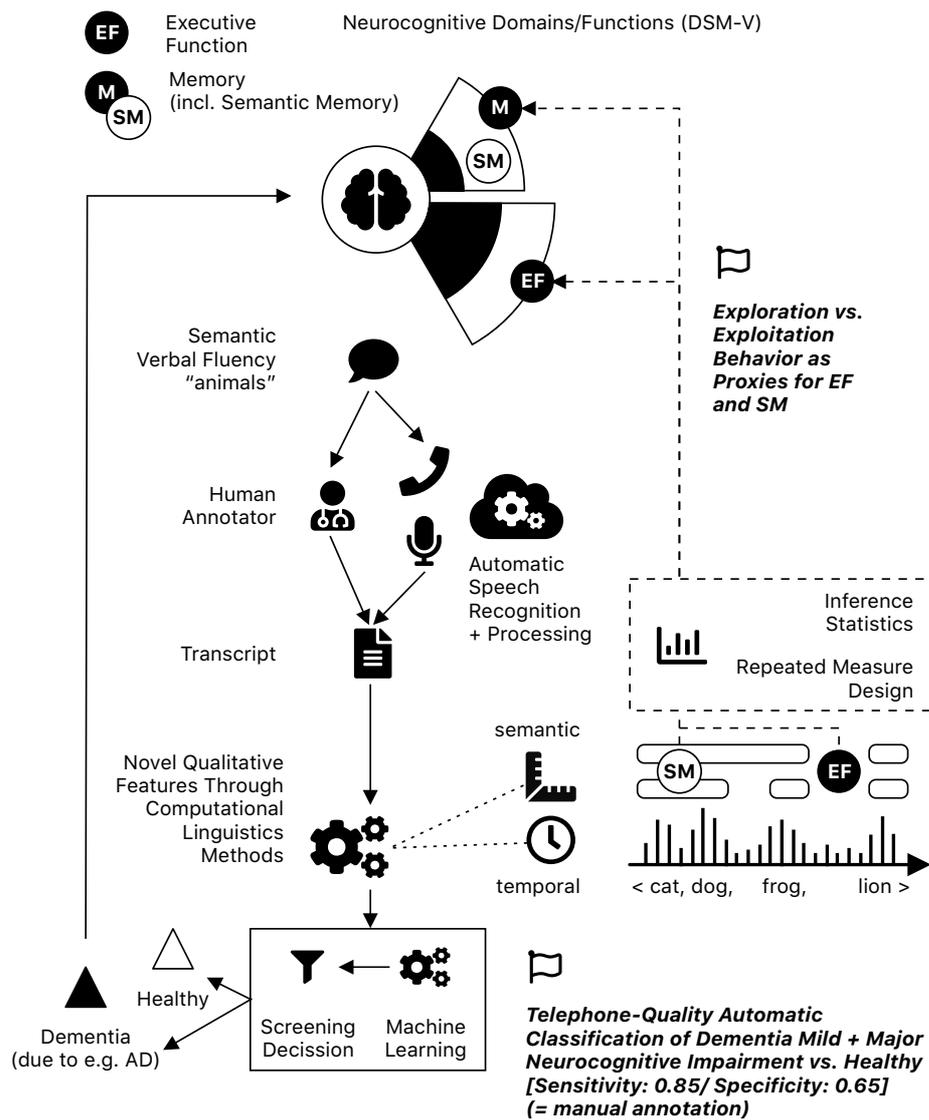


Figure 1: Visual abstract summarizing the major reasoning of this thesis.

This thesis is structured as follows: Following this introduction, the second chapter will provide background for AD as diagnostic entity, its spectrum of cognitive function decline, as well as an introduction to the semantic verbal fluency as one of the most commonly applied assessments in AD diagnosis. The third chapter will focus on the impairment of semantic memory and executive function in both AD clinical stages (prodromal AD/ aMCI & dementia) as assessed by the Semantic Verbal Fluency (SVF). This third chapter contains two articles that have been published in neuropsychological journals. The fourth chapter will focus on the implications of afore-described neurocognitive insights for clinical decision making, such as scalable screening for prodromal AD; this chapter also contains work that has been published in an applied clinical research journal on geriatric cognitive disorders as well as on a

conference for pervasive healthcare technologies. The fifth chapter will provide an overarching discussion and conclusion and the sixth chapter an outlook for adjacent future work (for an overview of publications included in this thesis see also Table 1).

Table 1: Publications included in this thesis with chapter correspondence.

Chapter	Publication
3.1	Tröger, J., Linz, N., König, A., Robert, P., Alexandersson, J., Peter, J., & Kray, J. (2019). Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in Alzheimer's disease. <i>Neuropsychologia</i> , <i>131</i> , 53-61.
3.2	Tröger, J., Lindsay, H., Mina, M., Linz, N., Klöppel, S., Kray, J., & Peter, J. (2021). Patients with amnesic MCI fail to adapt executive control when repeatedly tested with semantic verbal fluency tasks. <i>Journal of the International Neuropsychological Society</i> , accepted.
4.1	König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., & Robert, P. (2018). Fully automatic speech-based analysis of the semantic verbal fluency task. <i>Dementia and geriatric cognitive disorders</i> , <i>45</i> (3-4), 198-209.
4.2	Tröger, J., Linz, N., König, A., Robert, P., & Alexandersson, J. (2018, May). Telephone-Based Dementia Screening I: Automated Semantic Verbal Fluency Assessment. In <i>Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare</i> (pp. 59-66). — Best Paper Award

### Key theses and contributions of this work

Overall this thesis makes five major contributions from different scientific domains: neuropsychology, geriatric psychiatry as well as applied natural language processing. The five major contributions and their key statements are as follows:

1. In AD, semantic memory is structurally affected from an early clinical, aMCI, stage on and is even more affected in the later clinical dementia stage.
  - a. Chapter 3.1 shows heavily impaired semantic memory exploration in the SVF in AD dementia which probably reflects an interplay of semantic memory impairment and also lacking executive function for compensation. However, patients also show working semantic memory exploitation processes on an absolute lower performance level but fundamentally indicating a working semantic memory retrieval in the most frequently used parts of its structure.
  - b. Chapter 3.2 shows semantic memory structure loss in the first of three repeated assessments already at a prodromal amnesic Mild Cognitive Impairment (aMCI) stage. Also showing higher mean word frequency of the SVF

responses indicating more frequently used words which should be more accessible even in a structurally corrupted associative semantic memory.

2. The semantic memory impairment in AD is particularly worsened through the patients' inability to compensate by engaging executive functions.
  - a. Chapter 3.1 shows that the exploration of the semantic space through hampered switching behavior is the main driver behind the SVF impairment in the dementia group. There was impaired switching behavior for all modalities (impaired switching as temporally defined, semantically and traditionally taxonomic); this is in line with previous literature interpreted as mainly an executive control/ function problem.
  - b. Chapter 3.2 shows that the SVF impairment prevails even when patients are repeatedly confronted with the same task, which should normally result in better SVF scores through improved executive control strategies compensating an underlying semantic memory problem or a novelty effect. At the same time, healthy controls do improve significantly over the repeated assessment.

Hence, over the course of the disease, hampered executive functioning is found to be the main driver of later-stage AD patients' notably poor cognitive performance in the SVF. However, it is probably structurally preceded by semantic memory impairments. In the later clinical dementia stage of AD the interplay of both and especially the inability to compensate through executive function resources or cognitive control might be the reasons for the devastating cognitive impairment in the SVF and AD dementia in general.

3. Findings about the semantic memory and executive function impairment using the SVF are only made possible through computer-supported qualitative analysis of the semantic verbal fluency on an item-per-item level.
  - a. Through combining both computational semantic as well as temporal modalities in the qualitative analysis of the SVF, executive function and semantic memory impairments in this task can be better separated. This allows the insights from both Chapter 3.1 and 3.2.
  - b. SVF can be interpreted with regards to the underlying involved neurocognitive processes through computer-supported modelling of local semantic as well as temporal organization (clustering, switching and also within and between

cluster distance metrics) but also through global models of semantic word frequency/ word communality in a given language.

The evidence about the clinical significance of computer-supported qualitative analysis of the SVF leads the way towards potential applications in clinical decision support.

4. The more fine-grained qualitative analysis of the SVF is clinically valuable for AD diagnosis and screening but very time-consuming if performed manually. But automatic analysis pipelines can reliably and validly generate this diagnostic information from the SVF.
  - a. Chapter 4.1 shows that automatic transcription of speech plus automatic extraction of novel qualitative features (as investigated in Chapter 3.1 and 3.2) results in comparable clinical interpretation as compared to manual transcripts.
  - b. Chapter 4.1 also shows that the surplus of qualitative clinically important SVF features (AD-related importance has been established in Chapter 3.1) as compared to the traditional smaller set of SVF variables results in improved diagnostic decision support. This is simulated through machine learning classification experiments.
5. Ultimately the SVF could be used for cost-effective fully automated early clinical AD screening.
  - a. Chapter 4.2 shows that a fully automated pipeline based on phone-quality SVF speech input results in reasonably good diagnostic classification decisions. This points towards future epidemiological AD screening applications in healthcare.

## 2 BACKGROUND

This chapter gives an overview of diagnostic concepts and frameworks for AD and its progressive phases as well as the respective profiles of AD-related cognitive decline. This chapter also introduces the semantic verbal fluency task, provides background on novel qualitative analysis methods of it and discusses the ability to either identify clinical stages of AD with it or to characterize cognitive impairment.

### 2.1 AD AS DIAGNOSTIC ENTITY AND ITS STAGES

AD is a progressive neurodegenerative disease which is mainly characterized by cognitive decline (dementia syndrome) and resulting functional impairments. Dementia is not exclusively caused by AD but AD is the most common cause for dementia accounting for 60-80% of all dementia cases<sup>1</sup>. Other reversible conditions such as Depression or substance abuse can temporally cause dementia syndrome (see also Figure 2).

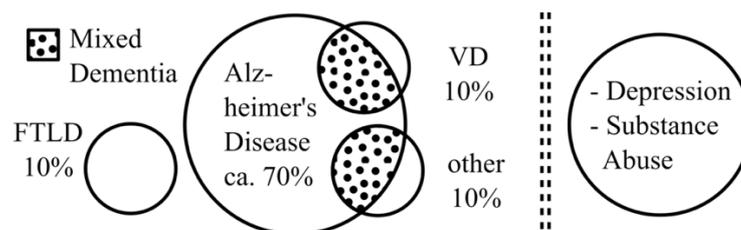


Figure 2: Dementias, according to their biological cause, including Fronto-Temporal Lobar Degeneration (FTLD), and Vascular Dementia (VD); the dotted areas indicate cases where more than one cause underlies the disorder—Mixed Dementia. On the right: mostly reversible, causes for dementia symptoms.

This thesis focuses on dementia as the prominent cognitive-behavioral phenotype of AD. Dementia is a syndrome associated with a loss of cognitive functions caused by underlying neuropathological conditions such as AD. As AD is a neurodegenerative disease worsening over time, AD-related dementia also progresses in severity. To clinically characterize the different phases of AD-related dementia, there are multiple classification systems considering different diagnostic concepts related to AD and dementia. Some of them primarily rely on a behavioral definition of dementia, some propose a hybrid approach also considering biomarkers to indicate degrees of certainty for a development towards AD.

<sup>1</sup> For a disambiguation see <https://www.alz.org/alzheimers-dementia/difference-between-dementia-and-alzheimer-s> (accessed 2021.01.14)

Across the board, dementia is defined as cognitive decline more severe than normal ageing would suggest. However, there are different approaches to what abnormal cognitive decline means and how it should be assessed. Some frameworks propose to differentiate between a mild form of dementia called *mild neurocognitive impairment* (DSM-5; American Psychiatric Association, 2013) or *mild cognitive impairment* (Petersen et al., 2014) that may or may not develop into a more severe form of dementia (in the DSM-5 called *major neurocognitive impairment*). Other approaches focus on the differentiation of dementia *stages* (as defined by the clinical cognitive phenotype) vs. dementia *states* (Dubois et al., 2016). But there are also frameworks that describe the cognitive impairment in more detail and even providing a unified evaluation framework for different dementia etiologies based on their distinct cognitive impairment profiles (DSM-5; American Psychiatric Association, 2013).

In an initial effort to streamline different dementia severity stages in one overarching diagnostic framework, Petersen described the concept of Mild Cognitive Impairment (Petersen, 2004; Petersen et al., 2014) as a diagnostic entity. This approach also encompasses MCI into the bigger picture of gradual progressing cognitive impairment towards dementia. MCI is marked by subjective cognitive complaint (either by the person itself or by a close family member), objective cognitive impairment (performing  $\geq 1$  but  $\leq 2$  standard deviations below age-stratified norm population in a certain cognitive domain) and otherwise preserved cognitive functioning as well as preserved independence in functional abilities. The concept of mild cognitive impairment (MCI) describes an interim stage between normal ageing and very early dementia. MCI thereby designates an early, but clinical, state of cognitive decline that would not satisfy the diagnostic criteria of dementia (yet). Additionally, clinical subtypes of MCI allow to encompass a variety of dementias as well as their respective prodromal forms (Petersen, 2004). These subtypes recognize not only mild memory impairments (amnestic, aMCI) but also MCI originating in other cognitive domains. Within this framework, MCI of the amnestic type (aMCI) as defined by focal memory impairments is presumed to be the early cognitive phenotype of later stage AD dementia; however, an AD diagnosis is typically established with additional proof through biomarkers and not all aMCI patients progress into AD.

In the Petersen criteria biomarkers are not required for the definition of MCI. Thereby the concept of MCI suggests an opposing view to other very biomarker-based definitions of

dementia by profiling dementia and its mild/ early stages solely on cognitive symptoms. For research purposes, however, they posit that biomarkers might be informative to identify etiological MCI subtypes by differentiating between MCI due to AD and MCI that is unlikely to be due to AD.

Adding the layer of biomarkers on top as source of information, there are adjacent diagnostic schemes that categorize MCI with multiple levels of certainty for progressing into clinical AD. In an important work, Dubois and colleagues (Dubois et al., 2016) proposed a unified nomenclature differentiating between AD dementia states and stages. This framework takes both types of evidence into account, biological as well as cognitive-behavioral diagnostic markers, and at the same time differentiates between the dementia phenotype (the *state* as defined by cognitive-behavioral markers) and the underlying biological etiology as indicator for further progression (*state*). This conceptual unification allows to recognize very early states of prodromal AD dementia that are actually still at a preclinical stage showing none or very little cognitive symptoms. This helps to identify the prime target for early and preventive pharmaceutical interventions. This thesis research focuses on the clinical stage of AD, encompassing the clinical prodromal AD stage as defined by Petersen's concept of amnesic MCI (2014) and the traditional AD dementia stage.

The Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association (DSM) introduced in its latest fifth version a novel neurocognitive domain approach as unified basis for neurocognitive disorders including also AD dementia (DSM-5; American Psychiatric Association, 2013). Recognizing in further detail the cognitive symptoms within the clinical stage of dementia, DSM-5 not only adopts the notion of early and late stage cognitive phenotypes of AD dementia (mild neurocognitive disorder similar to Petersen's concept of MCI and major neurocognitive disorder) but also different loci of the impairment. The DSM-5 proposes six distinct neurocognitive domains that are in line with neuropsychological research: complex attention, executive function, learning and memory, language, perceptual motor & social cognition (compare also Figure 3). The DSM-5 further details distinct subdomains for each neurocognitive domain and provides information on how to measure each of them. Eventually AD-related dementia affects all neurocognitive domains. But the one most relevant neurocognitive domain affected by AD across the full range of

clinical stages are learning & memory and executive function being also strongly related to a patient's abilities in daily living (Tomaszewski Farias et al., 2009).

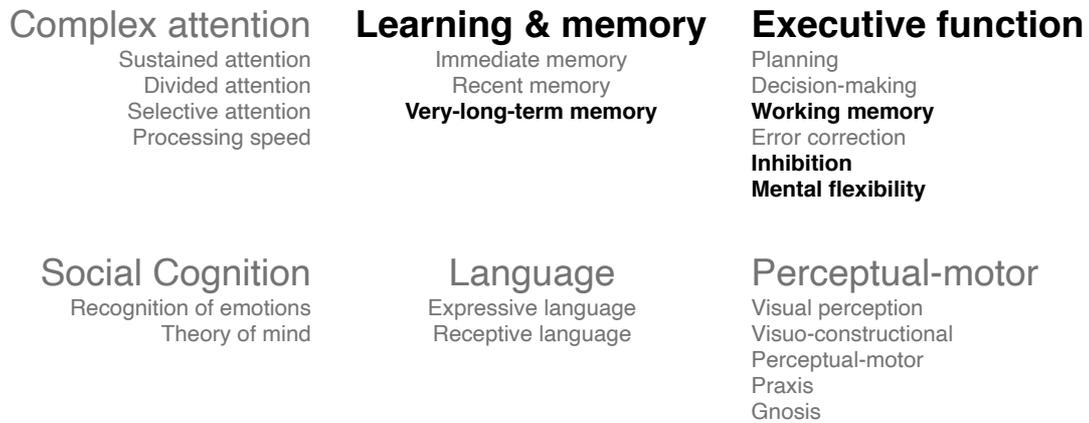


Figure 3: The DSM-5 six neurocognitive domains and their sub-domains. Highlighted in dark the domains and subdomains that are relevant for this thesis.

The recent DSM-5 framework of neurocognitive domains integrates well with the MCI concept of Petersen as different clinical subtypes of MCI are further specified in terms of their respective impaired cognitive domains as defined in the DSM-5.

Because of its closeness to research and proposed explainability structure of neurocognitive domains this thesis will follow the DSM-5 classification system to investigate different cognitive function impairments in dementia. For the definition of the disease progression this thesis will follow the Petersen criteria of MCI (Petersen et al., 2014) situating all experiments within the clinical stage of dementia (prodromal as well as acute dementia) as defined by Dubois and colleagues (Dubois et al., 2016); compare also Figure 4.

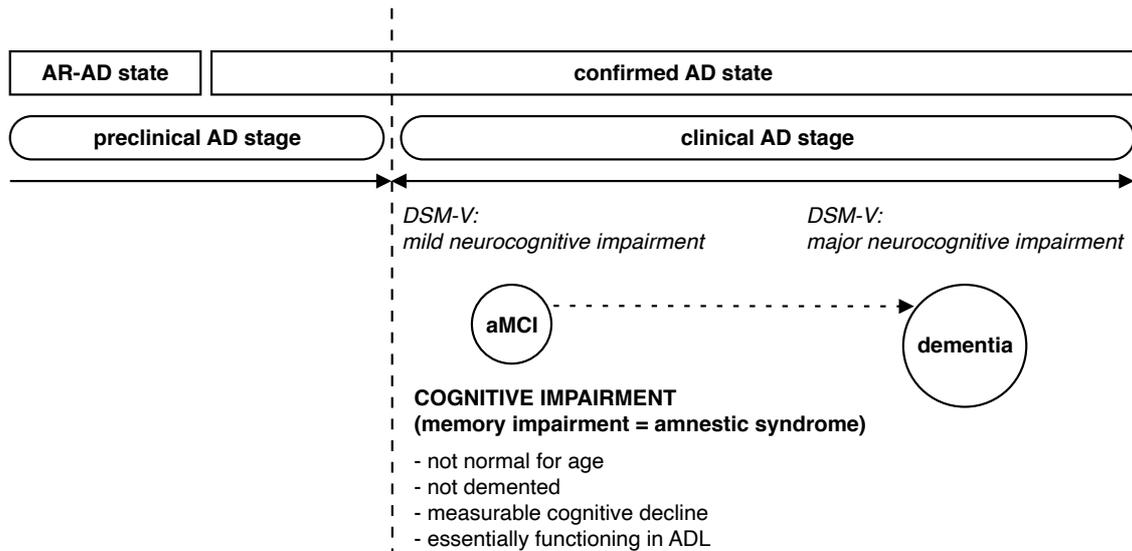


Figure 4: Differentiation between AD stage and state as introduced by Dubois and colleagues (2016). Displaying also the aMCI clinical AD stage as indicated by amnesic symptoms and as defined by Petersen, 2014 and encompassed into the latest DSM-V diagnostic nomenclature. The dashed line indicates the onset of the clinical cognitive AD phenotype. Note that the biomarker supported confirmed AD state can antecede the clinical stage/onset of cognitive symptoms. The dashed arrow indicates the increased risk for patients' progression from the amnesic MCI to the Dementia syndrome in AD. Abbreviations: AD=Alzheimer's Disease, AR-AD=at risk AD, aMCI=amnesic Mild Cognitive Impairment, ADL=Activities of Daily Living.

## 2.2 AD DEMENTIA'S SPECTRUM OF COGNITIVE FUNCTION DECLINE

This thesis investigates cognitive impairments in both prodromal and acute clinical AD stages and more specifically provides evidence for the impairment of executive function and semantic memory across both clinical AD stages (following the above-mentioned DSM-5 neurocognitive domain framework).

Eventually AD-related dementia affects all neurocognitive domains. But the one most relevant neurocognitive domain affected by AD dementia across the full range of clinical stages is memory, more precisely the subdomain of episodic memory<sup>2</sup> (Collie & Maruff, 2000; B. J. Small, Fratiglioni, Viitanen, Winblad, & Bäckman, 2000)—encompassed under the term recent memory by DSM-5 definitions. Episodic memory impairment often marks the earliest neuropsychological sign of (AD) (Galton, Patterson, Xuereb, & Hodges, 2000). In some rare clinical cases though, the disease also starts with other cognitive impairments than memory (Lambon Ralph, Patterson, Graham, Dawson, & Hodges, 2003). After this initial amnesic impairment, other neurocognitive domain impairments in semantic memory, language, complex attentional functions and executive function gradually become prevalent. There is the assumption that these gradual additional cognitive impairments reflect the neurological progression of the pathology (Braak & Braak, 1991). However, over the last decades, there has been work showing that also other neurocognitive (sub-)domains especially executive functioning are impaired early on (Crowell et al., 2002) and throughout the severe clinical progression of the disease in AD (Guarino et al., 2019).

### 2.2.1 *Semantic Memory*

DSM-5 defines semantic memory, a subdomain of Learning & Memory, as very-long-term memory which encompasses semantic or autobiographical knowledge as well as implicit learning. Semantic memory is the memory of entities around us or sometimes referred to the knowledge about the world around us. Semantic memory is built over the lifetime through experience and interaction with this surrounding world (Yee, Jones, & McRae, 2018).

A systematic review puts forward that semantic memory impairments are present throughout all stages of AD dementia even at a pre-clinical stage (Verma & Howard, 2012).

---

<sup>2</sup> Episodic memory is commonly defined as the neurocognitive system enabling people to remember events that have happened in the past (Tulving, 1993).

Semantic memory deficits in AD dementia often come in the disguise of language deficits and therefore are often misinterpreted as language impairments. Language tasks that show early impairment in AD dementia such as picture naming (J. Small & Sandhu, 2006) or categorial association tasks (like the Semantic Verbal Fluency task; SVF) clearly depend on semantic memory integrity (for an overview see Taler & Phillips, 2008). Overall, literature indicates that semantic memory impairments are (1) present at all stages of clinical AD dementia—mild to major (Verma & Howard, 2012) and (2) can be observed and assessed in and through language tasks (Taler & Phillips, 2008).

### 2.2.2 *Executive Function*

DSM-5 defines executive function as a set of subordinate functions such as planning, decision making, working memory, responding to feedback or error correction, overriding habits/ inhibition and mental flexibility. Similar to memory (episodic as well as semantic) impairments, executive function impairment has also been documented in AD, characterized by an impairment in inhibition and mental flexibility (for a comprehensive review see Guarino et al., 2019).

The impairment of the subdomain of mental flexibility is the best documented executive function impairment in AD dementia and draws from the extensive literature on verbal fluency impairments in AD (Henry, Crawford, & Phillips, 2004). The same pattern can be also found in aMCI (Nutter-Upham et al., 2008; Teng et al., 2013) as well as pre-clinical stages with subjective cognitive decline (Nikolai et al., 2017).

Overall there is conclusive evidence on both semantic memory as well as mental flexibility impairment throughout the clinical stages of AD dementia. But more importantly both draw evidence from partially the same psychometric assessment—the SVF task—which has been traditionally classified as a task that assesses language functions. The SVF task is one of the most commonly administered tasks in the assessment of dementia and especially AD dementia (Goldberg, Harvey, Wesnes, Snyder, & Schneider, 2015). Due to its high sensitivity it is present in multiple classic assessment batteries for dementia (e.g. DemTect—Kalbe et al., 2004; Addenbrookes Cognitive Examination-Revised—Mioshi, Dawson, Mitchell, Arnold, & Hodges, 2006) but is also very popular in clinical practice because of its short and easy use across a broad range of scenarios. However, at the same time neuropsychological research

argues that the composite nature of this task makes it one of the most difficult to interpret in research.

The above-mentioned helps to further specify the overall contribution of this thesis which is (1) furthering our understanding of executive function and semantic memory impairments across both clinical stages of AD (aMCI & dementia) and (2) at the same time help to efficiently identify AD patients early and cost-effectively. One can rightfully claim that the SVF is a prime clinical assessment perfectly qualifying for both mentioned aspects. However, this thesis also contributes to the challenge of disentangling the impairment of both neurocognitive constructs within the SVF in AD to make it fit for more sophisticated diagnostics and at the time same demonstrate this value in scalable and cost-effective real-world applications.

## 2.3 QUALITATIVE ANALYSIS OF THE SEMANTIC VERBAL FLUENCY TASK

Verbal fluency (VF) tasks are amongst the most widely applied neuropsychological tests for the assessment of neurocognitive disorders; especially for the diagnosis of clinical AD stages. The main strength of VF tasks is their ease of use (no testing material required and fully speech-based interaction) and brevity (1-2 minutes) given a high sensitivity for above-mentioned diagnostic purposes. Despite traditionally being widely adopted in clinical and diagnostic practice, there is an ongoing scientific discussion regarding what verbal fluency tasks actually measure in terms of neurocognitive domains. However, multiple studies show that VF tasks generate rich variance stemming from the interplay of multiple neurocognitive functions including executive function (EF) as well as memory and language components. Robustly identifying these components is crucial in order to understand how VF could be used to differentiate between multiple dementia sub-forms. VF as a test category comprises two major versions, the semantic verbal fluency (SVF) and the phonemic verbal fluency (PVF). Both follow similar rules: the tested person has to produce as many different words as possible within a given timeframe and a given constraint. The constraint for SVF is to produce only words belonging to a single semantic category (e.g. animals) and for the PVF the constraint is that all produced words should start with one letter (e.g. S). This thesis only focuses on the SVF as it has been shown that it is more sensitive to AD dementia (Henry et al., 2004).

The SVF is most-conducted in the 1-minute version and might be instructed as follows: 'Please name as many different animals as possible within 60 seconds'. It is important to stress two things: First, the categorical cue (i.e. animals), sets the focus on associative semantic memory, in this case the cue is 'animals' but multiple other version exists (sports, fruits, colors, supermarket items, etc.). Second, the constraint 'different' as well as the timed performance property set the focus on executive function involving inhibition as well as working memory and general processing speed. Answering to the mentioned instruction, a fictive piece of SVF production could be the following array of animals:

*frog - dolphin - donkey - monkey - gorilla - tiger - panther - aardvark - ant - crane*

SVF performance is traditionally scored based on the total number of correct in-category words without repetitions. As visible from the breakdown of the instruction itself, a healthy person's SVF performance requires the interplay of semantic memory as well as executive function (foremost mental flexibility). This is also reflected in research recognizing

the involvement of semantic memory (Birn et al., 2010; Rohrer et al., 1995; Shao et al., 2014) as well as the involvement of executive function, which is often regarded as the primary involvement (Amunts et al., 2020; Whiteside et al., 2016). This widely recognized compound nature of the SVF performance posits a major challenge for clinical research not only in AD: although the multiple reasons mentioned above make the SVF extremely popular in everyday clinical practice, it cannot be used effectively for clinical research that needs a more fine-grained resolution of engaged cognitive functions.

In order to differentiate between multiple cognitive processes involved in the SVF, it has been proposed to go beyond the classic quantitative analysis of the SVF (i.e. total number of correct words and/or errors) and analyze also *how* the words in the SVF are produced—the qualitative analysis of the SVF (A. K. Troyer et al., 1997). Two streams for qualitative analysis have been emerging since. The semantic qualitative analysis either looks at the semantic content of the produced words and compares them to the use of language in our culture in general (e.g. through analyzing frequency of words) or at the semantic relationships between the words produced themselves. The temporal qualitative analysis analyses the speed with which participants produce a certain sequence of words in this task. Both types of qualitative analyses are briefly described below.

It has been argued that production of words is normally organized in spurts, forming temporal clusters followed by pauses. This is interpreted as lexical search for semantic fields or subcategories between clusters, and retrieval/production of words within the cluster (Gruenewald & Lockhead, 1980; A. K. Troyer et al., 1997). Thus, researchers assume that between temporal clusters, executive search processes (i.e., switching) and within temporal clusters, semantic memory retrieval processes (i.e., clustering) are engaged. The underlying notion is that temporal clusters correspond to semantic clusters as words comprising temporal clusters are semantically related (A. K. Troyer et al., 1997).

Semantic and temporal dimensions are closely intertwined in this task and are mutually dependent. It has been argued that without the temporal information, the semantic information is rendered un-interpretable and vice versa (Mayr, 2002). Unfortunately, in clinical routine the SVF task is typically not recorded, but transcripts are manually analyzed afterwards, which results in the loss of the temporal dimension. Therefore, the relationship between temporal and semantic structure cannot be investigated. In this thesis both

dimensions are used for a comprehensive qualitative SVF analysis of word productions in subjects. In the following two sections, both semantic as well as temporal qualitative analysis will be explained in depth. Please note that the term qualitative analysis does not refer to the used metrics which are quantitative in nature but rather refers to the intention to analyze *how* the number of words in the SVF are produced. This is due to the fact that only by investigating how the words are produced one can gain insights into the different underlying neurocognitive processes which are in return relevant for a better understanding of AD-related cognitive function decline.

### *2.3.1 Semantic Qualitative Analysis*

For this type of analysis, a full transcript of patients' responses is needed. The typical approach is either modelling the semantic relation between produced SVF responses and language in general or semantic relation amongst the produced SVF responses themselves. This then helps derive insights into the semantic memory retrieval processes of a participant. Troyer and colleagues (A. K. Troyer et al., 1997) first introduced a systematic framework for qualitative semantic analysis. This method uses multiple human-defined taxonomic subcategories that are based on main categories to determine whether successively generated words belong to the same subcategory and thereby form a semantic cluster. Although it seems straightforward to interpret the size of a cluster as a proxy for semantic memory performance, this approach bears some methodological shortcomings. Manual analysis leaves significant room for interpretation of the annotator (the person that categorizes words in taxonomic groups or clusters), as produced words very often belong to one or more predefined subcategories. To better understand the ambiguity incorporated in this rating scheme, one might consider the example from the SVF with animals given earlier:

*frog - dolphin - donkey - monkey - gorilla - tiger - panther - aardvark - ant - crane*



Figure 5: Graphical representation of a classic SVF succession of named animals. Note that all animals can be semantically clustered within the same Troyer-subcategory: African animals (initially yellow). Frog and dolphin can be also clustered within another Troyer-subcategory: water animals (marked red). This approach does not cater for non-category-based associations such as: phonemically similar terms like donkey/monkey (marked green) or animals' co-occurrence in popular culture like the characters in the cartoon series *The Pink Panther* (marked in pink).

There are multiple ways how these utterances could be clustered following the subcategory-based approach introduced by Troyer and colleagues (A. K. Troyer et al., 1997):

- (frog - dolphin - donkey - monkey - gorilla - tiger - panther - aardvark - ant - crane) [all African animals]
- (frog - dolphin) [water animals], (donkey) [animals of burden], (monkey - gorilla - tiger) [jungle animals], (tiger - panther) [felines], (aardvark) [insectivores], (ant) [insects], (crane) [birds]
- ... (monkey - gorilla - tiger - panther) [jungle animals], ...

Hence, there is a need for a qualitative semantic SVF analysis scheme, which minimizes the impact of subjective semantic decisions (compare also Figure 5).

Recently, computational approaches to qualitatively analyze the SVF have been proposed (Woods, Wyma, Herron, & Yund, 2016). Hence, to avoid the above-mentioned shortcomings, statistical methods have been applied in order to obtain semantic clusters. Ledoux and colleagues (2014) have used Latent Semantic Analysis (LSA<sup>3</sup>) to compute similarity within and between clusters. Woods and colleagues (2016), on the other hand, used Explicit Semantic Analysis (ESA) (Clark et al., 2016)—a vector embedding trained on co-occurrence of

<sup>3</sup> <http://lsa.colorado.edu/>

words in Wikipedia articles—to identify chaining behavior for different demographics based on pairwise cosine similarity. Clark and colleagues (D. G. Clark et al., 2016) proposed novel semantic measure based on graph theory; most prominently, they put forward graph-based coherence measures which compare the patient's created sequence/path of words with the 'shortest' possible path through the fully connected weighted graph of all patient's words. This approach is quite similar to the idea of LSA/ESA or word-embeddings in general, comparing a patient's actual production sequence with an independent global representation. However, they normatively provide the graphs' weights (orthographic, phonological, and semantic similarity) and thereby influence the global representation, whereas distributional semantic and word embedding are directly learned from spatial representations of words from large corpora; the latter methods allow for semantic coherence measures without the need of normatively constructing the global space representation. Recently, Linz, Tröger, Alexandersson and König (2017) introduced a similar approach, leveraging neural word embeddings based on large word2vec models (Mikolov, Chen, Corrado, & Dean, 2013) which directly measure the semantic distance between two given words using Euclidean distance in the created embedding vector space. From the vantage point of scalability and feasibility for parallel versions, qualitative SVF analysis based on computational semantics represents a significant leap forward. For this thesis automatic qualitative semantic analysis was operationalized using a neural word embeddings approach to define semantic relationships.

For deriving semantic metrics, the semantic distance between produced words is used, calculated based on pre-trained neural word embeddings such as word2vec (Mikolov et al., 2013) or FastText (Joulin et al., 2016). Word2vec for example is based on a shallow, two-layer artificial neural network trained to embed words in a vector space, where the cosine distance is a measure for semantic similarity between words representations. Like other computational semantic approaches, neural word embeddings define words based on their context with an adjustable context window (the number of directly adjacent words that are considered to derive semantic meaning). However, word embeddings typically do not use distributional metrics, e.g., directly encoded co-occurrence to build the representation. For instance, in a classical distributional model (e.g. LSA or ESA) the word "queen" would be defined by how often it absolutely co-occurs in the context of other words like woman or king. As neural embeddings infer semantic distance directly from the vector embedding, it can render the

semantic association between king and queen even in case they never co-occurred in the training corpus simply through their embedding in the vector space.

Using such embeddings has the major advantage that an approximation of the overall density of a person's produced semantic network can be defined. Semantic proximity is calculated as the semantic distance between all possible word pairs of a person's SVF performance, which in return acts like a fully-connected graph or map of the successful lexico-semantic search items the person produced during the SVF assessment. This in return allows to draw conclusions about the lexico-semantic search process in general.

In order to interpret the semantic value of an SVF utterance as compared to the spoken language in general, computational linguistics measures can be used that define how common the word is in the language. Commonality of a word can be measured by its frequency of occurrence in a sufficiently large repository of text (e.g. how frequent/common is the word 'monkey' in the English language as measured by its occurrences in the whole body of English Wikipedia articles). Word frequency can be approximated using available packages like the Python wordfreq package (Speer, Chin, Lin, Jewett, & Nathan, 2018), which combines resources such as Wikipedia, news and book corpora, and Twitter.

Through above-presented computational approaches to modelling semantic relationships, the methodological issues that arise by using a human-/ taxonomy-based measure for semantic relatedness can be mitigated. However, Mayr (2002) highlights the interpretation problem that occurs when considering only qualitative semantic measures while disregarding the temporal alignment of produced words. Without the temporal information, the size of a semantically related cluster of words depends not only on the ability of a patient to come up with new words within clusters (semantic memory retrieval) but also on the difficulty a subject has with accessing a new semantic cluster (executive control or search strategies). To make best use of all modalities available, qualitative SVF analysis should combine semantic content of the produced word sequence with the temporal distribution of the same sequence.

### *2.3.2 Temporal Qualitative Analysis*

Early research has been shown a clear association between retrieval time and effective semantic memory recall (Collins & Quillian, 1969). Hence there should be a relation between temporal and semantic qualitative properties of participants' SVF performances.

Rohrer and colleagues (1995) proposed that by measuring the latencies of words produced in the SVF one can deduce the nature of the impairment. In this context latency of an SVF response word is defined as the elapsed time until a word is uttered as measured from the start of the SVF production. It's assumed that the mean latency of produced words can be modelled as an interaction of relevant semantic memory size (number of all animals known by the subject; i.e. search set) and the speed of accessing words in this semantic memory (time it takes to retrieve the word and utter it; i.e. sampling time); compare also Figure 6.

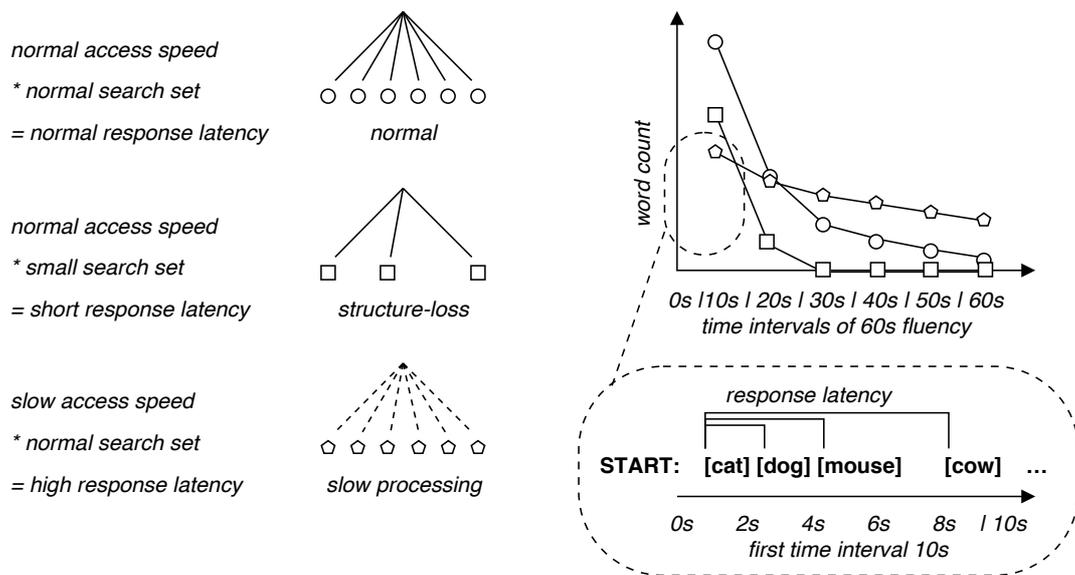


Figure 6: Loss of semantic memory structure vs. slow processing speed and the effect on mean response latency in the SVF.

Therefore, a short mean response latency (compare Figure 6) combined with a lower word count is interpreted as evidence for semantic memory structure-loss associated with AD and aMCI (Randolph, Braun, Goldberg, & Chase, 1993; Rohrer et al., 1995; Tröster, Salmon, McCullough, & Butters, 1989).

However, when plotting the produced words on a temporal axis, the distribution of the produced SVF words of AD patients resembles that of healthy subjects: the lion's share of words is produced early on and then the production flattens out over time asymptotically approaching 0 (Fernaes & Almkvist, 1998; Linz et al., 2019; Rohrer et al., 1995). Based on this, Fernaeus and Almkvist propose a two-processes model of the SVF production. They explain that SVF production is always driven by a fast and automatic retrieval process in the beginning producing the majority of words and a more effortful retrieval process that kicks in later producing the remaining words.

Besides the measures derived from response latencies, temporal information can be leveraged to determine *temporal clusters*. The time between two consecutive utterances has to be measured and then a threshold mechanism has to be applied: if the time is below a certain time threshold the consecutive words can be considered to be part of the same production cluster and if not, a new cluster starts. The difficulty lies in the definition of the threshold as it can be assumed that different subjects may have a different baseline speeds of producing words. Therefore, base thresholds should be determined on a per speaker basis accounting for inter-personal differences.

To conclude, this thesis combines temporal as well as semantic modalities in the qualitative analysis of the SVF to best differentiate between, executive function and semantic memory involvement and the corresponding pathological profiles of AD.

### 2.3.3 Qualitative Analysis of SVF in AD dementia

Research based on quantitative analysis of the SVF performance (focusing on the number of correct SVF responses) shows a clear impairment in patients with AD. However, research draws no conclusive picture regarding metrics from qualitative SVF analysis in AD: Longitudinal studies report a significant decline in switching processes, explaining the overall semantic fluency performance's decline, (Raoux et al., 2008), whereas other longitudinal studies identify clustering as the main impaired process (Mueller et al., 2015). Other cross-sectional studies report an impairment of both processes discriminating between patients with AD and healthy age-matched controls (Gomez & White, 2006; Murphy, Rich, & Troyer, 2006; A. K. Troyer, Moscovitch, Winocur, Leach, & Freedman, 1998) or neither one of them (Pakhomov, Eberly, & Knopman, 2016); for an overview see also Table 2. However, across multiple studies the quantitative SVF-count correlates strongly with both clustering and switching (Gomez & White, 2006; Robert et al., 1998).

*Table 2: Comparison of studies reporting SVF qualitative results for different group comparisons; effect size is reported as standardized mean difference (Cohen's d). Abbreviations: SVF-count='Semantic Verbal Fluency count', NOS='Number Of Switches', MCS='Mean Cluster Size', C='Controls', MCI='Mild Cognitive Impairment', AD='Alzheimer's Disease'. Keys for reported significant effects in the original articles: "/" means no significant result but reported, empty cell means not reported and "\*" means significance reported but not sufficient details reported to calculate d.*

	SVF-count			NOS			MCS		
	C vs. AD	C vs. MCI	MCI vs. AD	C vs. AD	C vs. MCI	MCI vs. AD	C vs. AD	C vs. MCI	MCI vs. AD
Troyer et al., 1998	2.29			1.23			0.94		
Murphy, Rich, & Troyer, 2006	2.01	0.48	1.40	0.95	/	0.82	0.75	/	/

EXECUTIVE FUNCTION & SEMANTIC MEMORY IMPAIRMENTS IN ALZHEIMER'S DISEASE

	SVF-count		NOS		MCS	
Gomez et al., 2006		1.19		0.93		0.57
March & Pattison, 2006	1.6		/		0.99	
Roux et al., 2008	1.04	0.59	0.93	0.33	/	/
Mueller et al., 2015		*		/		*
Pakhomov, Eberly, & Knopman, 2016	*	*	*			
Peter et al., 2016		1.30				/
Clark et al., 2016		0.65				
Price et al., 2012		1.46		/		/

### 3 CHARACTERIZING THE IMPAIRMENT OF EXECUTIVE FUNCTION AND SEMANTIC MEMORY

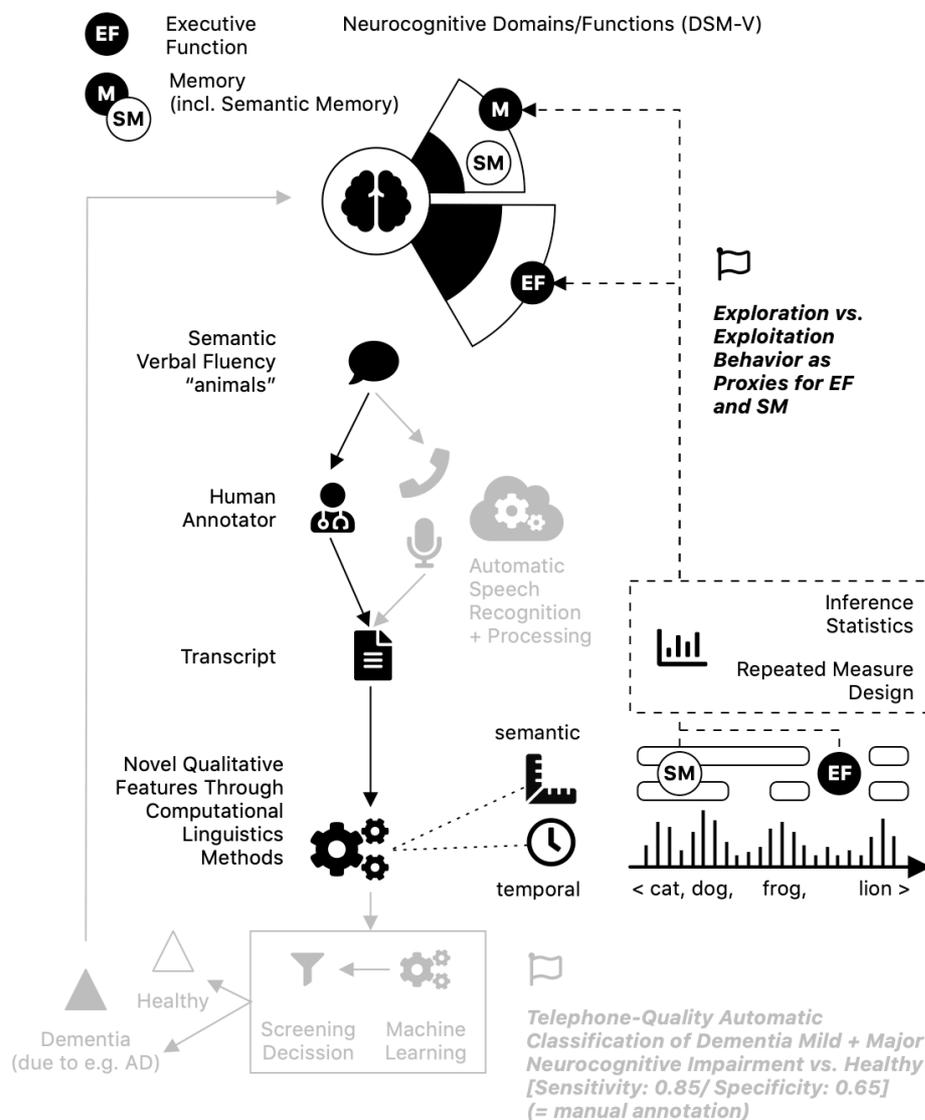


Figure 7: Visual abstract; highlighted in black the major contribution of this chapter as compared to the overall thesis.

Both, semantic memory and executive function have been shown to strongly relate to an AD patient's declining abilities in daily living (Tomaszewski Farias et al., 2009). Semantic memory impairment can be found as early as episodic memory changes and throughout the clinical AD stages dementia and aMCI (Verma & Howard, 2012). Executive function impairment however is sometimes found to be impaired at a prodromal early clinical stage (Crowell, Luis, Vanderploeg, Schinka, & Mullan, 2002) but is also suggested to have the most impact at a later dementia stage (Guarino et al., 2019) potentially accentuating other cognitive

impairments (Buckner, 2004). For the last decades, the SVF has been the established task to measure semantic memory especially in AD and dementia in general (Verma & Howard, 2012). However, more current findings strongly suggest that an AD-related SVF impairment is not only indicative of semantic memory but also of executive function break-down (Amunts, Camilleri, Eickhoff, Heim, & Weis, 2020; Shao, Janse, Visser, & Meyer, 2014). Differentiating between the contributions of executive function and semantic memory impairment within the SVF itself has become a renown challenge (Shao et al., 2014) that can only be solved on an item level using qualitative measures that go beyond the traditional quantitative performance counts such as number of correct words (Rohrer et al., 1995; A. K. Troyer et al., 1997). Automating this early work, recent approaches have been leveraging state-of-the-art computational measures to model influences of both semantic memory and executive function on a qualitative item-level (Linz, Tröger, Alexandersson, & König, 2017; Linz, Tröger, Alexandersson, Wolters, et al., 2017).

Based on this, this first set of studies investigates semantic memory and executive function impairments in both clinical stages of AD based on the SVF alone. The first paper in this chapter reports results from a cross-sectional comparison between both prodromal as well as acute AD stages and healthy controls whereas the second paper investigates closer cognitive impairments at a prodromal aMCI stage through repeated SVF assessment within patients and healthy controls.

3.1 EXPLOITATION VS. EXPLORATION—COMPUTATIONAL TEMPORAL AND SEMANTIC ANALYSIS EXPLAINS SEMANTIC VERBAL FLUENCY IMPAIRMENT IN ALZHEIMER'S DISEASE

Johannes Tröger<sup>1</sup>, Nicklas Linz<sup>1</sup>, Alexandra König<sup>2</sup>, Philippe Robert<sup>2</sup>, Jan Alexandersson<sup>1</sup>, Jessica Peter<sup>3</sup>, Jutta Kray<sup>4</sup>

<sup>1</sup> German Research Center for Artificial Intelligence (DFKI), Germany

<sup>2</sup> Memory Center, CoBTeK, IA CHU Université Côte d'Azur, France

<sup>3</sup> University Hospital of Old Age Psychiatry and Psychotherapy, University of Bern, Switzerland

<sup>4</sup> Chair for Development of Language, Learning & Action, University of Saarland, Germany

Impaired Semantic Verbal Fluency (SVF) in dementia due to Alzheimer's Disease (AD) and its precursor Mild Cognitive Impairment (MCI) is well known. Yet, it remains open whether this impairment mirrors the breakdown of semantic memory retrieval processes or executive control processes. Therefore, qualitative analysis of the SVF has been proposed but is limited in terms of methodology and feasibility in clinical practice. Consequently, research draws no conclusive picture which of these afore-mentioned processes drives the SVF impairment in AD and MCI. This study uses a qualitative computational approach—combining temporal and semantic information—to investigate exploitation and exploration patterns as indicators for semantic memory retrieval and executive control processes. Audio SVF recordings of 20 controls (C, 66–81 years), 55 MCI (57–94 years) and 20 AD subjects (66–82 years) were assessed while groups were matched according to age and education. All groups produced, on average, the same amount of semantically related items in rapid succession within word clusters. Conversely, towards AD, there was a clear decline in semantic as well as temporal exploration patterns between clusters. Results strongly point towards preserved exploitation—semantic memory retrieval processes—and hampered exploration—executive control processes—in AD and potentially in MCI.

*Keywords:* Alzheimer's Disease, MCI (mild cognitive impairment), Semantic Speech Analysis, Temporal Analysis

### 3.1.1 Introduction

Semantic verbal fluency (SVF), requires the verbal production of as many different items from a given category (e.g., animals) as possible within a given time. Multiple studies have shown the SVF's diagnostic sensitivity for dementia—due to Alzheimer's Disease (AD) or other etiologies—and its precursor Mild Cognitive Impairment (MCI) (Auriacombe et al., 2006; Gomez & White, 2006; Henry et al., 2004; Pakhomov et al., 2016; Raoux et al., 2008). However, for diagnostic purposes, it is crucial to identify which neurocognitive function drives SVF impairment, distinguishing between diseases with neurodegenerative origin from other etiologies (e.g. focal lesions or dysexecutive syndromes). Therefore, the SVF productions of patients with frontal lobe lesions would qualitatively follow a different pattern than those of temporal lobe lesions. Indeed, there is evidence indicating that performance in SVF tasks is predicted by individual differences in both semantic memory and executive control (Peter et al., 2016; Shao et al., 2014). Hence, attempts have been made to better differentiate between semantic memory and executive control impairments in the SVF: In addition to quantitative analysis of SVF performance (i.e. the count of correctly produced responses excluding repetitions), researchers have suggested qualitative analysis of SVF performance that aimed at separating semantic memory retrieval from executive control processes (Gruenewald & Lockhead, 1980; Henry et al., 2004; Robert et al., 1998; A. K. Troyer et al., 1997). However, this qualitative analysis scheme of the SVF tasks demands a substantial amount of manual work.

The major objectives of this paper are (1) to investigate—by the means of qualitative SVF analysis—how semantic memory retrieval and executive control processes influence the SVF performance in AD and (2) to advance the state of the art in qualitative SVF analysis by the means of novel computational measures.

#### Qualitative Analysis of the SVF

Generating words according to a given semantic category involves multiple cognitive processes including lexical retrieval, systematic lexical search, keeping track of generated words, and inhibiting automatic erroneous responses (Crawford & Henry, 2005). Considering semantic memory retrieval and executive control processes, Troyer and colleagues (A. K. Troyer et al., 1997) first introduced a systematic framework for qualitative analysis of the response behavior in the SVF. In general, the production strategy of words is organized in

spurts—temporal clusters—followed by pauses, implying the lexical search for semantic fields or subcategories between clusters—exploration—and retrieval/production of words within the cluster—exploitation (Gruenewald & Lockhead, 1980; A. K. Troyer et al., 1997). Words “that comprise these temporal clusters tend to be semantically related” (A. K. Troyer et al., 1997). Indeed, there is a clear association between retrieval time and effective semantic memory retrieval (Collins & Quillian, 1969). However, in most cases in research and clinical routine the SVF is transcribed and not recorded, meaning that the temporal dimension of the data is lost and the relationship between temporal and semantic structure cannot be investigated.

From the transcribed succession of words (e.g., animals), qualitative measures are calculated following the approach of Troyer and colleagues (A. K. Troyer et al., 1997). This method uses human-defined taxonomic subcategories to determine clusters. This analysis regime leaves significant room for interpretation of the annotator. Mayr (2002) highlights the interpretation problem that occurs when considering only qualitative clustering and switching measures while disregarding the temporal alignment of produced words and cluster boundaries. Without the temporal information, the switching measure is rendered as not interpretable in the sense of what cognitive process actually causes a potential low switching rate. This follows as “the number-of-switches score depends not only on the difficulty a subject has with accessing a new semantic cluster. The number of switches is reduced just as well when a subject has difficulties coming up with new words within clusters.” [p. 563] (Mayr, 2002). Thus, from a methodological perspective, there is a need for a qualitative SVF analysis scheme, which minimizes the impact of subjective semantic decisions and combines the semantic content of the produced word sequence with the temporal distribution of the same sequence. One important goal of the present study is to provide a new approach modelling exactly this combination of the temporal dynamic of produced words and their semantics.

#### *Quantitative and Qualitative SVF Impairments in MCI and AD*

Beyond the overall SVF performance, studies have reported the significance of errors in the SVF performance when comparing AD and controls, especially repetitions/perseverations (March & Pattison, 2006). However, there are also studies that reported insignificant repetition/perseverations in the SVF task comparing future AD subjects at an MCI stage and non-converters (Raoux et al., 2008).

A comprehensive overview of study results investigating MCI and AD in comparison to healthy controls in quantitative and qualitative measures of the SVF and their respective effect sizes is provided in the supplementary material. When focusing on quantitative analysis (i.e., the number of produced words), these studies indicated a clear impairment in patients with MCI and AD, as compared to healthy elderly controls. However, results are mixed regarding the qualitative SVF analysis in MCI or AD: Longitudinal studies reported a significant decline in switching processes, explaining the overall semantic fluency performance impairment (Raoux et al., 2008). Some studies interpreted impaired switching as dysfunctional executive control mechanisms rather than impaired semantic memory (Peter et al., 2016; Raoux et al., 2008). Other longitudinal studies identified clustering as the main impaired process (e.g. Mueller et al., 2015) arguing in favor for semantic memory degradation and a following impaired lexico-semantic access (March & Pattison, 2006; Murphy et al., 2006; Price et al., 2012; A. K. Troyer et al., 1998). Other cross-sectional studies report an impairment of both processes discriminating between patients with AD and healthy age-matched controls (Gomez & White, 2006; Murphy et al., 2006; A. K. Troyer et al., 1998) or neither one of them (Pakhomov et al., 2016).

Despite these contradictory findings across multiple studies, the quantitative SVF-count correlates strongly with both clustering and switching (Gomez & White, 2006; Robert et al., 1998). From a methodological standpoint, both the switching (Number of Switches; NOS) and cluster-size measures (Mean Cluster Size; MCS) are formally related to the overall SVF-count (note that the total number of switches (NOS) is always one unit smaller than the total number of clusters):

$$SVF_{count} = ((NOS + 1) * MCS) - Repetitions - Intrusions$$

Taking this into account, most studies rely on human annotators for determining the clusters manually by applying a taxonomic set of rules to define clusters. As mentioned above, this approach leaves room for interpretation and as a result the same SVF performance with a fixed SVF-count can be interpreted in favor for a larger or smaller mean cluster size or number of switches. This subjectivity in the calculation may explain some of the contradicting findings. Therefore, one aim of the present study is to provide a new approach of how qualitative measures can be extracted from the SVF that are formally less dependent on the

overall SVF count and can be extracted uniformly across studies and settings, rendering comparable results.

#### *Computational Qualitative Analysis as a Novel Approach to Analyze SVF*

Recently, computational approaches to analyze SVF have been proposed (D. G. Clark et al., 2016; Linz, Tröger, Alexandersson, & König, 2017; Woods et al., 2016). From the vantage point of scalability and feasibility for parallel versions, automatic qualitative SVF analysis based on computational semantics represents a significant leap forward. Considering recent computational semantics and temporal analysis approaches, the model of both semantic memory retrieval processes and executive control processes in the SVF can be re-framed as a trade-off between exploitation (i.e., semantic memory retrieval process) and exploration (i.e., executive control processes) (Hills, Todd, & Jones, 2015). Given the performance nature of the task (i.e., name as many different in-category items as possible within a fixed time), participants have to trade-off between exploitation and exploration in order to maximize their output. This is in line with the argument of Mayr (Mayr, 2002), stating that an SVF impairment is a trade-off between a retrieval impairment within clusters and an executive/strategic impairment between clusters.

#### *Study Goals and Research Questions*

Given the above-mentioned advances in computational linguistic, this paper addresses the following research questions: (1) Can computational semantic and temporal measures clarify the involvement of exploitation and exploration behavior and if so (2) which of these processes is the main driving force behind the progressive SVF impairment in AD. (3) This paper will also investigate how novel computational measures relate to traditional qualitative measures within MCI and AD patients and how the results relate to the current literature.

### *3.1.2 Methods*

#### *Participants*

Twenty patients with AD (age range = 66–82 years; 14 female), 55 patients with MCI (age range = 57–94 years; 25 female) and 20 Controls (C) (age range = 66–81 years; 15 female) participated in this study. All participants were recruited through the Memory Clinic located at the Institute Claude Pompidou in the Nice University Hospital. Written informed

consent was obtained from all subjects prior to the experiments. The study was approved by the local Ethics Commission and was conducted according to the Declaration of Helsinki.

All participants were native French speakers and were excluded if they had any major auditory or language problems, history of head trauma, loss of consciousness, psychotic or aberrant motor behavior, or history of drug abuse—according to anamnesis and clinical record.

All participants performed a cognitive test battery consisting of, amongst others, the following tests: The Frontal Assessment Battery (Dubois, Slachevsky, Litvan, & Pillon, 2000), the 5 Word Test (Cowppli-Bony et al., 2005), the FCSRT—only in case the 5 Word Test did not already reveal a memory impairment (Grober, Ocepek-Welikson, & Teresi, 2009), the Mini Mental State Examination (MMSE) (Folstein, Folstein, & McHugh, 1975), and the Clinical Dementia Rating (CDR) scale (John C. Morris, 1997).

Participants were assigned to one of three groups according to their diagnosis which was established prior to the study by a consensus of the medical team of the Memory Clinic, based on anamnesis, additional neuropsychological assessments, and clinical interviews with participant and caregiver. Participants with normal cognitive test performance and no evidence of functional decline were assigned to the cognitively healthy control (C) group. Patients with subjective cognitive decline and evidence of impaired cognitive function ( $\leq 1.5$  SD below norm in one neurocognitive domain), otherwise preserved cognitive functioning ( $\geq 0.5$  SD) according to criteria established by Petersen and colleagues (Petersen et al., 1999), intact routine activities of daily living and little or no evidence of functional decline were assigned to the mild cognitive impairment (MCI) group. Finally, AD diagnosis was determined using the NINCDS-ADRDA criteria (McKhann et al., 2011).

Information about the sample characteristics and cognitive test battery are given in Table 3. All three groups were matched for sociodemographic variables (see Table 3). There were no significant effects for age ( $p = .99$ ) nor years of education ( $p = .68$ ) between the groups.

*Table 3: Demographics for the respective groups of participants (M = mean, SD = standard deviation; Education in years; Overall group differences tested with non-parametric Kruskal–Wallis test). C='Controls', MCI='Mild Cognitive Impairment', AD='Alzheimer's Disease.*

C		MCI		AD	
M	SD	M	SD	M	SD

	C		MCI		AD	
N	20		55		20	
Age	77.3	4.0	77.0	7.5	77.4	4.2
Education	10.9	3.8	10.8	3.7	9.7	4.9
MMSE	28.3	1.7	25.9	2.7	17.5	4.3
CDR-SOB	0.5	0.7	2.3	1.5	7.8	3.3

### *Procedure*

Participants were asked to perform a battery of cognitive tests, including an SVF test (category: animals). In this test, participants were asked to produce as many different names of animals as possible in one minute and to avoid repetitions of animal names. Speech recordings of all participants were collected using an automated recording app on a tablet computer and were subsequently transcribed in PRAAT (Boersma & Weenink, Version 6.1.42) by trained students from the field of computational linguistics following the CHAT protocol (McWhinney, 1991). Following the CHAT protocol, every single utterance of the participants is transcribed, (including thinking aloud patterns and unintentional verbalizations of cognitive updating processes, e.g., 'what else is there' or 'cat, um, cat, cat, cat, what else, dog' 'did I say that already'). As a result, repetitions in the SVF task are systematically overestimated which may also influence assessment of other qualitative measures obtained from this task. In order to account for this, consecutive repetitions were deleted, but not repetitions in general.

### *Computational Semantic Verbal Fluency Measures*

From the transcribed audio recordings of the SVF, two different types of qualitative measures were automatically computed: (1) traditional taxonomic measures established by the approach of Troyer and colleagues (A. K. Troyer et al., 1997) and (2) novel measures derived from computational approaches.

### *Traditional Taxonomic Measures*

For the traditional measure, the *SVF-count*, reflecting the total number of produced animals excluding errors and repetitions, was determined. Additionally, *repetitions* were computed from this. Furthermore, semantic clusters were analyzed, a cluster being defined as sequences of successively generated words belonging to the same taxonomic subcategory. We adapted taxonomic subcategories one-to-one from the original research by Troyer and colleagues (A. K. Troyer et al., 1997) and sorted all named/transcribed animals in those categories, forming a lexicon that allows automatic scoring. A cluster consisted of a minimum

of two words belonging to the same category (cluster size: a single word = 0, two words = 1, three words = 2, etc.). The *mean cluster size* (MCS) was calculated as the sum of cluster sizes divided by the number of clusters. Finally, the *number of switching clusters* (NOS) was defined as the total number of switches between clusters, including single word clusters.

#### *Novel Computational Measures*

The novel computational measures included qualitative measures based on temporal information and measures based on computational semantics information (semantic proximity). Comparable to the traditional taxonomic measures, either temporal or semantic proximity information was used to determine clusters and switches.

In order to determine *temporal clusters*, each word  $w_i$  was assigned a start time  $s_i$  and an end time  $e_i$  according to its position in the speech recording. Clusters were then determined iteratively. The first word  $w_0$  started a new cluster. The next word  $w_j$  was part of the previous cluster, if the distance between its start time  $s_j$  and the previous words end time  $e_{j-1}$  was below a threshold  $t_j$  (i.e.,  $s_j - e_{j-1} < t_j$ ). A base threshold  $t$  was determined on a per speaker basis as the mean distance between any consecutive words produced by the speaker. To account for the fact that word production decreases towards the end of the task (Fernaes & Almkvist, 1998; Woods et al., 2016), this base threshold was linearly scaled by a maximum factor of two, based on the start of the current word  $w_j$  (i.e.,  $t_j = t * (\frac{s_j}{60} + 1)$ ). Please note that this approach automatically accounts for inter-personal differences in terms of production speed and reaction time.

Based on this clustering, the same MCS and NOS measures as described above were computed automatically. Moreover, to operationalize the efficiency of exploration and exploitation, the mean of all transition durations (in seconds) between temporal clusters, and the mean time between consecutive words, produced inside a temporal cluster, were calculated.

For deriving semantic metrics, the semantic distance between produced words was calculated based on a fastText (Joulin et al., 2016) neural word embedding, pre-trained on the French Common Crawl and Wikipedia corpora (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018; Linz, Tröger, Alexandersson, & König, 2017). This model is based on a shallow, two-layer artificial neural network trained to embed words in a vector space, where the cosine distance

between word representations is a measure for semantic similarity. Like other computational semantic approaches, fastText-models define words based on their context; based on previous studies the context window was set to five tokens in the corpus text (Linz, Tröger, Alexandersson, & König, 2017). However, fastText does not use distributional metrics, (e.g., directly encoded co-occurrence) to build the representation. For instance, in such a model the word “queen” would be defined by how often it absolutely co-occurs in the context of other words like woman or king. As fastText infers distance directly from the vector embedding, it can render the semantic association between king and queen even in case they never co-occurred in the training corpus simply through their embedding in the vector space.

Such embeddings have the major advantage of giving an objective approximation of the overall density of a person's produced semantic network. Semantic proximity is calculated as the semantic distance between all possible word pairs of a person's SVF performance, which in return acts like a fully-connected graph or map of the overall lexico-semantic search the person underwent during the SVF assessment. To determine the difference in effectiveness between the processes of exploitation of a semantic category (inside a temporal cluster) and the effectiveness of exploration of the semantic space (switching between temporal clusters), we calculated the inter- and intra-cluster semantic proximity. The *inter-semantic cluster proximity* was defined as the mean distance between the centroids (arithmetic mean position of all the points in the shape) of any pair of temporal clusters. Accordingly, the *intra-semantic cluster proximity* was determined as the overall mean distance between the mean distances of any pair of words occurring within a cluster. The difference between semantic proximity within clusters and semantic proximity between clusters was calculated accordingly; Figure 8 gives a visualization for better understanding.

## EXECUTIVE FUNCTION & SEMANTIC MEMORY IMPAIRMENTS IN ALZHEIMER'S DISEASE

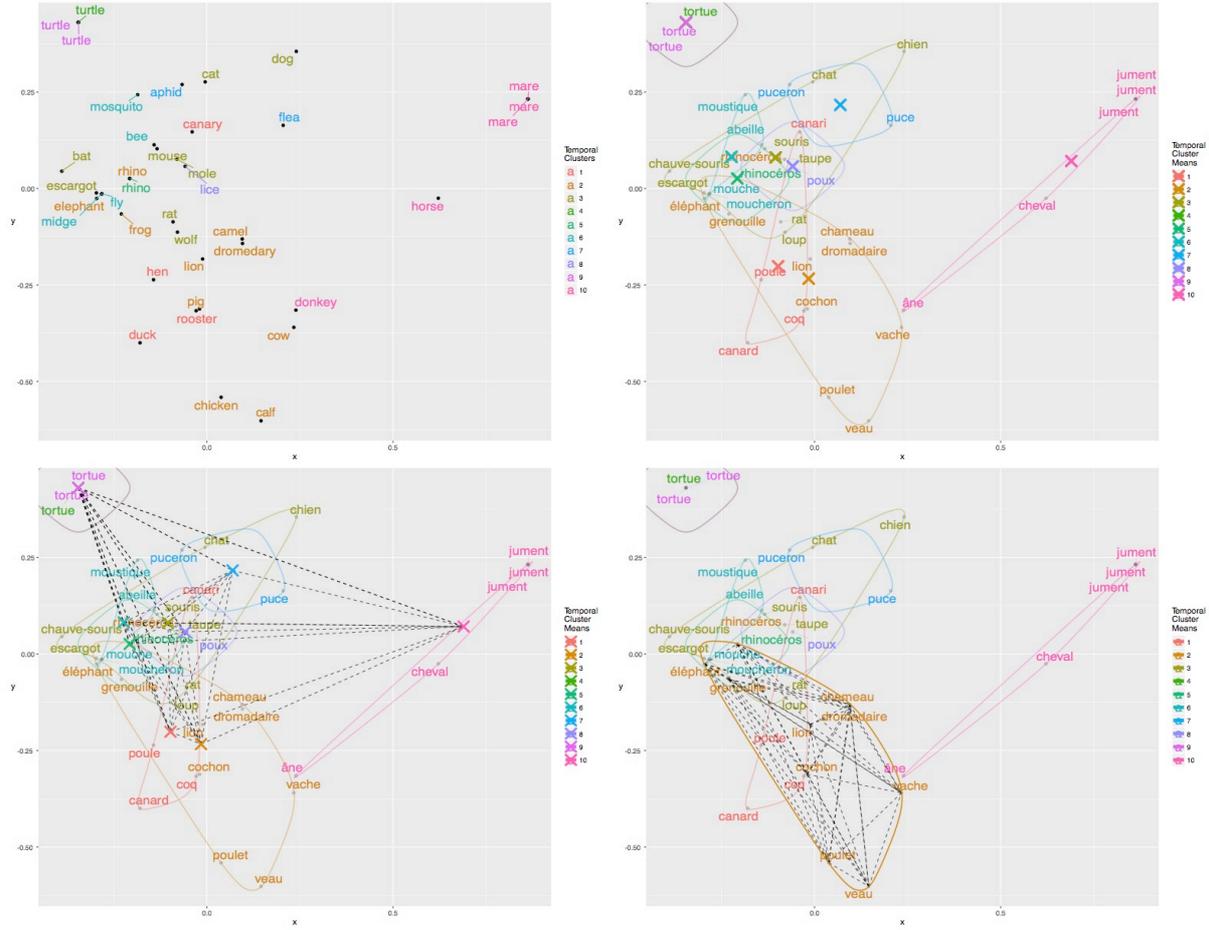


Figure 8: Semantic map of a patient's word production in the SVF task. Distance in the 2D-coordinate system directly represents semantic proximity. For visualization, the original vector space has been reduced to two dimensions using Principal Component Analysis.

In order to statistically determine *semantic clusters*, each word  $w_k$  was assigned a representation in the vector space. Let  $w_1, w_2, \dots, w_n$  be the sequence of animals produced by a participant  $p$ . Let  $\vec{w}_1, \vec{w}_2, \dots, \vec{w}_n$  be their representations in the vector space and let  $w_1, \dots, w_{n-1}$  form a semantic cluster. The next successive produced animal  $w_n$  is part of this cluster if

$$\left| \frac{\langle \vec{\mu}, \vec{w}_n \rangle}{\|\vec{\mu}\| * \|\vec{w}_n\|} \right| > \delta_p$$

With

$$\vec{\mu} = \frac{1}{n-1} * \sum_{\vec{x} \in \{\vec{w}_1, \vec{w}_2, \dots, \vec{w}_{n-1}\}} \vec{x}$$

One of the main problems of using computational semantic models to determine clusters is finding a sensible cut-off value  $\delta$ . Based on earlier experiments (Linz, Tröger, Alexandersson, & König, 2017) the mean distance between any animal produced by a subject

was used as cut-off. An ad-hoc global cut-off value would be hard to determine, since similarity scores tend to vary a lot. Based on this structure, the same MCS and NOS measures as described above were computed automatically. Moreover, to operationalize the efficiency of exploration and exploitation, the mean of all transition durations (in seconds) between semantic clusters (*switch transition duration*), and the mean time between consecutive words, produced inside a semantic cluster (*cluster transition duration*), were calculated. For an overview of all calculated measures consider the first column of Table 4.

#### *Statistical Analysis*

Statistical analysis was performed using R (software version 3.4.02). According to Levene's test for homogeneity of variances, homogeneity of variances was not met for around one third of the reported SVF measures. Consequently, the Wilcoxon signed-rank and ranked-sum tests for dependent and independent sample testing as well as the non-parametric Kruskal–Wallis test for multiple diagnostic groups, was computed. For adjacent pairwise comparisons,  $p$  values for each comparison were adjusted according to Bonferroni's method:  $p^* = p * n$  ( $n$  is the number of reported significant main effects multiplied by the three possible pairwise comparisons within the three diagnostic groups).

EXECUTIVE FUNCTION & SEMANTIC MEMORY IMPAIRMENTS IN ALZHEIMER'S DISEASE

Table 4: Descriptive statistics for the reported dependent variables. For significant main effects, pairwise comparisons with effect sizes are shown. Effect size is reported as Cohen's *d*, asterix indicate Bonferroni corrected significance: < .001\*\*\*, < .01\*\*, <

	C		MCI		AD		C vs. AD	MCI vs. AD	C vs. MCI
	M	SD	M	SD	M	SD			
Defined temporally									
Inter_cluster_proximity	.45	.05	.48	.07	.60	.14	<i>d</i> = 1.81**	<i>d</i> = 1.00**	<i>d</i> = 0.47
Intra_cluster_proximity	.63	.06	.63	.07	.64	.10	/	/	/
Proximity_difference	.18	.05	.14	.08	.03	.15	<i>d</i> = 1.67**	<i>d</i> = 0.86*	<i>d</i> = 0.39
Cluster_transition_duration	1.10	0.56	1.57	1.02	2.07	1.58	<i>d</i> = 1.17	<i>d</i> = 0.42	<i>d</i> = 0.52
Switch_transition_duration	5.85	1.68	7.53	2.92	12.53	5.30	<i>d</i> = 2.26***	<i>d</i> = 1.03**	<i>d</i> = 0.54
NOS_temporal	4.25	1.65	3.29	1.41	1.80	1.15	<i>d</i> = 1.90***	<i>d</i> = 1.06**	<i>d</i> = 0.54
MCS_temporal	2.57	0.79	2.27	0.79	2.19	0.97	/	/	/
Defined semantically									
Inter_cluster_proximity	.44	.04	.49	.09	.55	.15	<i>d</i> = 1.28*	<i>d</i> = 0.47	<i>d</i> = 0.60
Intra_cluster_proximity	.73	.06	.72	.19	.75	.27	<i>d</i> = 0.93	<i>d</i> = 0.55	<i>d</i> = 0.30
Cluster_transition_duration	1.60	0.70	1.96	1.41	2.47	2.08	/	/	/
Switch_transition_duration	3.47	0.98	4.95	2.34	7.28	3.97	<i>d</i> = 1.38*	<i>d</i> = 0.56	<i>d</i> = 0.70
Duration_difference	1.87	1.14	2.99	2.70	4.42	4.94	<i>d</i> = 0.68	<i>d</i> = 0.34	<i>d</i> = 0.44
NOS_semantic	6.20	2.07	4.76	2.36	4.35	2.58	<i>d</i> = 0.95	<i>d</i> = 0.12	<i>d</i> = 0.75
MCS_semantic	1.27	0.68	0.91	0.66	0.73	0.55	/	/	/
Traditional Measures									
SVF-count	17.10	4.40	12.65	4.38	7.05	3.20	<i>d</i> = 2.74***	<i>d</i> = 1.30***	<i>d</i> = 0.90*
Repetitions	0.90	1.12	0.95	1.34	1.40	1.31	/	/	/
NOS_troyer	8.90	3.16	6.42	2.78	3.10	2.53	<i>d</i> = 2.07***	<i>d</i> = 1.13**	<i>d</i> = 0.71
MCS_troyer	0.96	0.38	1.03	0.59	1.33	0.73	/	/	/

### 3.1.3 Results

The following section will report results from the traditional and novel qualitative SVF analysis. Means and standard deviations for all reported groups and measures including effect sizes for the main effect-driven group comparisons are reported in detail in Table 4.

#### *Combining Temporal and Semantic Measures*

Above all, this paper combines both computational temporal and semantic measures. This is done with two approaches: first, by using an automatically generated temporal architecture of clusters to enable an unbiased look at semantic organization and secondly, by using an automatically generated computational semantic architecture of clusters to enable an unbiased look at temporal organization.

#### *Analysis Based on Computational Temporal Architecture*

Temporal clusters provide the framework to compare semantic qualitative aspects within and between clusters. In general, the produced clusters of ADs are more semantically proximate in a semantic global space than those of MCI and C measured by semantic inter-cluster proximity [ $AD > MCI = C, \chi^2(2, N = 95) = 24.68, p < .001$ ]. Pairwise comparison revealed that ADs produced significantly more proximate clusters than both MCI [ $W = 228, p^* < .01$ .] and C [ $W = 43, p^* < .001$ .], and MCIs produced similarly proximate clusters as Cs [ $W = 384, p^* = .85$ .]. In contrast, correct words produced within a temporal cluster are across all groups similarly proximate [ $\chi^2(2, N = 95) = 0.13, p = .94$ .]. In general, across all groups, words within a temporal cluster were more semantically proximate than clusters among each other ( $Z = 7.70, p < .001$ ). Interaction-wise, there was a significant effect for the difference between intra and inter cluster proximity ( $AD < MCI = C, \chi^2(2, N = 95) = 18.52, p < .001$ ) showing for the AD group a significantly different intra-inter proximity distribution as compared to MCI [ $W = 836, p^* < .05$ .] and C [ $W = 350, p^* < .001$ .] and no difference for the comparison of MCI and C. For the semantic proximity interaction compare Figure 10.

#### *Analysis Based on Computational Semantic Architecture*

Computational semantic clusters provide the framework to compare temporal qualitative aspects within and between clusters. In general, patients with AD need more time

for switching between semantic clusters than Cs measured by the switch transition duration, with MCIs showing no significant difference in comparison to the other groups [ $AD > C$ ,  $\chi^2(2, N = 95) = 16.22, p < .001$ ]. Pairwise comparison revealed that ADs need significantly longer to switch between clusters than Cs [ $W = 67, p^* < .05$ ]. In contrast, the duration between correct words produced within a semantic cluster is similar across groups: short [ $\chi^2(2, N = 95) = 1.10, p = .58$ ]. In general, across all groups, transition duration of words within a semantic cluster was shorter than transition duration between clusters/ i.e. switches ( $Z = 7.42, p < .001$ ). Interaction-wise, there was a significant effect for the difference between switch transition duration and cluster transition duration ( $\chi^2(2, N = 95) = 6.07, p < .05$ ) showing trends that ADs have a greater difference between the transition times than MCI and C in a sense that switch transition duration is longer than cluster transition duration.

#### *Global SVF Performance Measures*

There was a significant effect for the SVF-count across the three diagnostic groups [ $AD < MCI < C$ ;  $\chi^2(2, N = 95) = 37.92, p < .001$ ]. There was no effect for SVF repetitions though:  $\chi^2(2, N = 95) = 2.67, p = .26$ .

#### *Clustering and Switching Measures*

Taxonomically as well as computationally derived clusters (based on temporal and semantic framework) showed the same pattern over the three diagnostic groups. Spearman's rank-order correlation between switching measures were moderate to strong: Troyer & temporal:  $r_s = .67$ ; Troyer & semantic:  $r_s = .57$ ; temporal & semantic:  $r_s = .52$ . On the contrary, correlation between clustering measures were rather weak or also negative: Troyer & temporal:  $r_s = -0.15$ ; Troyer & semantic:  $r_s = .19$ ; Temporal & semantic:  $r_s = .21$ . For a comparison and an example of taxonomic and temporal clustering and switching frameworks compare Figure 9 and Figure 10.

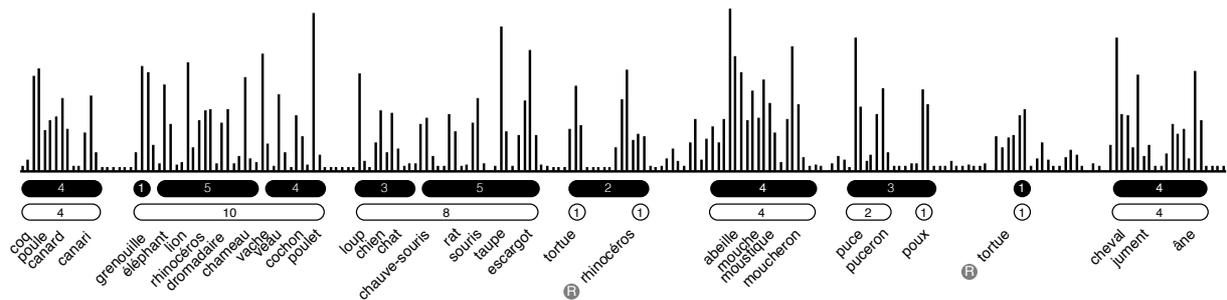


Figure 9: Juxtaposing traditional taxonomic (filled/black) and temporal (white) frameworks for clusters aligned with the speech signal and transcription of a healthy control subject.

NOS derived by the traditional taxonomic measure from Troyer and colleagues (1998) showed a clear effect over the diagnostic groups [ $AD < MCI = C$ ;  $\chi^2(2, N = 95) = 30.57, p < .001$ ]; pairwise comparisons revealed that ADs switch less than Cs [ $W = 367.5, p^* < .001$ ] and MCIs [ $W = 902, p^* < .01$ ] who switch similarly often as Cs. In contrast the MCS shows no significant between group differences [ $\chi^2(2, N = 95) = 3.77, p = .15$ ]. Similarly, NOS derived by the temporal measures showed a clear effect over the diagnostic groups [ $AD < MCI = C$ ;  $\chi^2(2, N = 95) = 26.36, p < .001$ ]; pairwise comparisons revealed that ADs perform less temporal switches than both Cs [ $W = 358.5, p^* < .001$ ] and MCIs [ $W = 880, p^* < .01$ ] who switch as often as Cs. No significant between group differences could be found for temporal MCS [ $\chi^2(2, N = 95) = 3.19, p = .20$ ]. The same pattern can be found for NOS derived by semantic architecture which also showed an effect over the three diagnostic groups [ $\chi^2(2, N = 95) = 10.57, p < .01$ ].

EXECUTIVE FUNCTION & SEMANTIC MEMORY IMPAIRMENTS IN ALZHEIMER'S DISEASE

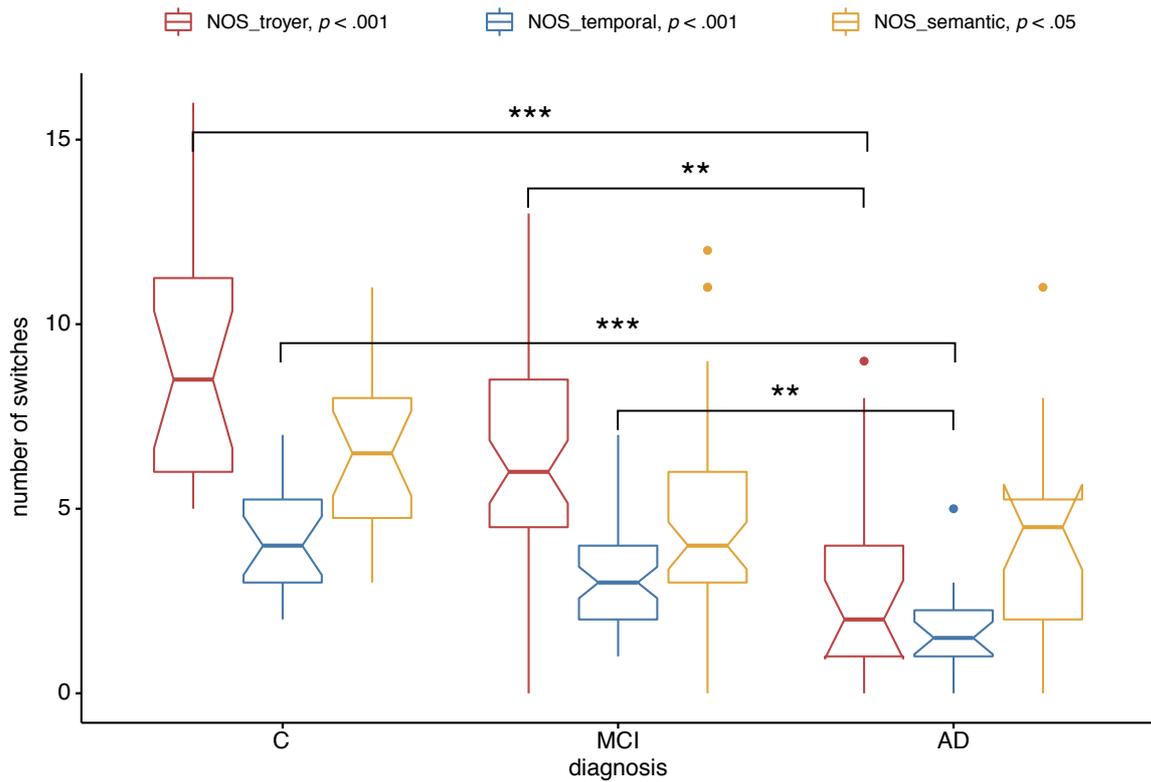


Figure 10: Comparison of qualitative switching measures (NOS) derived by traditional taxonomic (red), computational semantic (yellow) and computational temporal approaches (blue). Asterisks indicate significance of Bonferroni-corrected non-parametric group comparisons (< .001\*\*\*, < .01\*\*, < .05\*).

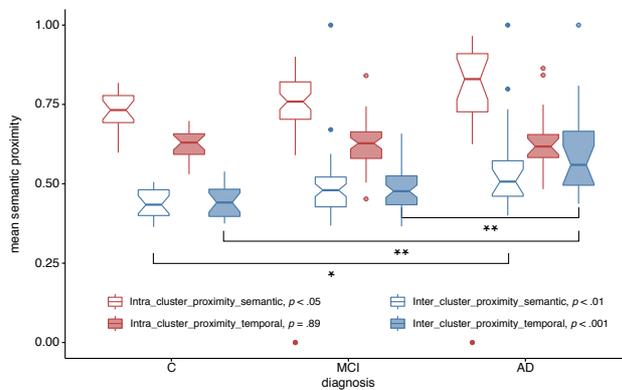


Figure 11: Semantic proximity between (blue) and within (red) semantic (empty) as well as temporal (filled) clusters. Asterisks indicate Bonferroni-corrected significance of non-parametric group comparisons (< .001\*\*\*, < .01\*\*, < .05\*).

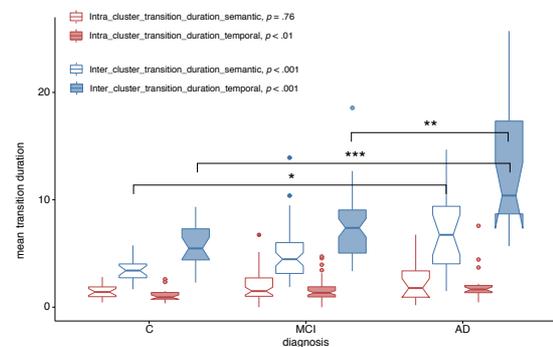


Figure 12: Transition duration in seconds between (blue) and transition duration of words within (red) semantic (empty) as well as temporal (filled) clusters. Asterisks indicate Bonferroni-corrected significance of non-parametric group comparisons (< .001\*\*\*, < .01\*\*, < .05\*).

### 3.1.4 Discussion

This is the first paper that uses computational semantic and temporal measures to clarify the involvement of exploitation and exploration behavior in semantic verbal fluency. These measures were used as a proxy for semantic memory retrieval and executive control processes to unravel which of these processes is the main driving force behind the progressive SVF impairment in AD. This study shows (a) intact or at least effective semantic memory retrieval processes in AD and MCI and (b) impaired, or rather inefficient and ineffective, executive control processes in AD and MCI; in AD, the results suggest that this extends to an overall inability to executively control the production strategy in the SVF task. This is in line with some of the existing literature (Peter et al., 2016; Raoux et al., 2008; Shao et al., 2014).

#### *Exploitation vs. exploration behavior in the SVF*

Qualitative analysis of the SVF has to consider two different layers of information: the semantic content as well as the temporal production of the words produced. In this paper semantic content of words was analyzed with a temporally defined cluster structure and temporal production of those words was analyzed with a semantically defined structure of clusters.

Importantly, neither ADs nor MCIs were found to produce smaller clusters than Cs—no matter whether clusters were defined by computational semantics, temporally or taxonomically—nor did they produce semantically or temporally different clusters. Furthermore, the semantic proximity, a measure for the semantic closeness of words produced within temporally defined clusters, was high for all groups. In other words, irrespective of the diagnostic group, all participants were able to produce a comparable amount of semantically related in-category items in direct succession. The exact same pattern was also found for temporal transition duration within semantically defined clusters; all groups produced semantically related words in fast temporal succession even though clusters were defined semantically.

On the contrary, the data showed a strong effect of semantic proximity increase between temporal clusters across the groups ( $C = MCI < AD$ ). More precisely, the clusters semantically converged over the groups in a way that, in the AD group, clusters were semantically as close as in-category words within a temporal cluster (compare also Figure 11). Importantly, this effect was accompanied by an increase in transition duration between

semantically defined clusters ( $C < AD$ ), showing that ADs took much more time switching from one semantic cluster to another (compare also Figure 12).

This was accompanied by a decrease in computational semantically, temporally as well as taxonomically defined switches ( $C = MCI > AD$ ). Notably, this effect has an entirely different importance than classic semantic-only qualitative approaches, where semantic closeness is inherent to clusters as defined by taxonomy (Mayr, 2002). The same accounts for the temporal transition duration effect where a cluster framework defined by computational semantics allows an independent analysis of temporal production.

By choosing a temporal (and semantic) clustering framework and modelling semantics (and transition duration respectively) on top of it, these results explicitly model the difficulties subjects have within or in-between clusters in terms of effectiveness and efficiency. Overall—as the SVF impairment increases from Cs over MCIs to ADs—the exploration behavior becomes more and more ineffective. This impairment in the *effectiveness* of exploration behavior was found to the extent that ADs completely fail at exploring ‘new semantically unrelated terrain’ (Hills et al., 2015) and thus perform poorly in this task, which demands to generate multiple *different* instances within one category. However, there was no impairment in effective exploitation behavior, quantified by neither traditional taxonomic cluster size metrics or temporal cluster size metrics, nor global semantic relatedness metrics. Importantly, this effect was not caused by an inflationary occurrence of repetitions in the AD group, as repetitions were constant across all groups.

Furthermore, across the three different groups, there is an impairment in the *efficiency* of exploration behavior, showing that ADs are not only less effective but also take more time to explore and thus performing poorly in this task, which additionally demands to generate multiple instances within one category in a *limited amount of time*. Results also show that ADs are marked by inefficiency in exploitation within a cluster. Both inefficient/time-consuming clustering and switching processes fit well the overall SVF impairment in AD; please note that the individual temporal cluster definition ensures that clusters are always temporally coherent, meaning that between cluster switches are always longer than within cluster word transitions.

After applying Bonferroni correction, semantic proximity as well as temporal transition duration between clusters and the related difference of within and between clusters, as well

as all three number of switching measures (NOS semantic, NOS temporal & NOS Troyer), showed no significant difference comparing MCI and C. However, effect sizes (quantified by Cohen's  $d$ ) are at least medium throughout. As there is a clear decline in SVF-count from C to MCI and all exploitation measures reveal neither main group effects nor important descriptive differences between C and MCI, these medium effect sizes serve as promising indicator for a similar but less pronounced pattern of exploration impairment in MCI as in AD. In other words, MCIs seem to show a structurally similar ineffective and inefficient exploration and inefficient exploitation pattern as compared to ADs. Despite medium effect sizes, these MCI-C pairwise comparisons remain insignificant after correction.

*Traditional qualitative metrics' relation to previous findings*

Overall, the reported traditional metrics are in line with previous work: the overall SVF counts lie within the range reported previously (Murphy et al., 2006; Price et al., 2012; Raoux et al., 2008), the observed cluster sizes are in line with some of the previous literature (March & Pattison, 2006; Murphy et al., 2006) as are the number of switches (Gomez & White, 2006; Price et al., 2012). However, there are studies that report smaller cluster sizes (Gomez & White, 2006; Raoux et al., 2008), but in these studies repetitions/perseverations had been excluded entirely from the analysis, which reduces the size of clusters in general. The pattern of main group effect for number of switches but no effect for cluster size has been reported previously (Peter et al., 2016; Raoux et al., 2008), but there are other studies which reported the exact inverse (March & Pattison, 2006; Murphy et al., 2006). As mentioned at the beginning, this inconsistency might be due to the lack of objective and reliable approaches for deriving cluster boundaries. Conversely this study used an algorithmic implementation of Troyer, Moscovitch, and Winocur's approach (A. K. Troyer et al., 1997) and an objective completely automated temporal approach as redundancy eliminating such shortcomings.

Though beyond the scope of this paper's novel methodology, future work should investigate how the here-presented novel qualitative temporal measures based on a computational semantics clustering framework would hold as computed on the basis of the traditional taxonomic semantic clustering framework.

*Methodological considerations and implications for the Assessment of AD*

This paper also introduces a new automatic qualitative analysis scheme of the SVF task which has implications for the neuropsychological practice. The results show that temporal

switches as well as semantic switches display the same pattern as classic taxonomic ones (A. K. Troyer et al., 1998). Moreover, words within a temporal cluster are semantically stronger related than between temporal clusters—hereby the notion of semantic relation is based upon objective/computationally derived semantic proximity. This shows that temporal clusters are capable of rendering semantic relations in the SVF performance without conducting time-consuming manual qualitative taxonomic semantic analysis. The fact that there is a higher overall semantic proximity within temporal clusters than between temporal clusters is in line with the original notion, that “words that comprise [...] temporal clusters tend to be semantically related” (A. K. Troyer et al., 1997) see also (Collins & Quillian, 1969). It is noteworthy that the methodological approach used represents an inverse approach to the original one (A. K. Troyer et al., 1997), which tried to infer successful/fast semantic retrieval through qualitatively analyzing the transcripts based on a manual taxonomy. Therefore, the authors highly recommend future studies to not only transcribe SVF data but record the audio, so that the temporal variance is not lost in the process as it has proven to be crucial for the qualitative interpretation of SVF impairments.

### 3.1.5 *Limitations*

The presented pairwise comparisons between MCI and C for the reported main effects show at least medium effect sizes (J. Cohen, 2013) but corrected  $p$ -values remain insignificant which might be due to the non-parametric analysis method. As for the hereunder reported novel measures the authors lacked clear hypothesis deducted from literature, future studies should take here-reported results as a starting point and focus explicitly on the MCI vs. C effects to better unveil potential differences in exploration and exploitation patterns between those groups.

The data presented in this study has been recorded in non-laboratory everyday clinical settings which might result in the relative in-homogeneity of the data. On the other hand, the results should have a relatively high ecological validity due to the same reasons.

Finally, a major limitation of these results is the missing direct link between the novel measures and gold-standard measures for executive control processes as well as semantic memory retrieval processes. Correlation with non-SVF markers for executive control processes and semantic memory retrieval processes could strengthen this.

### 3.1.6 Conclusion

Results of a combination of computational temporal and semantic measures were presented. Merging both approaches, the SVF can be analyzed more comprehensively by clearly modelling the difficulties subjects have within or in-between clusters in terms of effectiveness and efficiency. By the means of this new analysis scheme this paper provides evidence for *effective* but *inefficient* exploitation and *inefficient* and *ineffective* exploration in AD and most probably also MCI; given the related body of research this can be interpreted as effective but inefficient semantic memory retrieval processes and inefficient and ineffective cognitive control processes. This is in line with previous literature pointing towards impaired executive control processes in the SVF rather than semantic memory retrieval processes (Clark et al., 2016; Hills et al., 2015; Lerner, Ogrocki, & Thomas, 2009; Linz et al., 2019; Linz, Tröger, Alexandersson, & König, 2017). This approach presents a valuable and feasible extension to traditional qualitative SVF analysis, better differentiating between semantic memory retrieval and executive control processes. Future research is needed to demonstrate the validity of such an analysis scheme for etiologies with accentuated semantic memory retrieval impairment or executive functions impairment, as well as diagnosis of different focal lesions. From a methodological standpoint the authors conclude that—given the appropriate analysis methods—the SVF can be relevant for dedicated assessment of different neurocognitive functions.

**Acknowledgements** This research was partially funded by the EIT Digital Wellbeing Activity 17074, ELEMENT. The data was partially collected during the EU FP7 Dem@Care project, grant agreement 288199. The authors like to thank Hali Lindsay and Katja Häuser for helpful feedback on an earlier version of the manuscript.

3.2 PATIENTS WITH AMNESTIC MCI FAIL TO ADAPT EXECUTIVE CONTROL WHEN REPEATEDLY TESTED WITH SEMANTIC VERBAL FLUENCY TASKS

Johannes Tröger<sup>1</sup>, Hali Lindsay<sup>1</sup>, Mario Mina<sup>1</sup>, Nicklas Linz<sup>2</sup>, Stefan Klöppel<sup>3</sup>, Jutta Kray<sup>4</sup>, Jessica Peter<sup>3</sup>

<sup>1</sup> German Research Center for Artificial Intelligence (DFKI), Germany

<sup>2</sup> ki elements, Germany

<sup>3</sup> University Hospital of Old Age Psychiatry and Psychotherapy, University of Bern, Switzerland

<sup>4</sup> Chair for Development of Language, Learning & Action, University of Saarland, Germany

Semantic verbal fluency (SVF) tasks require individuals to name items from a specified category within a fixed time. An impaired SVF performance is well documented in patients with amnesic Mild Cognitive Impairment (aMCI). The two leading theoretical views suggest either loss of semantic knowledge or impaired executive control to be responsible. We assessed SVF three times on two consecutive days in 29 healthy volunteers (HC) and 29 patients with aMCI with the aim to answer the question which of the two views holds true. When doing the task for the first time, patients with aMCI produced fewer and more common words with a shorter mean response latency. When tested repeatedly, only healthy volunteers increased performance. Likewise, only the performance of HC indicated two distinct retrieval processes: a prompt retrieval of readily available items in the beginning of the task and an active search through semantic space towards the end. With repeated assessment, the pool of readily available items became larger in HC but not in patients with aMCI. The production of fewer and more common words in aMCI points to a smaller search set and supports the loss of semantic knowledge view. The failure to improve performance as well as the lack of distinct retrieval processes point to an additional impairment in executive control. Our data did not clearly favor one theoretical view over the other but rather indicates that the impairment of patients with aMCI in SVF is due to a combination of both.

*Keywords:* Semantic Verbal Fluency, amnesic MCI, Temporal Analysis, Semantic Loss, Executive Control, Practice effects

The previous chapter helped to establish a framework on the differentiation between executive function and semantic memory processes impaired in AD as measured by the SVF. However, the presented comparisons between MCI and healthy controls showed medium effect sizes but remained insignificant due to alpha value corrections. MCI or more precise amnesic MCI (aMCI) is considered to mark the early clinical stage of future AD dementia but qualitative differences in the SVF become considerably harder to measure because of the relatively intact level of cognitive functioning in aMCI. This is why, driven by clear hypotheses and with a slightly adapted methodological design, one should focus explicitly on the *MCI vs. C* effects to better unveil potential differences in executive functions and semantic memory impairments between those groups.

### 3.2.1 Introduction

In semantic verbal fluency (SVF) tasks, individuals need to generate and retrieve as many different items from a specified category as possible within a certain amount of time. Successful retrieval requires the interplay of at least two cognitive components: A semantic component, associated with the integrity of lexico-semantic networks and an executive component, related to strategic search and retrieval processes (Amunts et al., 2020; Shao et al., 2014). An impaired SVF performance is well documented in patients with dementia due to Alzheimer's Disease (AD) or its prodromal stage amnesic Mild Cognitive Impairment (aMCI) (Auriacombe et al., 2006; Gomez & White, 2006; Henry et al., 2004; Pakhomov et al., 2016; Raoux et al., 2008). However, there remains widespread disagreement as to what this impairment reflects. The two leading theoretical views either suggest loss of semantic knowledge (i.e., structural view; Rohrer et al., 1995) or impaired executive control mechanisms (i.e., procedural view; Fernaeus & Almkvist, 1998). These control mechanisms include a strategic or non-strategic search through the semantic space (D. G. Clark et al., 2016; Hills et al., 2015; Lerner, Ogrocki, & Thomas, 2009; Linz et al., 2019; Linz, Tröger, Alexandersson, & König, 2017) as well as monitoring processes to suppress previously mentioned items or items that do not belong to the category.

Evidence for the structural view stems from the latency of word production in SVF tasks. Rohrer and colleagues (1995) posit that verbal fluency performance depends on the number of words available in the semantic space and the time it takes to retrieve them. 'Latency' thus is the sum of the number of seconds from the first word to each of the

subsequent words divided by the number of words produced (see Figure 13 for an example). The structural view posits that associations between a category name and its members become weaker (i.e., the semantic space disintegrates) and thus, the activation of that category name as a retrieval cue will result in the activation of fewer category members (Figure 13). With fewer available category members within the semantic space, less time is needed to find them. As a consequence, the mean latency of word production becomes shorter. Patients with aMCI or AD typically show shorter mean response latencies in combination with a reduced word count (Randolph et al., 1993; Rohrer et al., 1995; Tröster et al., 1989). In addition, they typically generate highly semantically related words, which means that they stick to answers that are most commonly given by people in such a task, indicating that they are unable to fully explore their semantic space.

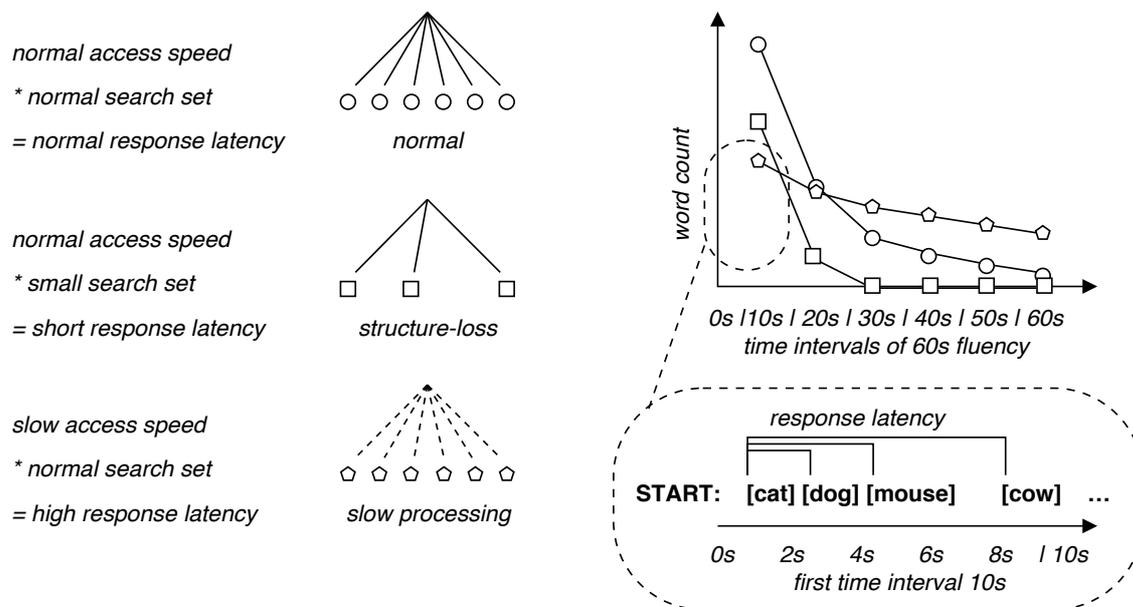


Figure 13: Influence of the structural basis of the semantic space and processing speed on the latency of word production. Structural loss results in a smaller semantic space. As a consequence, fewer words are available and less time is needed to retrieve them (i.e., the latency becomes shorter). In contrast, decreased processing speed without structural loss results in slower retrieval (i.e., longer latency, E). The first word (at 1s) is the starting point and the response latency of the second or third word is 3s or 7s, (i.e., 4-1 or 8-1).

The procedural view similarly posits that patients with aMCI or AD are unable to fully explore their semantic space (Tröger et al., 2019). In contrast to the structural view, however, the procedural view suggests that patients with aMCI or AD have difficulties adapting executive control. The majority of responses in SVF tasks are given very early in the task and considerably fewer, if any, towards the end. Two retrieval modes seem to be responsible for this pattern of word production; the majority of responses are given in an automatic retrieval

mode associated with rapid word production in the beginning of the task, while then a more effortful retrieval follows towards the end. Consequently, responses given early in the task are more common (i.e., frequent) in the respective language than responses given towards the end (Linz et al., 2019). Thus, in the production of words during SVF tasks the retrieval of 'easy-to-access' responses in the beginning can be distinguished from less common responses requiring more effort once the easy-to-access objects have been exhausted. Patients with aMCI or AD seem to have problems with adapting their search strategy towards this effortful retrieval.

So far, most studies assessed SVF tasks only once although the repeated assessment of SVF performance could help to answer the question which of the two views holds true. At first assessment, an impairment in SVF performance can reflect both structure loss and impairment in executive control (or a combination of both). Practising a task, however, can improve the way a person solves the task and thus performance. In SVF tasks, participants may improve by adapting executive control or by changing strategies to become more successful. Only few studies so far have investigated changes in SVF production with repeated assessment. These studies reported that patients with aMCI do not (or only slightly) improve compared to HC (Cooper, Lacritz, Weiner, Rosenberg, & Cullum, 2004; Duff et al., 2008, 2011). However, these studies focused on a quantitative analysis of SVF performance (i.e., the number of retrieved words) but did not consider qualitative aspects (e.g., retrieval modes, word frequency, or latency of word production). Thus, they did not try to provide an explanation *why* (and in what way) healthy volunteers improve and patients with aMCI do not. With the current study, we will close this gap, thereby possibly helping to elucidate which of the two views holds true.

### 3.2.2 *Methods*

#### *Participants*

We included n=58 participants in this study: 29 patients with aMCI (age range = 60–81 years) and 29 HC (age range = 61–81 years; Table 5). We recruited patients with aMCI from the Centre for Geriatric Medicine and Gerontology at the University Medical Centre Freiburg, Germany where they received their diagnosis. HCs were recruited via newspaper advertisement and flyers circulated in Freiburg, Germany. All participants gave written

informed consent prior to testing. The Ethics Committee of Freiburg University approved the study. The study conforms to the Declaration of Helsinki.

*Table 5: Sociodemographic characteristic of the sample (mean and standard deviations). HC = Healthy Controls, aMCI = Patients with amnesic Mild Cognitive Impairment, f/m = female / male, MoCA = Montreal Cognitive Assessment.*

	HC	aMCI	<i>t</i>	<i>p</i>
n	29	29		
Gender (f/m)	19/10	14/15		0.19 ( $\chi^2$ )
Age (years)	71.10 ± 4.74	73.21 ± 4.77	1.68	0.10
Education	14.66 ± 3.36	13.34 ± 3.31	1.49	0.14
MoCA	26.83 ± 1.91	22.07 ± 3.28	6.74	< 0.0001
Verbal intelligence	120.10 ± 12.44	114.90 ± 11.73	1.64	0.11

#### *Inclusion and exclusion criteria*

The study followed a standardised protocol. Participants were first screened over the phone. They had to be fluent in German, with normal or corrected-to-normal vision and no history of psychiatric or neurological disorders. Further exclusion criteria were current use of psychotropic medication, current or life-time drug abuse or addiction, brain damage, or sleep disorders. We evaluated depressive symptoms with the Geriatric Depression Scale (GDS; Yesavage & Sheikh, 1986) and included those with GDS ≤ 6 (Yesavage & Sheikh, 1986).

For patients with aMCI, cognitive functioning was evaluated with the neuropsychological battery from the Consortium to Establish a Registry of Alzheimer's disease (CERAD-plus) (J. C. Morris et al., 1989) during the diagnostic process in the memory clinic. They also received MR Imaging, laboratory diagnostics, and functional assessment during the diagnostic process. To be diagnosed with aMCI, they had to show impairment in the delayed recall of a previously learned list of words (1.5 SD below age-, gender-, and education adjusted norms). Additionally, they needed to a) report memory complaints, b) show no impairment in activities of daily living, and c) no dementia according to established criteria (Petersen, 2004). They also needed to fulfil criteria for a diagnosis of MCI due to AD with intermediate certainty according to revised criteria (Petersen et al., 2014). That is, they needed to exhibit signs of neuronal injury (i.e., hippocampal volume or medial temporal atrophy by volumetric measures of visual rating). Healthy elderly volunteers were included with a MoCA score ≥ 23 as recommended by Carson, Leach and Murphy (2018).

*Procedure*

We collected data on two consecutive days. On day one, participants completed the Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005) and then we administered other neuropsychological tests including a test of verbal intelligence (Lehrl, 2005) as well as SVF. After a pause, we applied the SVF task a second time. On day two, we administered the SVF task a third time and applied other cognitive tests.

*Semantic verbal fluency task*

We instructed participants to produce as many different four-legged animals as possible within 60 seconds and to avoid repetitions. We collected speech recordings of all participants with a microphone on a computer and trained students from the field of computational linguistics subsequently to transcribe these recordings in PRAAT (Boersma & Weenink, Version 6.1.42 ). We obtained the following measures for statistical analyses:

*Word Count.* We calculated the number of words produced within 60 seconds, excluding the number of repetitions. We followed the approach suggested by Linz and colleagues (2019) and included only unique, correct words to the participants' word count. To examine the change in participants' performance over the 60s, we segmented the transcript into six 10s time intervals. We sorted words into these time intervals based on their speech onset. Given that they performed the task three times, we obtained 18 data points for each participant (6 intervals\*3 assessments).

*Mean response latency.* We computed the mean response latency ( $\tau$ ) according to Rohrer and colleagues (1995). The first uttered word ( $w_1$ ) was used as the onset of the semantic verbal fluency production sequence. Then, we calculated the time that had elapsed since the onset of this word (i.e.,  $w_1$ ) until the onset of any other word in the production sequence ( $w_i$ ), which would represent the subsequent response latency for these other words, according to Rohrer and colleagues. Next, we calculated the sum of all response latencies and divided it by the total number of words ( $n$ ).

$$\tau = \sum_{i=2}^n \frac{w_i - w_1}{n}$$

*Mean Word Frequency.* Comparable to our previous study (Linz, Tröger, Alexandersson, Wolters, et al., 2017), we calculated the mean word frequency (MWF) using the Python wordfreq package (Speer et al., 2018), which combines resources such as

Wikipedia, news, and book corpora as well as Twitter. We calculated the MWF, by summing up the frequencies ( $a$ ) of all correctly produced words ( $i$ ) divided by the word count ( $n$ ).

$$MWF = \frac{1}{n} \sum_{i=1}^n f(a_i)$$

*Mean Temporal Cluster Size and Number of Temporal Switches.* Comparable to our previous work (Tröger et al., 2019), we computed clusters and switches temporally. A cluster consisted of a minimum of two words belonging to the same cluster as defined by a temporal threshold. In order to determine temporal clusters, each word  $w_i$  was assigned a start time  $s_i$  and an end time  $e_i$  according to its position in the speech recording. Clusters were then determined iteratively. The first word  $w_0$  started a new cluster. The next word  $w_j$  was part of the cluster, if the distance between its start time  $s_j$  and the previous words' end time  $e_{(j-1)}$  was below a threshold  $t_j$  (i.e.,  $s_j - e_{(j-1)} < t_j$ ). A base threshold  $t$  was determined for a speaker as the mean distance between any consecutive words produced by the speaker. To account for a decrease in word production towards the end of the task, this base threshold was linearly scaled by a maximum factor of two, based on the start of the current word  $w_j$  (i.e.,  $t_j = t * (s_j / 60 + 1)$ ). This approach automatically accounts for inter-personal differences in terms of production speed and reaction time. The mean cluster size was calculated as the sum of cluster sizes divided by the number of clusters. Finally, the number of switching clusters was defined as the total number of switches between clusters, including single word clusters.

#### *Statistical Analysis*

We performed statistical analysis with R (software version 3.4.02). As dependent variables, we used word count, mean response latency or mean word frequency. All analyses consisted of two within-subject factors assessment (t1, t2, and t3) and time interval (0s-10s, 10s-20s, 20s-30s, 30s-40s, 40s-50s, and 50s-60s) as well as one between-subjects factor diagnosis (aMCI and HC); resulting in an overall experimental design of 3\*6\*2. For the analysis of main effects and interactions, we used analysis of variance. For the analysis of repeated assessment effects, we used two planned contrasts (t1, t2, t3 [1, -1, 0] and t1, t2, t3 [0, 1, -1]). Statistical significance levels were set to  $p < 0.05$  and we corrected for multiple testing with the Bonferroni-Holm procedure.

### 3.2.3 Results

#### Word count

Patients with aMCI produced significantly fewer words than HC [i.e., main effect of diagnosis;  $F_{(1, 56)} = 66.04, p < .001$ ]. In addition, we found that the production of words changed significantly over repeated assessment [ $F_{(2, 952)} = 3.99, p < .05$ ] as well as across the six time intervals [ $F_{(5, 952)} = 142.28, p < .001$ ]. However, a significant interaction between assessment\*diagnosis indicated that the effect of repeated assessment was different for HC and patients with aMCI. Indeed, we found no significant improvement in the aMCI group but HC significantly increased their word count when doing the task repeatedly. Planned contrasts indicated that this was due to a significant change between t2 and t3 [ $t_{(1, 320)} = 4.31, p < .01$ ; Figure 14]. An exploratory analysis in this group revealed a clear distinction between two retrieval modes across the time intervals at first assessment [i.e., significant interaction between assessment\*time interval,  $F_{(5, 308)} = 3.2, p < .01$ ]: Healthy volunteers produced significantly more words during the first 20 seconds of the task than during the remaining 40 seconds [ $t_{(1,28)} = 5.0, p < .001$ ; Figure 14B]. At t2 and t3, this clear distinction was no longer visible. When doing the task repeatedly, the increase in word count happened primarily during the first half of the task, while no increase was observed during the second half (Figure 15).

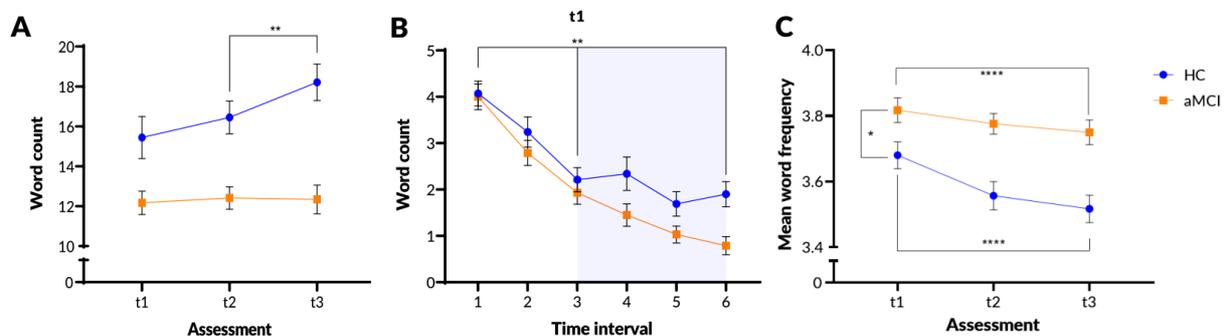


Figure 14: Change in semantic verbal fluency performance (i.e., word count) in healthy controls (HC) and patients with amnesic Mild Cognitive Impairment (aMCI) over three assessments (t1-t3; A). The performance in HC suggests two different retrieval modes at first assessment (B). Mean word frequency was higher in aMCI but changed similarly to the HC group with repeated assessment (C). Error bars represent the standard error of the mean.

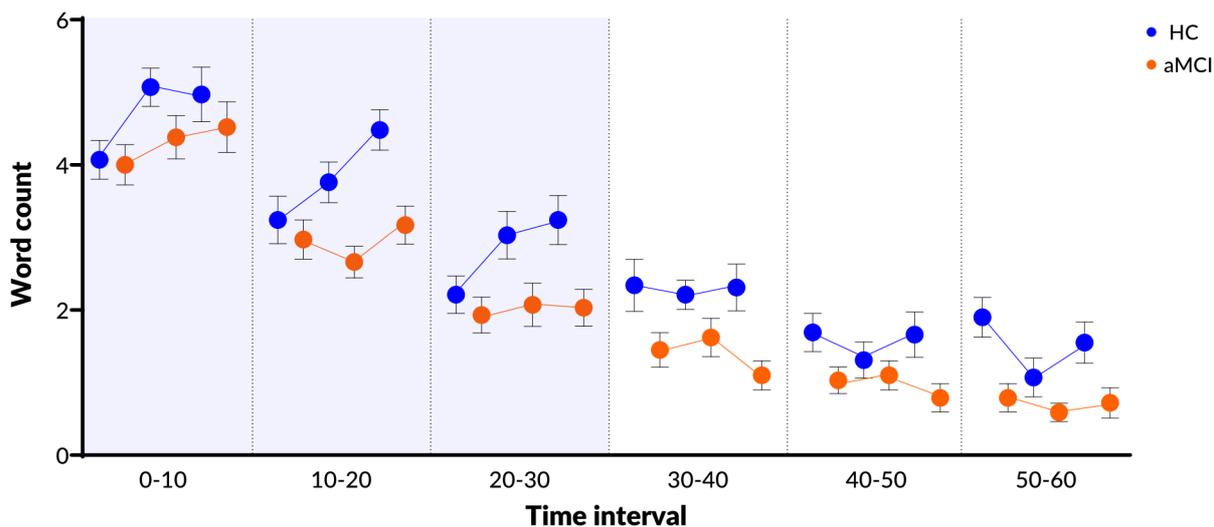


Figure 15: Change in semantic verbal fluency performance (i.e., word count) in healthy controls (HC) as well as patients with amnesic Mild Cognitive Impairment (aMCI) over three assessments. The performance was split into six time intervals with 10 seconds each. Healthy controls particularly increased performance during the first half of the task as highlighted in blue. Error bars represent the standard error of the mean.

#### Mean response latency

We found a shorter mean response latency in patients with aMCI than in HC [HC:  $23.32 \pm 4.88$  s, aMCI:  $18.99 \pm 5.03$  s;  $F_{(1,56)} = 8.02$ ,  $p < .05$ ] and a shorter mean response latency as both groups did the task repeatedly [ $F_{(2,112)} = 3.37$ ,  $p < .05$ ]. The latter was comparable for HC and patients with aMCI since we found no significant interaction between time interval\*diagnosis.

#### Mean Word frequency

At first assessment, the mean frequency of words was higher in patients with aMCI than in HC [ $F_{(1,56)} = 3.95$ ,  $p < .05$ ; Figure 14]. Both groups retrieved less frequent words towards the end of the task [ $F_{(5,952)} = 76.2$ ,  $p < .001$ ]. The latter was comparable for HC and patients with aMCI since we found no significant interaction between time interval\*diagnosis.

#### Mean Temporal Cluster Size and Number of Temporal Switches

We found a significant main effect of diagnosis for both mean cluster size [ $F_{(1,56)} = 8.26$ ,  $p < .01$ ] and number of switches [ $F_{(1,56)} = 8.92$ ,  $p < .01$ ]. This indicates that patients with MCI showed significantly smaller clusters and switches less often than healthy volunteers (Figure 16; Table 6). For the number of switches, we additionally found a statistical trend for an interaction between diagnosis\*assessment [ $F_{(2, 112)} = 2.62$ ,  $p = .07$ ]. With repeated

assessment, healthy volunteers switched significantly more often, while patients with MCI did not (Figure 16; Table 6).

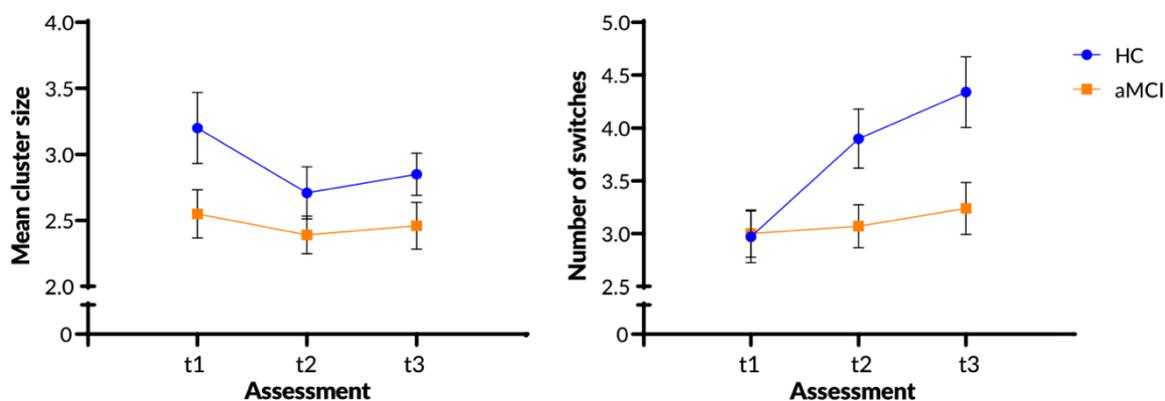


Figure 16: Number of clusters (left) and number of switches (right) during repeated assessment of semantic verbal fluency in healthy volunteers (HC) and patients with Mild Cognitive Impairment (aMCI). Error bars represent the standard error of the mean.

Table 6: Means  $\pm$  standard deviations of mean cluster size and number of switches during semantic verbal fluency tasks at three different assessments (t1, t2, t3). aMCI = patients with amnesic Mild Cognitive Impairment.

	Healthy Controls			aMCI		
	t1	t2	t3	t1	t2	t3
Mean Cluster Size	3.20 $\pm$ 1.45	2.71 $\pm$ 1.06	2.85 $\pm$ 0.86	2.55 $\pm$ 0.99	2.39 $\pm$ 0.77	2.46 $\pm$ 0.95
Number of Switches	2.97 $\pm$ 1.32	3.90 $\pm$ 1.50	4.34 $\pm$ 1.80	3.00 $\pm$ 1.22	3.07 $\pm$ 1.10	3.24 $\pm$ 1.33

### 3.2.4 Discussion

In the current study, we examined whether the well documented impairment in SVF task performance in patients with aMCI that we also found in the current study reflects a loss of semantic knowledge (i.e., structural view) or a failure to adapt executive control (i.e., procedural view). Therefore, we had patients with aMCI and healthy volunteers perform an SVF task repeatedly. In line with previous research we found that only the performance of healthy volunteers improved while that of patients with aMCI did not (Cooper et al., 2004; Duff et al., 2008, 2011).

When patients with aMCI did the task for the first time, they retrieved fewer (and more frequent) words (Figure 14) with a shorter mean response latency than healthy volunteers. A reduced word count alone can be explained by either a smaller semantic space (due to a loss of semantic knowledge) or by a slower word production. Since patients with aMCI produced fewer words in combination with a shorter mean response latency, the results of the first assessment are more in favour of a loss of semantic knowledge (or at least a less accessible

semantic space). This is because successful SVF task performance requires the location, activation, and retrieval of specific members of a category. If associations between the category and its members become weaker (i.e., the semantic space disintegrates), the activation of that category name as a retrieval cue will activate fewer members. With fewer members within the semantic space, less time is needed to find them. Thus, the response latency becomes shorter. An alternative explanation for the reduced performance in patients with aMCI could be that the search process in the beginning of the task has been compromised and has thus been more effortful in patients with aMCI. If correct, they would have been expected to be slower at initiating search processes and retrieving words from memory even for easily accessible words and, hence, have a mean response latency *longer* than healthy volunteers (see Figure 13). However, patients with aMCI had a *shorter* mean response latency than HC. The relative distribution of responses during the task determines the mean response latency and therefore, shorter mean response latencies indicate that patients with aMCI retrieved items predominantly at the beginning of the task and quickly exhaust their pool of accessible items. Hence, the combination of having produced fewer and more frequent (i.e., more common) words with a shorter mean response latency points to a smaller semantic space containing more commonly used words and supports the structural view - at least at first assessment (Randolph et al., 1993; Rohrer et al., 1995; Tröster et al., 1989).

We also found evidence to support the procedural (i.e., executive) view. Comparable to other studies, we found that healthy volunteers employed two different retrieval strategies; an automatic retrieval in the beginning of the task and an effortful retrieval towards the end (Fernaes & Almkvist, 1998; Linz et al., 2019)(Figure 14). In the beginning of the task, the automatic retrieval occurs from a pool of readily available words; that is, these are commonly used and easily accessible items. As time passes by and this initial pool of words is exhausted, word generation becomes more challenging, thus requiring more cognitive effort (i.e., more executive control). Our results suggest that healthy volunteers were able to make these additional efforts (at least at first assessment) but we did not observe this in patients with aMCI. Their performance at first assessment did linearly decrease and did not suggest that they employed different retrieval strategies (Figure 14). As already mentioned, this could indicate both structural and procedural deficits. However, they also did not engage more executive control with repeated assessment. With repeated assessment, a change in executive control seems more likely than a change in semantic knowledge (at least when the

repeated assessment happens in short succession). Our group of healthy volunteers particularly improved during the first 30 seconds of word production when doing the task repeatedly (Figure 15, highlighted in blue). Since the mean word frequency did not significantly change during that time interval, the increase in performance was most probably not due to an increased production of less frequent words. Instead, the pool of readily available – easy-to-access - items seems to have become larger in HC. Probably, the activation of a member of the category simultaneously activated the semantic neighbourhood. Since members of a category are organized in networks, the members are represented as a system of nodes and links, as opposed to isolated pairs (Goñi et al., 2011). Thus, a mental process (i.e., location, activation, and/or retrieval) operating on one member of the category may have changed the states of related words in the network, thereby enhancing the likelihood of these related words to be activated and retrieved when tested repeatedly. Consequently, with every repetition, more easy-to-access words became readily available. This was not the case in patients with aMCI. When doing the task repeatedly, they did neither improve in the first 30 seconds of the task nor in the final 30 seconds (Figure 15).

Another explanation may be that healthy volunteers became more familiar with the task due to repeated practice. It could be that they remembered the responses from the previous assessment, became quicker in retrieving them with enough time to search for new items that they had not retrieved previously. This would, however, again require increased executive control since they would need to keep every response in their working memory (i.e., monitor every response) and inhibit answers already given – with increasing word count, this would become more difficult and would require more executive control. Patients with aMCI did not show this, which supports that they exhibit a problem with executive control that becomes most apparent with repeated assessment. The procedural view is also supported by data from phonemic fluency although this task was only used in patients with aMCI during the diagnostic process in the memory clinic. For phonemic verbal fluency, an individual is asked to generate as many different items starting with a certain letter (e.g., 'F') as possible. Semantic verbal fluency requires a strategic search through the semantic knowledge store (i.e., the semantic space), while phonemic fluency depends more on knowledge of word spelling and phonemic relatedness. Patients with AD typically show larger impairment in SVF than in phonemic verbal fluency. Yet, for patients with aMCI, this is not necessarily the case (Brandt & Manning, 2009; Nutter-Upham et al., 2008). Comparable to previous studies,

patients with aMCI in our sample were better in their SVF performance than in their phonemic verbal fluency performance. This may indicate that they had more problems with executive functions than with a search through the semantic space. However, at first assessment the reduced word count in combination with the shorter mean response latency rather points to a loss of semantic knowledge. Therefore, our data suggests that patients with aMCI are impaired in SVF tasks due to loss of semantic knowledge in combination with a failure to adapt executive control.

### 3.2.5 Limitations

Our study may have several limitations. First, we posited that responses given earlier in the task are typically more common and that this reflects an automatic retrieval mode. However, the cultural milieu of the participants or other variables (e.g., education) may also influence the order of word generation. For the current study, we matched participants according to education and they all needed to be fluent in German. In addition, all of them were Caucasian and none of them had an immigrant background. Therefore, it seems unlikely that a difference in culture (or education) explains our findings.

Another possible limitation might be that we assessed phonemic fluency performance only in patients with aMCI and not in healthy controls. Therefore, a direct comparison between both groups was not possible.

Finally, our study may be limited by the fact that we included both single-domain amnesic MCI (n=8) and multiple-domain amnesic MCI (n=21). However, we found no significant differences between both groups regarding age, education, premorbid intelligence, or MoCA score. We also did not find any significant differences in word count, mean latency, or word frequency.

**Acknowledgements** All authors declare that there are no conflicts of interest. This work was supported by EIT-Digital, Area 2. Innovation and Research, Segment 2.3 WEL (Digital Wellbeing), Grant number 19141.

### 3.3 CHAPTER CONCLUSION

The aim of this chapter was to investigate how different neurocognitive processes (i.e. executive functions and semantic memory) are impaired at different clinical stages of AD as measured by the SVF. In order to disentangle the impairment of both neurocognitive constructs within the SVF AI-driven qualitative temporal and semantic analysis methods have been employed.

In two studies reported above, SVF impairments were found at both clinical cognitive stages of (probable) AD: aMCI and AD dementia. In the first study computational temporal and semantic measures profiled the effectiveness and efficiency in patients' SVF performance. The results indicate that AD patients' impaired SVF production patterns imply effective but inefficient semantic memory retrieval processes and inefficient and ineffective executive functions. In other words, the findings support the thesis that in AD dementia the impairment observable via the SVF is majorly caused by hampered executive functions. This is in line with previous research on executive function impairments in AD clinical stages (Peter et al., 2016; Raoux et al., 2008; Shao et al., 2014). There were strong effect sizes in MCI, displaying the same pattern as in AD dementia, albeit statistically remaining a trend and not reaching statistical significance. Hence in the second study an experimental setup with a higher sensitivity for measuring different cognitive function impairments in MCI was chosen. Through the repeated administration of the SVF task, a) pure novelty effects (Thorgusen, Suchy, Chelune, & Baucom, 2016) were ruled out and b) the influence of executive function was made more visible. Novelty effects are defined as an initial suppression of performance in face of a novel task and represent a common confounder in cognitive assessment at older ages (Suchy, Kraybill, & Franchow, 2011). Through repeated testing the novelty of the task is lost and participants recover to their actual best-case performance. In addition to reduced novelty effects, practice effects can be expected through the repeated task exposure. Depending on the task, practice effects can be an indicator of memory (memory tasks in multiple sessions using the same material) but in a complex open task like the SVF they can be indicative of improved cognitive control mechanisms adapting to the task and in return can be interpreted as a proxy of executive function. Indeed, in the second study steadily improving SVF performance was only found for the healthy control group but not for the aMCI group. This improvement was especially pronounced during the first half of the SVF word production in

the consecutive assessments, suggesting that healthy controls increased in processing speed. Additionally, healthy controls switched a lot more during repeated assessments, indicating an increasingly more efficient exploration strategy. In contrast, the aMCI group did not improve in their performance and also started at a lower baseline in the initial assessment. Hence, the lack of improvement in aMCI reflects a structural *and* a procedural deficit. Their deficit in SVF tasks may have been structural in the first place, but they also did not employ appropriate executive control mechanisms (e.g., adjusting their retrieval strategy) to become more successful.

To conclude, the data of the two above-reported studies suggests that aMCI and AD dementia is marked through both semantic memory as well as executive function impairments. However, executive function impairments become more notable in the later dementia stage, whereas semantic memory seems to be most responsible for the initial SVF impairment at an aMCI stage. Nevertheless, the executive functions impairment also becomes apparent at an aMCI stage if repeated testing paradigms or advanced computational qualitative measures are applied.

Novel computational qualitative measures have proven to be able to disentangle the impairment of both neurocognitive constructs within the SVF in AD and bear great potential for automated diagnosis support solutions. The next chapter presenting a second pair of studies leverages exactly this potential and demonstrates the value of computational SVF analysis in scalable and cost-effective clinical applications.

### 4 IMPLICATIONS FOR CLINICAL DECISION MAKING

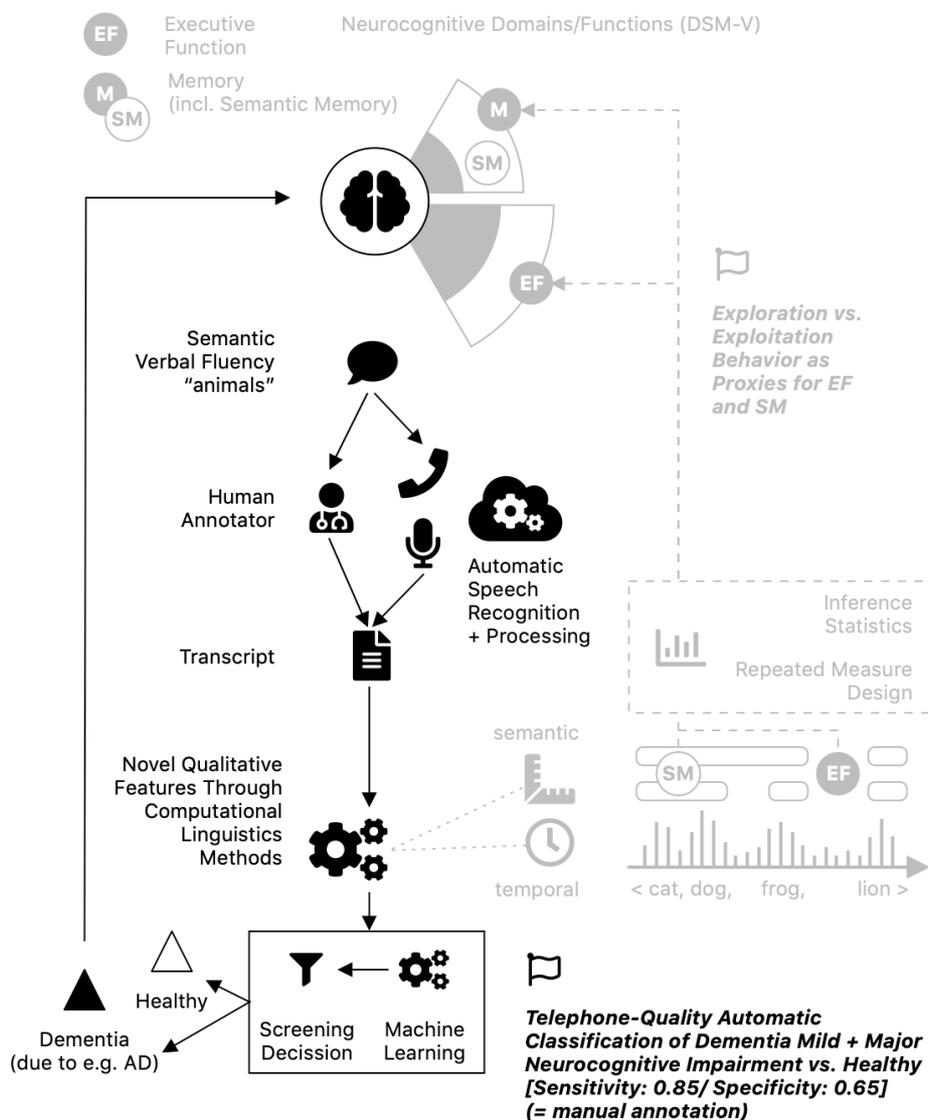


Figure 17: Visual abstract. The major contribution of this chapter is highlighted in black.

Although the SVF is common practice and has been proven extremely sensitive to AD dementia and its probable precursor aMCI, the SVF remains challenging when it comes to differentiating the impairment of different neurocognitive functions and their AD-related profiles. This thesis employed a framework of computational AI-driven qualitative temporal as well as semantic analysis methods improving the signal on the two distinct neurocognitive functions. Despite having great value for future clinical and especially pharmaceutical research, these results have little immediate impact on today’s societal challenge of fighting AD. For this purpose, the afore-established qualitative computational measures should be

harnessed to serve the societal mid-term goal in the fight against AD: To screen for early (cognitive) signs of AD early and at scale.

In the scope of this thesis, *early* means at an early but clinical stage (i.e. MCI). *At scale* means requiring little to no human resources for performing as well as evaluating the assessment and also requiring a minimum technical setup. With the automatization of evaluation processes, the SVF will make for a highly effective screening tool that at the same time poses minimal patient burden (i.e. normally 60s of relatively free speech).

It is a visionary long-term goal for AD research as well as healthcare systems, to rely on a fully automated assessment system that could reach out to a large population on a regular basis (e.g. every month). This requires two separate efforts: (1) establish the feasibility of automatic evaluation and decision making (no human resources for the data evaluation) and (2) establish the feasibility of remote low-tech data collection. This part of the thesis contributes to both efforts: First, by showing how fully automated analysis of the SVF results in comparable downstream diagnostic decision support as compared to human evaluations. Second, by proposing a remote telephone- and SVF-based screening concept and demonstrating its feasibility.

#### 4.1 FULLY AUTOMATIC SPEECH-BASED ANALYSIS OF THE SEMANTIC VERBAL FLUENCY TASK

Alexandra König<sup>1</sup>, Nicklas Linz<sup>2</sup>, Johannes Tröger<sup>2</sup>, Maria Wolters<sup>3</sup>, Jan  
Alexandersson<sup>2</sup>, Phillippe Robert<sup>1</sup>

<sup>1</sup> Memory Clinic, Association IA, CoBTek Lab, CHU Université Côte d'Azur, Nice, France

<sup>2</sup> German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

<sup>3</sup> School of Informatics, University of Edinburgh, Edinburgh, UK

Semantic verbal fluency (SVF) tests are routinely used in screening for mild cognitive impairment (MCI). In this task, participants name as many items as possible of a semantic category under a time constraint. Clinicians measure task performance manually by summing the number of correct words and errors. More fine-grained variables add valuable information to clinical assessment, but are time-consuming. Therefore, the aim of this study is to investigate whether automatic analysis of the SVF could provide these as accurate as manual and thus, support qualitative screening of neurocognitive impairment. SVF data were collected from 95 older people with MCI ( $n = 47$ ), Alzheimer's or related dementias (ADRD;  $n = 24$ ), and healthy controls (HC;  $n = 24$ ). All data were annotated manually and automatically with clusters and switches. The obtained metrics were validated using a classifier to distinguish HC, MCI, and ADRD. Automatically extracted clusters and switches were highly correlated ( $r = 0.9$ ) with manually established values, and performed as well on the classification task separating HC from persons with ADRD (area under curve [AUC] = 0.939) and MCI (AUC = 0.758). The results show that it is possible to automate fine-grained analyses of SVF data for the assessment of cognitive decline.

*Keywords:* Alzheimer's disease, Dementia, Mild cognitive impairment, Neuropsychology, Assessment, Semantic verbal fluency, Speech recognition, Speech processing, Machine learning

##### 4.1.1 Introduction

As life expectancy across the globe increases, the incidence of age-related cognitive impairment is soaring. Relevant current research focuses on early intervention to slow the

progression of cognitive decline with a long-term goal of helping to find a cure for (reduce the occurrence of) Alzheimer's disease and other dementias (Bateman et al., 2012; Langbaum et al., 2013; Reisa A. Sperling et al., 2011). It has been demonstrated that prevention at prodromal stages targeting disease-modifying risk factors shows promising results and are more likely to be effective (Sindi et al., 2015).

While a full assessment of cognitive function requires a trained clinician, the increasing prevalence of dementia and milder forms of cognitive impairment warrant large-scale screening of the population. Even in high-income countries, as many as 50% of all relevant cases remain undiagnosed (Prince et al., 2016). New approaches to screening and monitoring are needed (Laske et al., 2015; Snyder et al., 2014).

In order to address this problem, we need new tools that are fast, do not need a laboratory, and can automatically indicate which patients might need to be referred to a specialist (Tröger, Linz, Alexandersson, König, & Robert, 2017). Such tools are highly scalable, and can be made accessible to healthcare professionals with little to no specialised training in old age psychiatry. Ideally, it should be possible to administer them remotely, and they should integrate easily with existing telehealth and telecare solutions for older patients. Automated analysis of speech, in particular speech that is produced during a standard clinical assessment, might be a prime candidate for such a tool (Hoffmann et al., 2010; König et al., 2015; López-de-Ipiña et al., 2018; Roark, Mitchell, Hosom, Hollingshead, & Kaye, 2011; Satt, Hoory, König, Aalten, & Robert, 2014; Tóth et al., 2018). Several research groups demonstrated the interest of adopting an automated approach to speech analysis for clinical assessment of older people (Fraser, Meltzer, & Rudzicz, 2015; König et al., 2018; López-de-Ipiña et al., 2013; Meilán, Martinez-Sanchez, Carro, Carcavilla, & Ivanova, 2018; Tóth et al., 2015). Overall, reported work either uses speech from conversations, spontaneous speech tasks, reading or repetition tasks, and fluency tasks.

However, if natural language is analysed, considerable effort has to be spent on pre-processing the data, e.g. annotating turns, or trimming the audio file, in order to prepare it for further computational learning which is not useful for an application in daily clinical practice. Moreover, in order to detect in speech early subtle changes of cognition, it seems crucial to induce a minimum of cognitive effort in a vocal task (König et al., 2018; Szatlóczki, Hoffmann, Vincze, Kalman, & Pakaski, 2015).

Category fluency, or semantic verbal fluency (SVF) task, requires the verbal production of as many different items from a given category, e.g. animals, as possible in a given time period. The SVF task is one of the most widely used neuropsychological test comprising both executive control and semantic memory retrieval processes. It is relatively short and part of standard dementia screens such as the Addenbrookes Cognitive Examination-Revised (ACE-R) (Mioshi et al., 2006) and often used in assessing cognitive function in older people (Canning, Leach, Stuss, Ngo, & Black, 2004; Marczyński & Kertesz, 2006; Peter et al., 2016). SVF performance can distinguish people with dementia from healthy controls (HC) and people with mild cognitive impairment (MCI) (Auriacombe et al., 2006; L. J. Clark et al., 2009; Henry et al., 2004; Pakhomov et al., 2016; Raoux et al., 2008), and additionally can be sensitive to early phases of neurodegenerative disease (Costa et al., 2017).

Most studies of SVF performance use the total number of correct words produced. However, in order to differentiate between multiple pathologies and gain detailed information on underlying cognitive processes, more elaborate measures have been established which serve as additional indicators (Gruenewald & Lockhead, 1980).

One prominent approach, popularized by Troyer and colleagues (A. K. Troyer et al., 1997, 1998), assumes two processes are involved in the production of SVF word sequences, the lexical search for a word from the category to be produced, and the retrieval of other lexical items that are semantically related to the original word. The temporal sequences of semantically related words are called clusters, and the executive search process between clusters is called switching. Typically, semantic clusters are determined using predefined semantic subcategories, often according to Troyer and colleagues (A. K. Troyer et al., 1997). After determining cluster boundaries, the mean size of clusters and the number of switches between clusters are computed. Various parameters related to the size of clusters and number of switches have been shown to be sensitive to cognitive decline and differentiate between different types of dementia.

Unfortunately, any analysis of SVF data that goes beyond word counts is too time consuming for daily clinical practice, especially for general practitioners and family physicians, who are typically the first point of contact for people who suspect that they or one of their loved ones has a cognitive impairment. In addition, any analysis strategy that is based on fixed, pre-defined categories is open to subjective judgement. This might explain some of the

variation in cluster sizes and switch counts reported in the literature (Gomez & White, 2006; Mueller et al., 2015; Murphy et al., 2006; Pakhomov et al., 2016; Raoux et al., 2008; A. K. Troyer et al., 1998).

While automatic analysis introduces its own systematic biases, it is objective, replicable, and yields almost immediate results for clinicians to act on. Thus, computational approaches to analyse the SVF task have been proposed (D. G. Clark et al., 2016; Pakhomov & Hemmy, 2014; Woods et al., 2016) for which statistical methods have been applied in order to obtain semantic clusters. Pakhomov and Hemmy (Pakhomov & Hemmy, 2014), as well as Ledoux and colleagues (2014) use latent semantic analysis, to compute the strength of semantic relations between pairs of words produced (Gabrilovich & Markovitch, 2009). Woods and colleagues (Woods et al., 2016), use explicit semantic analysis (Gabrilovich & Markovitch, 2009) – a vector embedding trained on co-occurrence of words in Wikipedia articles – to identify chaining behaviour for different demographics based on pairwise cosine similarity. Clark and colleagues (D. G. Clark et al., 2016) propose novel semantic measures based on graph theory; most prominently, they put forward graph-based coherence measures which compare the patient's created sequence/path of words with the "shortest" possible path through the fully connected weighted graph of all patient's words. Neural word embeddings based on large word2vec models (Gabrilovich & Markovitch, 2009) allow to directly measure the semantic distance between two given words using simple geometry in the created vector space (Mikolov et al., 2013). In terms of scalability and feasibility for parallel versions of categories, qualitative SVF analysis based on computational semantics may represent a promising step forward. However, before this method could make its way into daily clinical practice, it should be demonstrated to provide reliable and valid data for a regular use.

For this, we set out in this research, to investigate whether fully automatic analysis of the SVF task can be (1) considered as reliable as the manual one, (2) can be used for automatic qualitative assessment of neurocognitive impairment within this task and the corresponding domain, and (3) in the end could be used as a valid fast and scalable screening tool, based on a classification experiment.

#### 4.1.2 *Methods*

##### *Recruitment*

Within the framework of a clinical study carried out for the European research project Dem@care, and the EIT-Digital project ELEMENT, speech recordings were conducted at the Memory Clinic located at the Institut Claude Pompidou and the University Hospital in Nice, France. The Nice Ethics Committee approved the study. Each participant gave informed consent before the assessment. Speech recordings of participants were collected using an automated recording app which was installed on an iPad. The application was provided by researchers from the University of Toronto, Canada, and the company Winterlight Labs.

##### *Clinical Assessment*

Each participant underwent the standardized process used in French Memory clinics. After an initial medical consultation with a geriatrician, neurologist or psychiatrist, a neuropsychological assessment was performed.

Following this, participants were categorized into 3 groups: Control participants (HC) that were diagnosed as cognitively healthy after the clinical consultation, patients with MCI, and patients that were diagnosed as suffering from Alzheimer's disease and related disorders (ADRD). For the ADRD and MCI group, the diagnosis was determined using the ICD-10 classification of mental and behavioral disorders (World Health Organization, 1992). Participants were excluded if they were not native speakers or had any major hearing or language problems, history of head trauma, loss of consciousness, addiction including alcoholism, psychotic or aberrant motor behavior or were prescribed medication influencing psychomotor skills.

The cognitive assessment included (among others) the Mini-Mental State Examination (MMSE) (Folstein et al., 1975), phonemic verbal fluency (letter "f"), SVC (animals), and the Clinical Dementia Rating Scale (O'Bryant et al., 2010).

Each participant performed the SVF task during a regular consultation with one of the Memory Center's clinicians who operated the mobile application. For the Dem@care data, the vocal tasks were recorded with an external microphone attached to the patient's shirt and for the ELEMENT data, with the internal microphone. Instructions for the vocal tasks ("Pouvez-vous me dire le plus possible de noms d'animaux pendant une minute?/Can you

please give me in one minute as many animal names as you can think of?”) were pre-recorded by one of the psychologists of the centre ensuring standardised instruction over both experiments. Administration and recording were controlled by the application and facilitated the assessment procedure.

#### *Speech Data Processing and Transcription*

Recordings of patients were analyzed manually and automatically. For manual analysis, a group of trained speech pathology students transcribed the SVF performances following the CHAT protocol (McWhinney, 1991) and aligned the transcriptions with the speech signal using PRAAT (Boersma & Weenink, Version 6.1.42 ). For the automatic transcription, the speech signal was separated into sound and silent parts using a PRAAT script based on signal intensity. The sound segments were then analysed using Google's Automatic Speech Recognition (ASR) service, which returns several possible transcriptions for each segment together with a confidence score. The list of possible transcriptions was searched for the one with the maximum number of words that were in a predefined list of animals in French. In case of a tie, the transcription with the higher confidence score was chosen.

#### *Features*

Word count was defined as the number of animal names produced minus the number of repetitions. Clusters were determined based on statistical word embeddings, a commonly used technique in computational linguistics, calculated with word2vec (Mikolov et al., 2013) based on the French FraWac corpus (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009) as described by Linz and colleagues (Linz, Tröger, Alexandersson, & König, 2017). Let  $a_1...a_n$  be their representations in the vector space and let  $a_1...a_{n-1}$  form a semantic cluster.  $a_n$  is part of this cluster if

$$\left| \frac{\langle \mu, a_n \rangle}{\|\mu\| \times \|a_n\|} \right| > \delta_p$$

with

$$\mu = \frac{1}{n-1} \times \sum_{x \in \{a_1...a_{n-1}\}} x$$

$$\delta_p = \frac{n!}{(n-2)!} \times \sum_{x,y \in \{a_1...a_n\}} \left| \frac{\langle x, y \rangle}{\|x\| \times \|y\|} \right|$$

Mean cluster size was computed as the average number of words per cluster, and the number of switches was the number of clusters – 1.

### *Classification*

In order to evaluate the feasibility of the automatic approach, we performed 2 analysis that aimed to replicate existing results in the literature on differences in SVC performance between people with no impairment, mild neurocognitive impairment/MCI, and major neurocognitive impairment/AD (Murphy et al., 2006; St-Hilaire et al., 2016). The first used a staging approach using validated normative data provided by St-Hilaire and colleagues (2016), and the second used machine learning (ML) classifiers.

*Automatic Norm-Based Neurocognitive Evaluation* For simulation of a real-world clinical application scenario, word counts from manual and automatic transcripts were compared using normative data for SVF. First, normative equations (Clark et al., 2016) were used to determine a z value, based on manual word counts, age and education level, and people were staged in accordance with diagnostic categories of DSM-5 ( $z > -1$  = no impairment,  $z > -2$  = minor impairment,  $z \leq -2$  = major impairment). In a second step, people were staged using the normative equations, based on automatic word count, age and education level. The first staging was considered the ground truth and the second was compared to the first using classification metrics.

*ML Automatic Diagnosis Classification* To give an idea of how the collected features could be combined to make a diagnostic decision, an ML classifier was trained. Each person in the database was assigned to a label relating to his or her diagnosis (HC, MCI, and ADRD). The features described above (Features section) were used, either calculated from automatic or manual transcripts, depending on the scenario. In all scenarios, we use support vector machines (SVMs) (Cortes & Vapnik, 1995) implemented in the scikit-learn framework (Pedregosa et al., 2012). Leave-one-out cross validation was used for testing. In this procedure, the data are split into 2 subsets (“folds”). One fold contains only one sample, the other contains all other samples. For each of the folds, the classifier is trained on the second fold and evaluated on the held-out sample. To find a well-performing set of hyperparameters, parameter selection using cross-validation on the training set of the inner loop of each cross-validation iteration was performed. All features were normalized using z-standardization, based on the training fold of each iteration.

*Performance Measures*

The performance of ASR systems is usually determined using Word Error Rate (*WER*) as a metric. *WER* is a combination of the mistakes made by ASR systems in the process of recognition. Mistakes are categorized into substitutions, deletions and intrusions. Let *S*, *D*, and *I* be the count of these errors respectively, and *N* be the number of tokens in the ground truth. Then

$$WER = \frac{S + D + I}{N}$$

We only calculated *WER* for words describing animals, not for off-task speech, which also occurs in our data. We refer to this metric as *VFER* (verbal fluency error rate).

As performance measures for prediction of each class in the ML classification experiment, we report the receiver operator curve (*ROC*), as different trade-offs between sensitivity and specificity are visible. We also report area under curve (*AUC*) as an overall performance metric (Bateman et al., 2012).

*Table 7: Demographic data and clinical scores by diagnostic group. Data are presented as mean (standard deviation) or as stated. HC, healthy control; MCI, mild cognitive impairment; ADRD, Alzheimer's disease and related disorders. \*  $p < 0.05$ , significant difference from the control population (Wilcoxon-Mann-Whitney test).*

	HC	MCI	ADRD
Subjects, n	24	47	24
Age, years	76.12 (4.41)	76.59 (7.6)	77.7 (3.99)
Sex	5M/19F	23M/24F	8M/16F
Education, years	10.50 (4.05)	10.81 (3.6)	9.75 (4.69)
MMSE	28.21 (1.82)	26.02* (2.5)	18.83* (4.99)
CDR-SOB	0.46 (0.67)	1.68* (1.11)	7.5* (3.7)

#### 4.1.3 Results

Relevant demographic characteristics of the HC group ( $n = 24$ , age 76.12 years, MMSE 28.21, CDR-SOB 0.46), the MCI group ( $n = 47$ , age 76.59 years, MMSE 26.02, CDR-SOB 1.68), and the ADRD group ( $n = 24$ , age 77.7 years, MMSE 18.83, CDR-SOB 7.5) are presented in Table 7. The total number of participants was 95. Excluding MMSE and CDR-SOB, no significant effects between the groups were found.

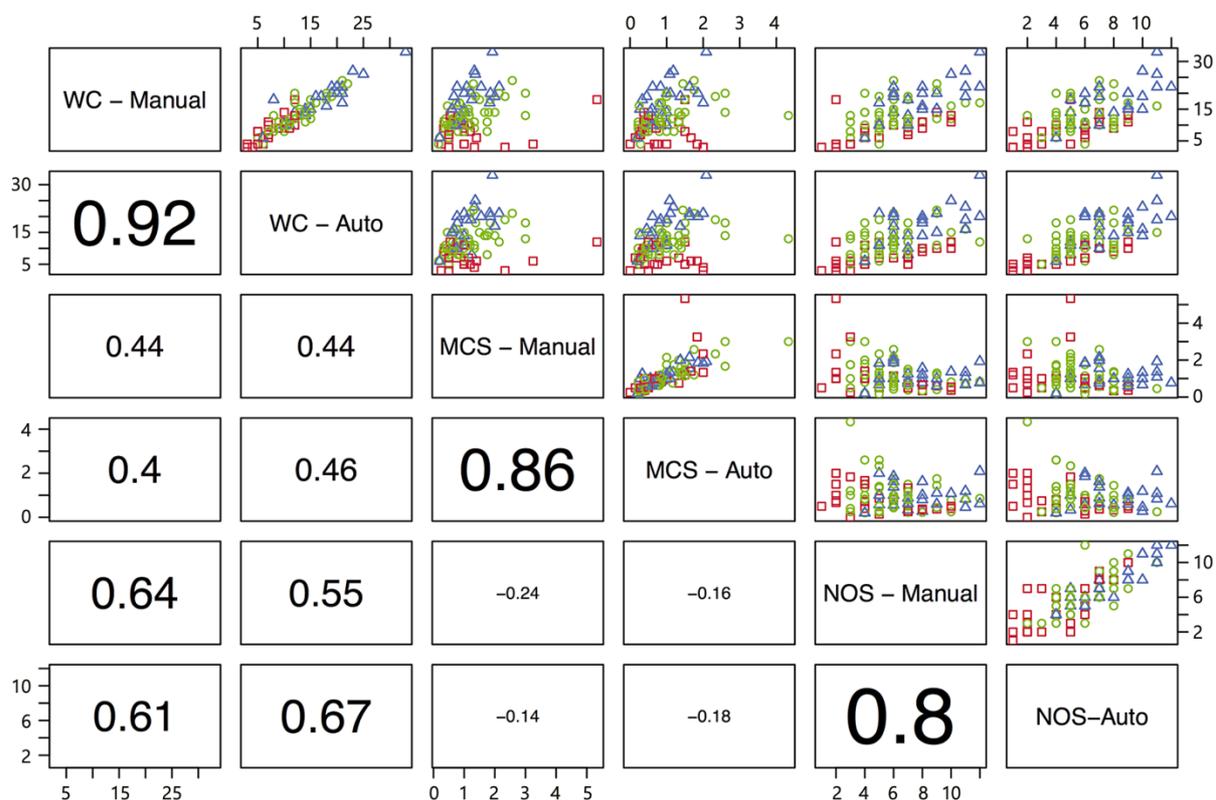


Figure 18 Correlation matrix and scatter plots for features based on manual and automatic transcripts. Spearman's correlation coefficients are reported. WC, word count; MCS, mean cluster size; NOS, number of switches. Diagnostic groups are encoded on the scatter plot as healthy controls = blue triangles, MCI = green circles, AD = red squares.

#### Automated Speech Recognition

Evaluation of all samples in the corpus yielded a VFER of 20.01%. Since not all types of errors might have the same impact on analysis (e.g., word count is not influenced by substitutions in all cases), the proportion of types of error made are considered. 50.3% of all errors were deletions, 29.8% were substitutions, and 19.9% were insertions.

#### Correlation

The relationship between features extracted from automated transcripts and manual ones was examined.

Consequently, Spearman's correlation coefficient was computed. All relationships are reported in Figure 18. The correlation between manual and automatic SVF analysis was strong across all 3 relevant features with a correlation of  $\rho = 0.921$  for the main clinical feature in this task, the word count.

*Automatic Norm-Based Neurocognitive Evaluation*

Neurocognitive disorder evaluations (no impairment, minor and major impairment) determined with the automatic word count agree with labels based on the manual word count with an accuracy of 0.831, weighted precision of 0.83, weighted recall of 0.83 and *F1* of 0.83. When looking at sensitivity and specificity in a one versus all scenarios, using HC as the negative class, the model achieves a sensitivity of 0.914 and a specificity of 0.833. A detailed confusion matrix is depicted in Figure 19.

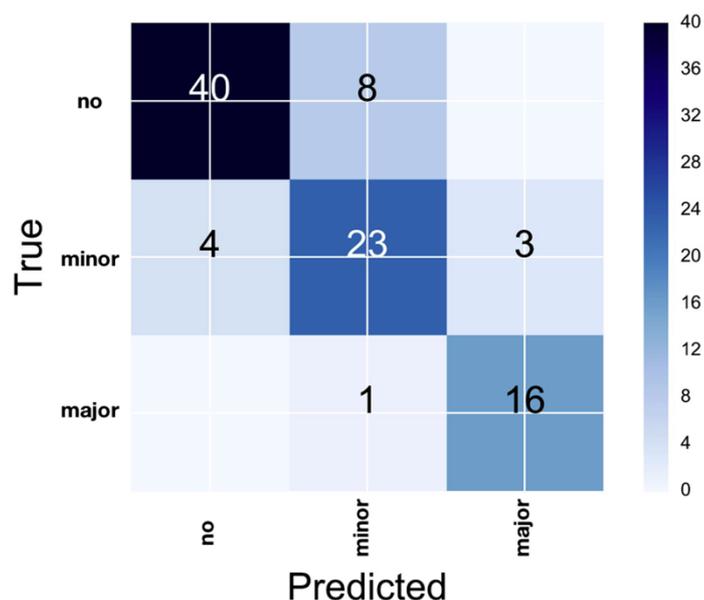


Figure 19: Confusion matrix for diagnosis based on normative data, automatic word count (WC), and manual WC (*no* =  $z > -1$ , no impairment; *minor* =  $z > -2$ , minor impairment; *major* =  $z \leq -2$ , major impairment).

*ML Automatic Diagnosis Classification*

ROC curves for all scenarios are reported in Figure 20. Classifiers trained on automatic measures and manual ones perform comparably or better for 2 of 3 scenarios.

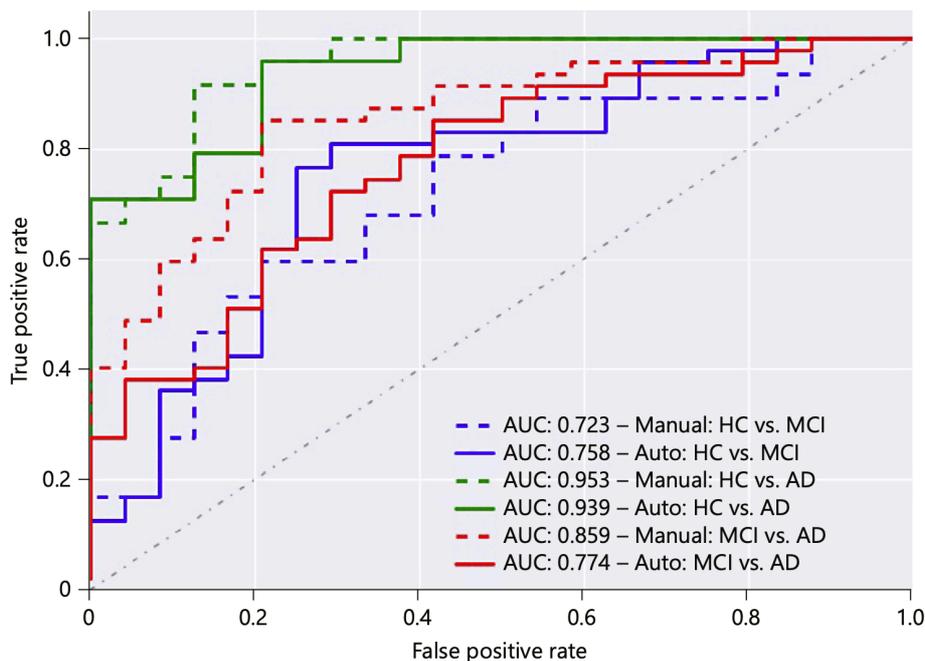


Figure 20: Receiver operator curve of classification models for different scenarios. Models trained on manually extracted features are displayed as dashed lines, ones based on automatic features are displayed as solid lines. Colour indicates the classification scenario, as coded in the legend. Area under the curve (AUC) reported in the legend for each scenario and feature set.

#### 4.1.4 Discussion

In this paper, we describe an automated analysis method for the fine-grained analysis of SVF data in terms of clusters and switches and validate it for the category of animals. Clusters and switches, determined by the tool correlate well with clusters and switches that were determined manually using a strict annotation procedure. Both manually and automatically derived statistics were successful in distinguishing between HC, people with MCI and people with Alzheimer's disease or related disorders.

Considering the reliability of the fully automated pipeline, the ASR is often considered to be the main limiting factor (Tóth et al., 2015). Our results show an overall low error rate of 20.01% for the automated system, compared to the manual transcripts. This in itself represents an improvement over results of other authors using ASR systems for evaluating the SVF tasks (Lehr, Prud'hommeaux, Shafran, & Roark, 2012; Pakhomov, Marino, Banks, & Bernick, 2015). In line with previous research, diagnostic groups differ significantly in the number of errors made by ASR (Kruskal-Wallis,  $\chi^2 = 13.7$ ,  $df = 2$ ,  $p < 0.001$ ). More word errors are produced by the ASR for AD patients, compared to healthy subjects. Since persons with AD are expected to produce less words in an SVF task, this does not negatively affect further

analysis. Closely looking at the types of errors, insertions and deletions are both problematic for further analysis. Both skew the raw word count, which still is the single most predictive performance indicator in SVF for dementia detection. Substitutions only affect qualitative measures such as the mean size of clusters and the number of switches between clusters, but do not affect the word count.

#### *Automatic Norm-Based Neurocognitive Evaluation*

Even though the ASR produced word errors, mainly deletions, which negatively affect the overall word count and thereby the main clinical measure of SVF, the correlation between the automated and manual systems is very strong, i.e. 0.921. This shows that although the ASR system introduces some errors, it does not greatly affect the overall clinical measure, since the errors are not correlated to cognitive status. In the first experiment, we benchmarked the automatic pipeline for a norm-based neurocognitive evaluation. The performed neurocognitive evaluation based on automatic word count agreed strongly with labels based on the manual word count. The confusion matrix (Figure 19) shows that the automatic approach tends to systematically underestimate the performance of a person in the SVF task. This can be attributed to the deletions of the ASR. Thus, the automatic pipeline can be considered conservative, showing high sensitivity, which is of great importance to its use as a screening tool.

#### *Automated ML Diagnosis Classification*

For both the HC versus AD and HC versus MCI scenarios, the performances of models trained on automatic and manual features have comparable AUC (0.723 vs. 0.758 and 0.953 vs. 0.939). In the MCI versus AD scenario, the AUC of models trained on automated features deteriorated (0.859 vs. 0.774). The difference of the previous experiment can be explained by the flexibility of ML models to learn decision boundaries, in contrast to pre-determined diagnostic norms. ML models are also able to accommodate the previously mentioned systematic errors of ASR.

A similar approach has been suggested by (D. G. Clark et al., 2016), studying the utility of an automatic SVF score for the prediction of conversion with the result that higher prediction accuracy was obtained with the classifiers trained on all scores, rather than on manual scores. Overall, it can be stated that using automatic analysis of the SVF task allows immediate access to reliable and clinically relevant measures such as the word count,

switches, and clusters. This is potentially useful for differentiating between deficits in either executive or semantic processing. The automation of recording, transcription, and analysis streamlines test administration and ultimately leads to more robust, reproducible data.

In addition to the assessment of cognitive decline, these qualitative measures extracted from the SVF performances may be of great interest as well for other neurocognitive disorders affecting verbal ability and executive control such as frontotemporal dementia or primary progressive aphasia (Van Den Berg, Jiskoot, Grosveld, Van Swieten, & Papma, 2017).

Costa and colleagues (Costa et al., 2017) state that we are far from having available reliable tools for the assessment of dementias, since one of the main problems is the heterogeneity of the tools used across different countries. Therefore, a working group of experts recently published recommendations for the harmonization of neuropsychological assessment of neurodegenerative dementias with the aim to achieve more reliable data on the cognitive-behavioral examination. Automated speech analysis of the SVF could be one potential tool to assist in harmonizing test procedures and outcomes. It also provides additional quantitative measurements extracted from speech signals for cognitive screening without increasing time, costs or even workload for the clinician. Such a tool could be used as an endpoint measurement in clinical trials to assess intervention outcome and monitor disease progress, even remotely over the phone.

#### *Limitations*

A few limitations of this study should be considered. We did not recruit healthy participants from the general elderly population, but were limited to include persons who came for clinical consultation to the memory clinic cognitively healthy but with some subjective complaints. It should be further noted that the data set for this study is only in French, thus, limiting transferability of its results to other languages. A major goal for future work is the collection of SVF recordings in multiple languages and within the framework of longitudinal studies.

#### *4.1.5 Conclusion*

To conclude, the study demonstrates the feasibility of automatic analysis of SVF performance in elderly people to assess and monitor cognitive impairment. Furthermore, new measures beyond simple word counts such as word frequencies could be investigated in the

future, possibly giving way to a deeper understanding of underlying cognitive functions and changes due to neurodegenerative disease.

**Acknowledgements** This research was partially funded by the EIT Digital Wellbeing Activity 17074, ELEMENT. The data were collected during the EU FP7 Dem@Care project, grant agreement 288199.

4.2 TELEPHONE-BASED DEMENTIA SCREENING I: AUTOMATED SEMANTIC VERBAL FLUENCY ASSESSMENT

Johannes Tröger<sup>1</sup>, Nicklas Linz<sup>1</sup>, Alexandra König<sup>2</sup>, Phillipe Robert<sup>2</sup>, Jan Alexandersson<sup>1</sup>

<sup>1</sup> German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

<sup>2</sup> Memory Clinic, Association IA, CoBTek Lab, CHU Université Côte d'Azur, Nice, France

Dementia has a large economic impact on our society as cost-effective population-wide screening for early signs of dementia is still an unsolved medical supply resource problem. A solution should be fast, require a minimum of external material, and automatically indicate potential persons at risk of cognitive decline. Despite encouraging results, leveraging pervasive sensing technologies for automatic dementia screening, there are still two main issues: significant hardware costs or installation efforts and the challenge of effective pattern recognition. Conversely, automatic speech recognition (ASR) and speech analysis have reached sufficient maturity and allow for low-tech remote telephone-based screening scenarios. Therefore, we examine the technologic feasibility of automatically assessing a neuropsychological test—Semantic Verbal Fluency (SVF)—via a telephone-based solution. We investigate its suitability for inclusion into an automated dementia frontline screening and global risk assessment, based on concise telephone-sampled speech, ASR and machine learning classification. Results are encouraging showing an area under the curve (AUC) of 0.85. We observe a relatively low word error rate of 33% despite phone-quality speech samples and a mean age of 77 years of the participants. The automated classification pipeline performs equally well compared to the classifier trained on manual transcriptions of the same speech data. Our results indicate SVF as a prime candidate for inclusion into an automated telephone-screening system.

*Keywords:* Dementia, Screening, Speech Analysis, Phone-based, Machine Learning

The previous paper helped to understand the feasibility of automatic qualitative analysis of SVF performance in elderly people to screen and monitor for neurocognitive impairment. After manual and automatic annotation, qualitative semantic measures (clusters and switches, as introduced in the previous chapter) have been extracted and the overall performance has been evaluated in a machine learning experiment (HC vs. MCI vs. ADRD) for both automatic as well as manual annotations. Automatic annotation yielded comparable qualitative features and subsequently performed comparably on the overall screening task.

With this result, the previous paper represents an essential step towards screening for early (cognitive) signs of AD dementia early and at scale. The results showed that automatic qualitative analyses of SVF data can be used for screening. However, it primarily helped to prove the cost-effectiveness of the approach (reducing the manual resource-demanding pre-processing steps of the speech input) but did not suggest an overall technical solution that is also scalable (requiring little to no human resources needed for performing as well as evaluating the assessment and also requiring minimum technical setup). The following paper will introduce a scalable technical concept and provide evidence of its feasibility in order to arrive at a truly scalable approach to population-wide early screening for AD dementia signs.

#### *4.2.1 Introduction*

Dementia has a large economic impact on our society: according to the World Alzheimer Report 2016, dementia is about to become a trillion-dollar disease by 2018 (Prince, Comas-Herrera, Knapp, Guerchet, & Karagiannidou, 2016). Since many clinical trials have failed to find a cure, a conceptual shift has occurred considering Alzheimer's disease (AD) as a continuum for which early intervention may offer the best chance of therapeutic success (Dubois et al., 2016). This urgent need to identify a treatment that can delay or prevent AD has increased the number of preventional trials targeting disease modifying risk factors for which early screening of subjects at risk to develop cognitive impairment is highly relevant (Aisen et al., 2017). Recent research has shown that prevention at prodromal stages targeting disease mechanisms show promising results and are more likely to be effective (Sindi et al., 2015). Many challenges remain detecting these 'silent' stages, where clinical signs are not yet very obvious since our understanding of the pathological mechanism is still quite limited (Auriacombe et al., 2006; Gomez & White, 2006; Pakhomov et al., 2016; Raoux et al., 2008) and current tools may lack sufficient sensitivity to detect subtle but meaningful changes.

This approach has led to the current discussion on creating and approving more clinically relevant measures for early population-based screening with low-cost tests of high sensitivity and lower specificity (Dubois et al., 2016). For instance, currently, just 50% of cases are diagnosed in Europe and the US (Auriacombe et al., 2006; Gomez & White, 2006; Pakhomov et al., 2016; Raoux et al., 2008). This can be attributed to effective screenings for early signs of dementia (mild neurocognitive disorder) having not reached sufficient coverage. Especially in areas with low population density, clinical facilities and experts are too distributed to screen populations effectively, as this is still done in a face-to face manner today. Many clinical trials suffer from high dropout rates partly due to visit frequency and study length (Grill & Karlawish, 2010). This translates into a medical supply resource problem and highlights the opportunities for telemedicine applications.

It has been put forward that new tools may address this need fast, require neither laboratory setup nor external material, and automatically evaluate and indicate potential clinically relevant persons. Therefore, research should focus on innovative computerized tools that reveal robust psychometric properties for early detection of neurocognitive disorder significantly decreasing the workload of expert clinicians, which represent a very rare resource in most cases. Thus, automatic, inexpensive and remote solutions allowing a broad frontline screening of cognitive abilities in the general population should be developed.

There is growing evidence for the feasibility of automatic speech analysis in addressing exactly this need (Hoffmann et al., 2010; Lehr et al., 2012; Tóth et al., 2015). Speech-based solutions can be remotely administered via telephone and therefore have minimal technical user interface requirements. This makes them a very attractive solution in the mentioned frontline screening context.

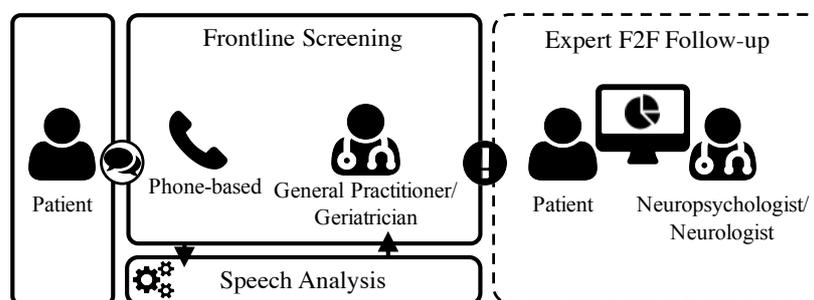
Neuropsychological studies comparing a video and telephone based psychometric dementia screening with a face-to-face assessment, reported good ecological validity for the telemedicine application (Munro Cullum, Hynan, Grosch, Parikh, & Weiner, 2014). However, such studies do not fully exploit the combined opportunities of telemedicine neuropsychological screening empowered by automatic speech analysis and machine learning classification.

Our aim is to develop technology with which raw speech data can be processed via the telephone—facilitated by computational linguistic techniques and machine learning—in order

to give a simple risk assessment for dementia. Instead of using free, unconstrained speech, we hope to achieve better performance and shorter assessment times, through analyzing performances of cognitive tests.

Semantic Verbal Fluency (SVF) tasks are neuropsychological tests in which patients are given limited time (e.g. 60 seconds) to name as many items belonging to a certain semantic category as they can. SVF has been shown to be sensitive to even early forms of dementia (Auriacombe et al., 2006; Gomez & White, 2006; Pakhomov et al., 2016; Raoux et al., 2008). SVF can be considered a multifactorial task, comprising both semantic memory retrieval and executive control processes (Henry & Crawford, 2005; Robert et al., 1998; A. K. Troyer et al., 1997). Previous studies have concluded the feasibility of automatically analyzing SVF performances (König et al., 2018; Pakhomov et al., 2015), although no study known to the authors has investigated analysis of telephone quality recordings.

The aim of this study is therefore to benchmark a solution processing raw telephone quality SVF data suitable for inclusion in a fully automated dementia frontline screening for global risk assessment (compare also Figure 21).



*Figure 21: Telephone-based frontline screening scenario: speech gets sent to the analysis server which automatically indicates the general practitioner (GP) or the geriatrician (G) in charge the risk for neurocognitive disorder, the GP/G also checks via phone for excluding/confounding conditions (e.g. substance abuse) and forwards the patient to the specialist who would efficiently continue with the in-depth assessment.*

#### 4.2.2 Related Work

The following section gives an overview of efforts aiming at the automated detection of dementia based on multiple different sensor solutions. For this paper, we would like to differentiate between solutions based on classic *pervasive sensing* such as home monitoring systems and speech analysis as a special subcategory of pervasive sensing.

*Computerized cognitive screening*

Digital tests that seek to assess cognitive functions, briefly and globally, are being developed with the aim to be administered remotely (Brando, Olmedo, & Solares Canal, 2017). The exhibited advantages of these tests are standardization of administration and stimulus presentation as well as the measures (e.g. reaction times and latencies) are more accurate: performances can be compared to established norms (Wild, Howieson, Webbe, Seelye, & Kaye, 2008) allowing the clinician to concentrate on a personalized analysis of the patients' needs.

For instance, the CogState Brief Battery (CogState) is a brief computerized test which assesses reaction and processing speed, episodic memory, attention, working memory, learning, and decision-making. (de Jager, Schrijnemaekers, Honey, & Budge, 2009) examined the specificity and sensitivity of the CogState test for the diagnosis of mild cognitive deterioration, comparing it with classical pen and paper tests with the result that it reaches similar discrimination level as traditional tests.

CANTAB, one of most known cognitive screening tools, offers specialized AD test battery versions for assessing prodromal states, or mild dementia. The batteries measure motor skills, executive function, episodic memory, visual memory information processing and sustained attention. CANTAB has been shown to be highly sensitive to cognitive dysfunction and ties in closely with current neurobiological models for MCI (Égerházi, Berecz, Bartók, & Degrell, 2007; Fowler, Saling, Conway, Semple, & Louis, 1997).

The TDAS (Touch Panel-type Dementia Assessment Scale) (Inoue, Jimbo, Taniguchi, & Urakami, 2011) based originally on the pen and paper ADAS-cog test (Rosen, Mohs, & Davis, 1984), measures word recognition, instruction compliance, temporal orientation, visuospatial skills, recognition of object use, naming, planning of the writing process, money computation, and recognition of the time indicated by an analogue clock. This digital test can be administered in 30 minutes, just two-thirds of the time that ADAS-cog requires.

The CNSVS (CNS Vital Signs) (Gualtieri & Johnson, 2006) is a digital screening test, assessing working memory, mental flexibility, psychomotor speed, verbal and visual memory, set shifting and inhibition and vigilance and sustained attention. The authors studied test-retest reliability as well as concurrent and discriminant validity concluding that it can be used as a reliable screening tool in medical contexts.

Phone-based screening has been investigated by Castanho and colleagues (2016) comparing the delayed recall task and a classical neuropsychological assessment with the Telephone Interview of Cognitive Status (TICS) in a population of older adults. The TICS consists of 13 items evaluating spatial, temporal and personal orientation, working memory, attention, and verbal and semantic memory. TICS showed high correlation levels with global scores of classical tests as well as a satisfactory internal consistency. This method could allow faster access to assessment for people living in rural areas producing similar results as the usual pencil and paper screening tests.

*Automated Screening Based on Pervasive Sensing*

Manifold research has been done into the feasibility of home monitoring systems for modelling domestic circadian activities (activity patterns following a biological 24h rhythm). As such rhythms are typically disturbed by dementia—especially nocturnal activity patterns—these techniques provide a useful basis for automatic dementia detection/screening. Using infrared sensors to monitor nocturnal activities, studies have found significant differences between dementia patients and healthy controls (e.g. Suzuki, Murase, Tanaka, & Okazawa, 2007). Similarly, the same technical setup has been shown to effectively model daily routines (Franco, Demongeot, Villemazet, & Vuillerme, 2010). Following the same rationale and technique König and colleagues (König et al., 2015) leveraged automatic detection of instrumental activities of daily living (IADL) in patients with MCI and healthy participants. Besides promising results, such studies are often carried out with very small sample sizes ( $N < 50$ ) and focus mainly on the automatic classification of activities rather than the actual neurocognitive disorder. Moreover, the installation of home-monitoring systems requires significant resources and a person's consent to be monitored in their private life; two issues that render such a solution unrealistic in broad population frontline screening.

Also focusing on circadian rhythm monitoring but using less complex wrist-worn technology, Paavilainen and colleagues (2003) found significant correlations between sleep patterns and common dementia staging scales. However, similar to the above-mentioned studies, sample size is relatively small and the main automatic analysis effort was spent on activity monitoring rather than prognostic classification problems.

Beyond such passive sensing approaches, there is also research on the diagnostic use of pro-active sensing situations: situations that are framed by some task/instruction producing

more diagnosis related variance. Leveraging virtual reality technology, (Tarnanas et al., 2013) used a realistic virtual reality (VR) fire evacuation task to predict amnesic Mild Cognitive Impairment (MCI; often considered as the precursor of dementia), Alzheimer's disease (AD) and controls from task performance reaching area under the curve (AUC) values of more than 80%. Though very sensitive, the classification setup requires a lot intervention from technicians to analyze the VR task performance. Moreover, the VR screening setup has similar limitations as the classic neurological assessment: it requires the expensive VR laboratory and test persons have to leave their home.

Other studies combine gait and balance analysis through a hip- /foot-worn accelerometer and specific walking tasks (Chung et al., 2012; Hsu et al., 2014). Such approaches take advantage of classic geriatric assessments showing age-/dementia-related gait irregularities when confronted with a simple straight-line walking task or dual task paradigms (e.g. walking and mental arithmetic task).

These pervasive sensing approaches reveal several shortcomings for our use case. They are either very technology-heavy, which implies significant investments, and rely heavily on activity recognition which represents an ongoing classification research challenge in itself. Alternatively, they have to be done in laboratories far away from peoples' homes. Conversely, automatic speech analysis recently has reached a technical readiness level that renders it very attractive for speech based pervasive solutions. Moreover, the only technical requirement is a working telephone which can be considered as ubiquitous in most countries even for an aged population such as the dementia screening target group.

#### *Automated Screening Based on Speech*

Authors have reported studies on automated dementia screening with possible applications in phone-based telemedicine scenarios. Tröger and colleagues (2017) extracted paralinguistic features from speech based cognitive tests and trained classifiers to discriminate between healthy controls and patients with AD. Furthermore, Lehr and colleagues (2012) used ASR to extract features from a story retelling task and was able to discriminate between MCI and healthy controls with an Area Under the Curve (AUC) score of 80.9%. Satt and colleagues (2014) used four spoken cognitive tests (Countdown, Picture description, Repetition and SVF), extracted paralinguistic features to discriminate individuals with MCI, early AD and healthy controls (HC). Trained models achieve an accuracy of 87% for

early AD vs. HC and 81% for MCI vs. HC. Not focusing on dementia detection but on Parkinson's Disease, Klumpp and colleagues (Klumpp et al., 2017) report an application which is phone-based and acts as a passive listener to monitor speech over time. However, as soon as an anomaly is detected the app also uses classic cognitive speech tasks to elicit richer and more controlled variance (i.e. a psychomotor task: continuously repeating pa-ta-ka during a given period of time)

Multiple studies report approaches that are less feasible in phone-based screening scenarios but provide strong evidence for the effectiveness of speech-based screening for dementia patients, including early stages. Overall, reported work either uses speech from conversations, spontaneous speech tasks, reading or repetition tasks, and fluency tasks.

The most liberal setting consists of conversations with clinicians. Audio files of spontaneous speech from conversations (Dodge et al., 2015; Khodabakhsh, Yesil, Guner, & Demiroglu, 2015), or classical autobiographic patient interviews (Hoffmann et al., 2010) have been used in small setups, yielding significant effects. For such data, considerable effort has to be spent on preprocessing the data (e.g. annotating turns or trimming the audio file) in order to prepare it for further computational learning.

Tasks, eliciting spontaneous speech, are slightly more restricted and therefore easier to process; descriptions of the Cookie Theft Picture or comparable visual material, allows for extracting a wide variety of features and yields very good results (Al-hameed, Benaissa, & Christensen, 2016; Fraser, Rudzicz, & Hirst, 2016; König et al., 2015; Orimaye, Wong, & Golden, 2014). Similarly, some researchers report positive results from speech samples based on an animated film free recall task (Gosztolya et al., 2016).

Reading or repetition tasks are the handiest to deal with, in the sense of automated processing, as they need little transcription and provide an inherent ground truth. Simple sentence reading has been shown to provide enough variance to effectively discriminate between AD and HC with an accuracy of 84% (Meilán et al., 2014).

Verbal fluency tasks, such as the semantic animal fluency task, have produced rich variance to discriminate between AD patients and HC (Lehr et al., 2012; Linz, Tröger, Alexandersson, & König, 2017; B. Yu, Quatieri, Williamson, & Mundt, 2015). The benefits of semantic vs. phonemic fluency tasks have been discussed in multiple publications and there is a large body of neuropsychological evidence reporting dementia patients' difficulties in

semantic fluency tasks, concluding that dementia patients and MCI patients have significant more difficulties in semantic, e.g., animal, fluency tasks compared to other psychometric standard tests.

In summary, speech analysis provides a powerful opportunity to broad dementia screening as it has minimal technical requirements and leverages a mature technology—ASR—and can be done remotely in almost all geographic areas. Sensitivity can even be increased through the use of specific psychometric speech tasks, such as the semantic verbal fluency task. Therefore, our aim is to benchmark an entirely automatic pipeline for dementia screening using telephone-quality audio recordings of a classic dementia screening speech task, ASR and machine learning classifiers on top.

#### *4.2.3 Methods*

In order to address the above-mentioned challenges, this section will elaborate on the technical pipeline of the proposed system and provide evidence for its feasibility. In the following, the telephone-based speech data processing and the machine learning experiment will be described.

##### *Participants*

Within the framework of a clinical study carried out for the European research project Dem@care, and the EIT Digital project ELEMENT, speech recordings were conducted at the Memory Clinic located at the Institut Claude Pompidou and the University hospital in Nice, France. The Nice Ethics Committee approved the study. Each participant gave informed consent before the assessment. Speech recordings of elderly people were collected using an automated recording app which was installed on a tablet computer. Participants underwent a clinical assessment including a battery of recorded speech-based tasks.

Each participant went through an assessment including: Mini-Mental State Examination (MMSE) (Folstein et al., 1975), the phonemic and semantic verbal fluency (Tombaugh, Kozak, & Rees, 1999), and the Clinical Dementia Rating Scale (O'Bryant et al., 2008). Following the clinical assessment, participants were categorized into three groups: control participants that complained about having subjective cognitive impairment (SMC) but were diagnosed as cognitively healthy after the clinical consultation, patients with MCI and patients that were diagnosed with dementia (D), including AD. For the AD group, the diagnosis

was determined using the NINCDS-ADRDA criteria (McKhann et al., 2011). Related mixed/vascular dementia was diagnosed according to the ICD 10 (World Health Organization, 1992). For the MCI group, diagnosis was conducted according to Petersen criteria (Petersen et al., 1999). Participants were excluded if they had any major audition or language problems, history of head trauma, loss of consciousness, psychotic or aberrant motor behavior. For an overview of demographic data see Table 8.

Each participant performed the SVF task during a regular consultation with one of the Memory Center's clinician who operated the mobile application which was installed on an iPad tablet. Instructions for the vocal tasks were pre-recorded by one of the psychologists of the center ensuring a standardized instruction over the experiment. Administration and recording were controlled by the application and facilitated the assessment procedure.

*Table 8: Demographic data and clinical scores by diagnostic group; mean (standard deviation); SMC='Subjective Memory Complaints', MCI='Mild Cognitive Impairment', D='Dementia', MMSE='Mini Mental State Examination', CDR-SOB='Clinical Dementia Scale - Sum of Boxes'.*

	SMC	MCI	D
N	40	47	79
Age	72.65 (8.3)	76.59 (7.6)	79.0 (6.1)
Sex	8M/32F	23M/24F	39M/40F
Education in years	11.35 (3.7)	10.81 (3.6)	9.47 (4.5)
MMSE	28.27 (1.6)	26.02 (2.5)	18.81 (4.8)
CDR-SOB	0.47 (0.7)	1.68 (1.11)	7.5 (3.7)

#### *Speech Data Processing*

Speech was recorded through a mobile tablet device using the built-in microphone. The recordings were digitized at 22050 Hz sampling rate and at 16 bits per sample. To simulate telephone conditions, the recordings were downsampled to an 8000 Hz sampling rate, using the Audacity<sup>4</sup> software. Since the tablet device's microphone is used in mobile phones, no further transformations were applied.

Recordings of patients were analysed manually and automatically. For manual analysis, a group of trained speech pathology students transcribed the SVF performances following the CHAT protocol (McWhinney, 1991) and aligned the transcriptions with the speech signal using PRAAT (Boersma & Weenink, Version 6.1.42 ). For the automatic

<sup>4</sup> <http://www.audacityteam.org/>

transcription, the speech signal was separated into sound and silent parts using a PRAAT script based on signal intensity. The sound segments were then analysed using Google's Automatic Speech Recognition (ASR) service<sup>5</sup>, which returns several possible transcriptions for each segment together with a confidence score. The list of possible transcriptions was searched for the one with the maximum number of words that were in a predefined list of animals in French. In case of a tie, the transcription with the higher confidence score was chosen.

### *Features*

We extracted a variety of features from the generated transcripts. All hereunder reported features are either clinically accepted (i.e. word count), have been proven to have diagnostic power based on previous medical research (i.e. clusters and switches) or proved to have diagnostic power based on research in the field of computational linguistics (i.e. semantic metrics). Moreover, all features are firmly based on clinical research and therefore explicable and understandable by medical experts.

**Word Count:** The count of distinct correct responses (animals), excluding repetitions, is the standard clinical measure for evaluation of SVF. Its diagnostic power for even early stages of cognitive impairment has been shown in countless studies.

**Clusters and Switches:** Many previous researchers (Gruenewald & Lockhead, 1980; Linz, Tröger, Alexandersson, & König, 2017; Raoux et al., 2008; A. K. Troyer et al., 1997) have shown that production in SVF is guided by so called clusters—clusters of words that are produced in rapid succession and often shown to be semantically connected. We determine clusters in multiple ways—taxonomy-based (A. K. Troyer et al., 1997) and statistical (Linz, Tröger, Alexandersson, & König, 2017) semantic, as well as temporal analysis (Fernaes, Östberg, Hellström, & Wahlund, 2008)—and compute mean cluster size and number of switches between clusters as features.

**Semantic Metrics:** Many purely semantic metrics have been suggested for the analysis of SVF, that look at the type of words produced. We include frequency norms (Linz, Tröger, Alexandersson, Wolters, et al., 2017) estimated from large text corpora and computed as the mean frequency of any produced word and semantic distance (Linz, Tröger, Alexandersson,

---

<sup>5</sup> <https://cloud.google.com/speech/>

Wolters, et al., 2017) approximated using neural word embeddings trained on external text resources. We include the mean semantic distance between any produced word, the overall mean of means of semantic distances inside a temporal cluster and the mean semantic distance between any temporal cluster.

### *Classification Experiment*

In order to evaluate the feasibility of using SVF in a telephone screening scenario, we performed a machine learning experiment. We built classifiers that discriminate the healthy population from the impaired samples. People were counted into the impaired population, when they belonged to either the MCI or dementia groups. First, we established a performance baseline, training models based on features extracted from manual transcripts. After that we used the transcripts from ASR to extract features and constructed models.

In all scenarios we used Support Vector Machines (SVMs, Cortes & Vapnik, 1995) implemented in the scikit-learn framework (Pedregosa et al., 2012). Due to our limited amount of data—166 samples—we could not keep a separate hold-out set for testing and instead used leave-one-out cross validation. For each sample, the data is split into a training-set—all samples but the one—and a test-set—the one held-out sample. The classifier is trained on the test set and evaluated on the held-out training set. To find a well-performing set of hyperparameters for the SVM (i.e., kernel,  $C$ ,  $\gamma$ ), we performed parameter selection using cross-validation on the training set of the inner loop of each cross validation iteration. For an overview of the complete pipeline spanning from speech recording to automatic screening classification see Figure 22.

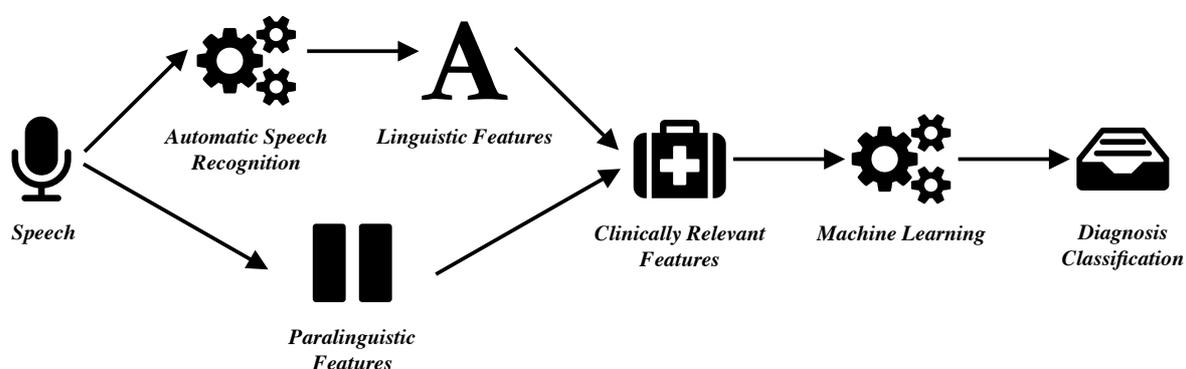


Figure 22: Technical pipeline: the automatic frontline screening using machine classification and feature selection of clinically relevant features feeding the machine learning classifier for neurocognitive screening.

*Performance Measures*

The performance of ASR systems is usually determined using Word Error Rate (WER) as a metric. WER is a combination of the types of mistakes made by ASR systems in the process of recognition. Mistakes are categorized into substitutions, deletions and intrusions. Let  $S$ ,  $D$  and  $I$  be the count of these errors and  $N$  the number of tokens in the ground truth. Then

$$WER = \frac{S + D + I}{N}$$

Since WER considers all utterances, including off-task speech which is not reflected in any of our features, we used a slightly adapted version. Instead of comparing the ground truth annotation of the recording and the ASR results, we transformed both into a list of animals and calculate the WER for these sequences. We refer to the result as the Verbal Fluency Error Rate (VFER) in further discussion.

As performance measures for prediction of each class in the ML classification experiment, we report the receiver operator curve (ROC), as different tradeoffs between sensitivity and specificity are visible. We also report area under curve (AUC) as an overall performance metric.

*4.2.4 Results*

We first evaluate the VFER on the automatic transcript, which is determined to be 33.4%. Of the errors made by the ASR, 69% are deletions, 22% are substitutions and 9% are intrusions. Substitutions are the least problematic error, since they only skew the word count—the single most predictive feature—in rare cases, where a word is substituted with a previously named one.

Figure 23 shows the receiver operator curve (ROC)—a plot of true positive rate vs. false positive rate—for both classification experiments. Models based on features extracted from manual transcripts have an AUC of 0.852 and models built on features extracted from automatic transcripts show an AUC of 0.855. Since a high sensitivity is key for screening applications, a sensible sensitivity-specificity trade-off for the automatic model could be at a sensitivity of around 0.85 and a specificity of 0.65.

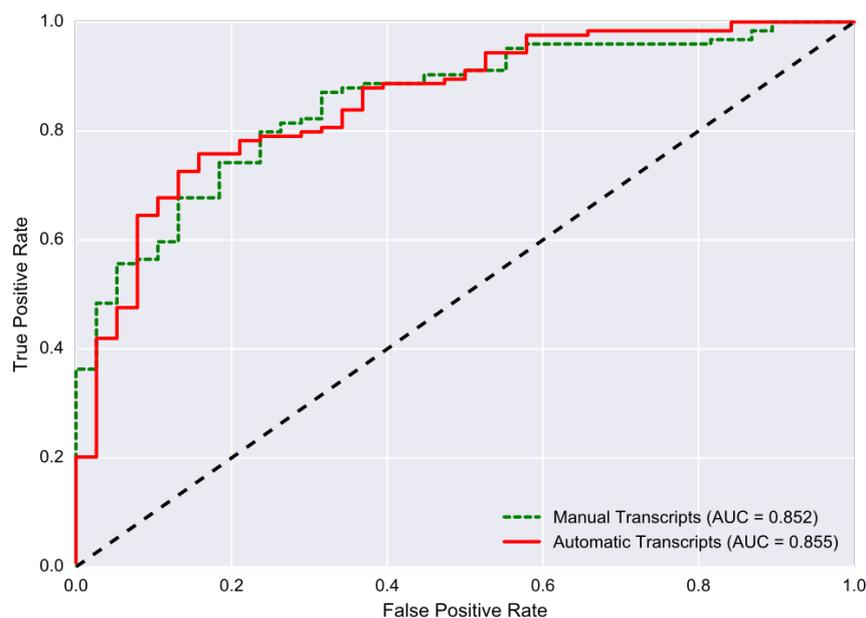


Figure 23: Receiver Operator Curve (ROC) for features based on manual transcripts (green) and on automatic transcripts (red). Area under curve (AUC) is reported in the key.

#### 4.2.5 Discussion

The results of our experiments show, that (1) the fully automated analysis of phone-based SVF is feasible for dementia screening, (2) the phone-based pipeline produces classification results comparable to the gold-standard manual transcription-based classifiers and (3) the word error rate for the ASR approach is acceptable despite the reduced telephone bandwidth and the aged population.

In general, regarding screening scenarios, high sensitivity scores are important. Our classification experiment based on the fully automated pipeline shows a good AUC and for screening scenario a good sensitivity of 0.85 and decent specificity of 0.65. For achieving better specificity results, it may be necessary to include additional tasks, especially focusing on the differentiation of MCI and healthy controls. Nevertheless, this is not the main goal for broad screening, as false positives are less expensive for a health-care system than false negatives.

In our experiments, the automated ASR-/phone-based screening pipeline and the pipeline based on manually transcribed speech reach comparable classification results. This is very encouraging, as the transcription of speech is the number-one resource-straining factor,

showing that an automatic speech-based system has become a powerful alternative to manual analysis of speech-based psychometric tests.

ASR is often considered to be the main weakness in speech based automatic screening approaches (Tóth et al., 2015). Our results show an overall error rate of 33.4 % for the automated system, compared to the manual transcripts. This result represents an improvement over results of other authors using ASR systems for evaluating the SVF tasks (Lehr et al., 2012; Pakhomov et al., 2015). In line with previous research, more word errors are produced by the ASR for dementia patients, compared to healthy subjects, which can be explained by age-related speech erosion. Considering the types of errors, insertions and deletions are both problematic for further analysis, as they skew the raw word count, the single most predictive performance indicator in SVF for dementia detection. Substitutions affect the word count less, only in rare cases, where a word is substituted with a previously named one, generating a false repetition.

#### 4.2.6 *Conclusion*

In this paper we set out to benchmark a telephone-based analysis of SVF for inclusion into a fully automated dementia frontline screening for global risk assessment. Our results show that SVF is a prime candidate for inclusion into an automated pipeline, providing decent sensitivity and specificity scores. Additionally, we show that the phone-based classification is as effective as the gold-standard manual transcription-based classifier displaying an acceptable ASR word error rate despite telephone setup and the aged sample for the experiments.

Further research will be directed into finding additional tests, that offer increased sensitivity and specificity in combination with SVF. The idea of this series is to validate and construct a system, that solely based on the telephone as a technological interface and administrable in less than 10 minutes, perfectly fits the need of broad dementia screening tools. It should also serve epidemiological research studies and inclusion for pharmaceutical trials, which aim at including representative shares of the population by cost-effective screening for persons with early onset neurocognitive impairments.

#### 4.3 CHAPTER CONCLUSION

This chapter demonstrates how afore-established theoretical insights on the neurocognitive profile of AD patients at different clinical stages of the disease (MCI & Dementia) can be harnessed within the greater challenge of efficiently fighting AD as a societal challenge. Clinical research agrees that as long as pharmaceutical industry has not yet brought effective treatment to the market, the most promising mid-term strategy would be to screen cost-effectively for early signs of AD at scale (Dubois et al., 2016; Laske et al., 2015). Whereas early cognitive signs of the disease as measured by the ecologically highly valid SVF can be validly extracted even at a neurocognitive function level (as shown in Chapter 3) this does not guarantee cost-effectiveness and scalability. On the one hand, the need for human resources both in the annotation as well as in the evaluation process of the SVF is incompatible with both aspects, cost-effectiveness as well as scalability. On the other hand, requiring testees to come to dedicated medical facilities and burdening the healthcare infrastructure, means an additional major obstacle to scalability.

The first paper in this chapter therefore investigated how automatic processing of SVF speech data could potentially lower the human manual processing resources involved (i.e. speech transcription and annotation of qualitative SVF markers). In the first paper, a combination of ASR and novel computational qualitative SVF metrics reached comparable screening decisions as compared to human transcriptions and evaluations of the SVF. The screening decision between mild and major neurocognitive impairment was in this case experimentally defined as one and two standard deviations below norm population in the SVF word count. This is partially in line with Petersen (2014) but of course not considering impairments in other domains. This was taken one step further and the confirmed diagnosis (HC vs. MCI vs. AD) was tried to automatically infer in an ML classification scenario. In this final experiment, again features extracted from ASR-grade speech annotation performed on-par with the manual transcript ones in ML classification scenarios.

Providing a proof of concept for a maximum-scalable AD screening scenario, the second paper shows evidence for the technical feasibility of a telephone-based dementia screening, using the SVF and its advanced qualitative analysis as introduced earlier. Accordingly, the second paper introduces a screening scenario in which the SVF would be recorded over an ordinary landline telephone (a technical prerequisite nearly every household

in a developed country fulfills). The recordings are then ran through an ASR system and fed into an automatic feature extraction pipeline extracting afore-mentioned qualitative SVF features. Finally, these features are being evaluated by an ML classifier for a screening decision. The results show, that the fully automated analysis of phone-quality SVF is feasible for dementia screening if connected to a ML classification approach. Furthermore, results show that the performance of an ASR-based approach is acceptable despite the reduced telephone bandwidth of the speech sample and the aged population. Screening based on the fully automated pipeline showed a good sensitivity of 0.85 and decent specificity of 0.65 for impaired cognition (MCI + Dementia) vs. HC. A high sensitivity is the to-optimize performance measure in a screening scenario.

To conclude, chapter 4 showed that an automated ASR-/phone-based screening pipeline based on the SVF produces comparable results to manually transcribed speech. This is very encouraging, as it shows how the qualitative analysis of the SVF detecting early differentiated cognitive changes in clinical stages of AD dementia can be transferred into a cost-effective scalable real-world solution. Thereby this work opens up exciting new possibilities for the mid-term strategy in fighting AD on a societal level by enabling population-wide screening at low cost and low patient burden.

## 5 OVERARCHING DISCUSSION AND CONCLUSION

This thesis set out to achieve an ambitious two-fold objective: (1) improve our understanding of the progressive executive function and semantic memory impairment and their interplay in clinical AD as measured by the Semantic Verbal Fluency (SVF) and (2) harness those insights into the different neurocognitive function AD profiles for applied early AD screening; for a visual overview of this thesis see Figure 24 below. This overarching chapter will contain a discussion about the impairment of semantic memory and executive function in AD as measured through the SVF and its implication for clinical decision support in screening.

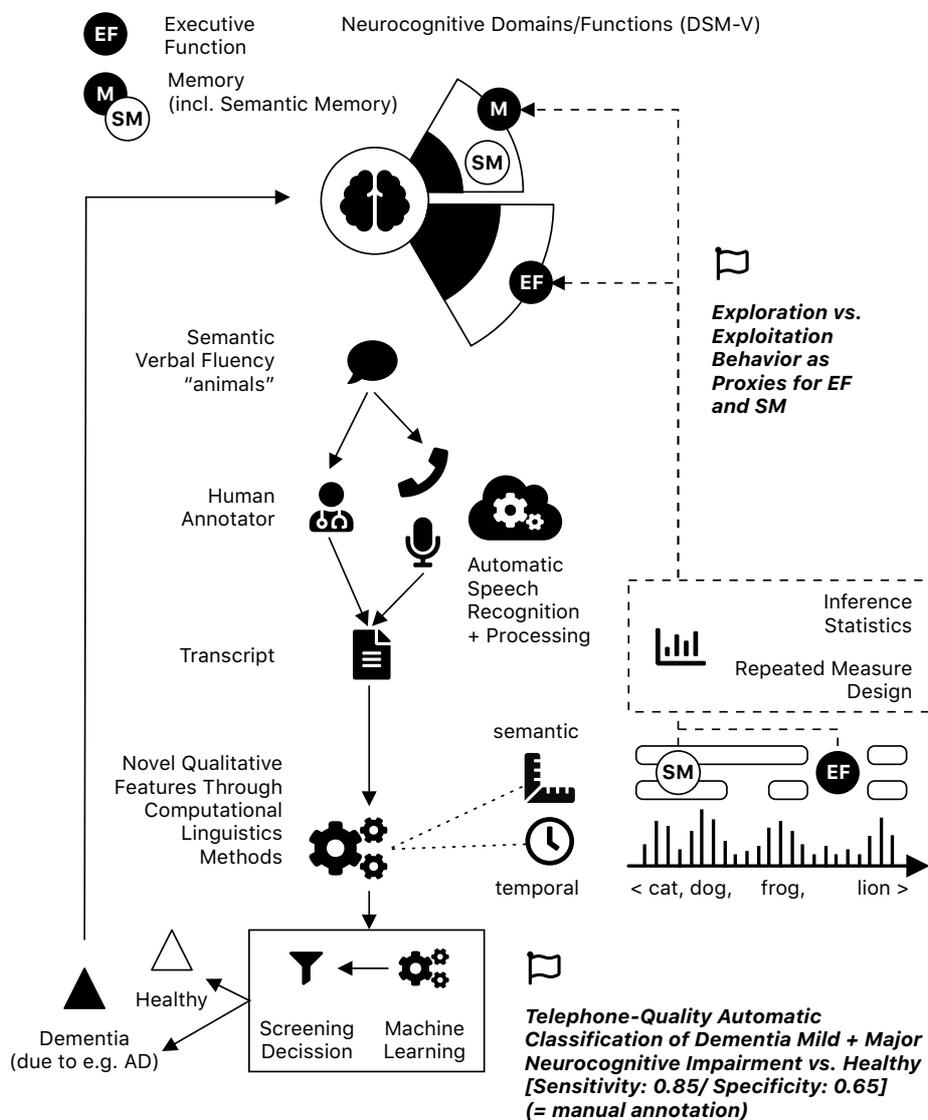


Figure 24: Visual abstract of this thesis' major scientific building blocks and results. Dotted lines represent the main line of research from chapter 3—establishing the neurocognitive basis of novel computational qualitative features. The solid lines represent the main line of achievements underlying chapter 4—proving real-world AD screening societal impact.

## 5.1 NEUROCOGNITIVE AD PROFILES FROM SVF—EMBEDDED INTO NEUROSCIENCE

This thesis helps to better understand the role of executive function and semantic memory impairment at both prodromal as well as acute clinical AD stages measured by the SVF. Combining both computational semantic as well as temporal modalities in the qualitative analysis of the SVF, the results show that semantic memory is structurally affected from an early aMCI stage and particularly worsened through the inability to compensate by engaging executive function. This effect prevails even when repeatedly confronted with the same task. Hence, over the course of the disease, hampered executive functioning is assumed to be the main driver of later-stage AD patients' notably poor cognitive performance in the SVF.

The SVF is a compound task drawing from multiple neurocognitive functions including, amongst others, primarily executive function and semantic memory (Amunts et al., 2020; Rohrer et al., 1995). However, using traditional analysis methods of the SVF—i.e. word count of correct responses or error count—it is difficult, if not impossible, to disentangle the involvement of both mentioned neurocognitive functions. This limits the usability of the SVF for differentiated neuropsychological assessments (Shao et al., 2014). Receiving a clean signal on both neurocognitive functions is a major challenge in utilizing the SVF as more sophisticated neuropsychological assessment. However, novel item-level qualitative computational SVF measures give an insight to disentangling the involvement of executive function and semantic memory, as they objectively model SVF production patterns using temporal as well as semantic modalities. This thesis combines both modalities in a qualitative analysis of the SVF to best differentiate between executive function and semantic memory involvement and thereby investigating the pathological profiles of AD dementia at different stages.

### 5.1.1 *Distinguishing AD-Related Executive Function & Semantic Memory Impairment*

Both executive function as well semantic memory impairment are relevant throughout the clinical stages of AD and can be measured by qualitative aspects of the SVF. Therefore, in a first step, the involvement of both neurocognitive functions has to be disentangled. This thesis achieves this through combining temporal as well semantic modalities for qualitative computer-supported analysis of SVF-performances. Within this frame, two studies were conducted. One study investigated the exploration and exploitation patterns in patients' SVF performance as a proxy for executive function and semantic memory. The second study

monitored the quality of improvement in repeated SVF assessment, putting additional focus on procedural (related to executive function) aspects of the SVF performance. The first study shows, that at an acute clinical dementia stage of AD, hampered executive function might be the main driver behind the starkly impaired SVF performance which is in line with some previous research (Peter et al., 2016; Raoux et al., 2008; Shao et al., 2014). Due to small effect sizes, the first study only showed trends of similar neurocognitive functions profile in a prodromal clinical stage, Mild Cognitive Impairment (MCI). Building on this trend, the second study introduced a new paradigm in which participants conducted the SVF three times in a row. This repeated assessment of classic cognitive tasks (such as the SVF) has recently been shown to be one of the most underestimated yet sensitive cognitive assessment paradigms to robustly identify AD-related early cognitive decline (Jutten et al., 2020). With this respect, repeated testing is supposed to be delivering higher sensitivity for measuring the respective cognitive function impairments in MCI. The second study ruled out pure novelty effects (Thorgusen et al., 2016) and made the influence of executive functions more visible. The results show that amnesic MCI (aMCI) patients did not improve their SVF performance over the three trials (in contrast to healthy controls). More importantly aMCI patients did not seem to possess the executive function resources to improve their production strategy as measured by improved exploration patterns and a steadier production of words over time (in contrast to healthy controls). Nonetheless at first assessment the structural deficits in the sense of hampered semantic memory became obvious in the second study even at an aMCI stage.

These results show that AD manifests in impaired executive function as well as impaired semantic memory at both stages of clinical cognitive symptoms (prodromal stage, aMCI & acute stage, dementia) and that this drives the well-documented SVF impairment in clinical practice. However, there is a notable stage-related difference in the visibility of both neurocognitive functions' impairment in the SVF. At an early clinical aMCI stage, executive function deficits are present but less obvious. To nevertheless model them, more sensitive markers such as the lack of practice effects or lack of improvement in the SVF performance are needed. The semantic memory impairment however, seems accentuated also at this early aMCI stage. At an acute clinical dementia stage, the semantic memory impairment remains present but the SVF performance—as measured by computational qualitative markers—gets particularly hampered through a progressively prominent executive function impairment.

Earlier research has shown that AD patients' ineffective exploration patterns represent the main effect underlying the overall reduced SVF count (Murphy et al., 2006; Raoux et al., 2008). This has also been shown for aMCI (Gomez & White, 2006; Raoux et al., 2008) although effect sizes are typically smaller. This thesis however, found that the impaired exploration pattern in aMCI is initially of a structural nature pointing towards a breakdown of associative structure within the semantic memory system of aMCI subjects. During the repeated testing, the observed structure loss is compounded by the secondary impairment in executive function which is one hypothesis for why aMCI patients perform poorly, as they cannot compensate for the primary memory impairment with more efficient search patterns. This, similar to previous research, manifests as a lack of SVF practice effects as compared to strong practice effects in healthy control subjects (Cooper et al., 2001; Duff et al., 2008, 2011).

In sum, the presented results highlight that an SVF test requires a compound performance between ready to use semantic memory structures with nodes and links between them that are activated through their associative neighborhood (Goñi et al., 2011) and executive function effort that can be invested by the testee to faster or more efficiently navigate this structure. As AD progresses, not only does the structure continue to break down, but secondarily the patients are deprived of their capabilities to compensate for this semantic memory structure loss resulting in devastatingly exponential cognitive decline. This interpretation of the purely behavioral findings from this thesis should be put into perspective of other neurological research on AD providing additional important explanations for the presented behavioral observations.

#### *5.1.2 Computer-Supported Qualitative SVF Markers and Classic AD Neuroscience*

It is worth combining here-presented behavioral findings with evidence from research on non-symptomatic AD through cognitive reserve as well as neurological research on the progressive nature of AD. Some elderly individuals, on the one hand, satisfy the AD neuropathology (e.g. AD pathology established postmortem) but others show little to no cognitive symptoms during their lifetime (Dubois et al., 2016). This phenomenon suggests the involvement of additional mediating factors such as cognitive reserve. Cognitive reserve is commonly defined as a broad set of characteristics that render an individual less vulnerable to AD neurodegeneration such as increased baseline cognitive capacity, education, or the extensive use of compensation strategies (Rentz et al., 2010; Roe et al., 2011). This should be

put into relation to chapter 3.2, showing that aMCIs overall did not—as by this thesis' understanding—engage additional executive function (compensation) strategies. The repeated SVF assessment in combination with the here-established computational qualitative markers (see chapter 3.1), could be a perfect setup to measure cognitive reserve in aMCI or AD through modeling compensation strategies. In this context, (Buckner, 2004) stated that “evidence suggests that compensation for brain decline in aging [especially accelerated through AD] may partly account for why some older adults age gracefully and others decline rapidly” (p. 204). Classic rapid AD dementia-related cognitive decline is therefore considered as a multifactorial interaction of executive function as well as (semantic) memory impairment that results in an exponentially deteriorating global cognition, as both cognitive impairments mutually affect each other. Neuroscience research suggests a two-factor model explaining pathological ageing (i.e. especially because of AD) suffering particularly from the interaction between impaired prefrontal regions (associated with executive function) and impaired medial temporal regions (associated with memory function). In this model, normal ageing would primarily affect prefrontal areas but spare medial temporal regions from accelerated decline (Braak & Braak, 1991; Hedden & Gabrieli, 2004).

The studies comprised in this thesis reflect these neurological theories on a behavioral level, as we find that semantic memory structurally is affected at an early aMCI stage of the disease, particularly worsened through the inability to compensate by engaging executive function. This is the case even if testees are repeatedly confronted with the same task. This merges with the first study (section 3.1) finding hampered executive functioning to be at first glance the main driver of later-stage AD patients' notably poor SVF performance (i.e. through ineffective exploration patterns).

Next to functional imaging, today's gold-standard assessment of AD neuropathology makes use of proteomics-based AD biomarkers (Drummond & Wisniewski, 2020). Specifically, an increased amount of beta amyloid plaques is associated with early clinical AD pathology that can be identified through means of Cerebrospinal Fluid (CSF) probing or Positron Emission Tomography (PET) (Vlassenko, Benzinger, & Morris, 2012). Over the last decades, the so-called '*amyloid hypothesis*' puts forward that an accumulation of amyloid beta is the toxic cause of AD neurodegeneration. However, nowadays increased amyloid beta levels are rather regarded as a downstream result and not the initial cause of AD (Drachman, 2014).

Independent of the causal relationship, the amyloid beta level is still a valuable AD indicator. Hence, amyloid beta positivity ( $A\beta+$ ) is regarded as one of the gold-standard biomarkers for AD pathology indicating especially at-risk subjects before clinical symptoms set in (Dubois et al., 2016). Recent findings show that  $A\beta+$  is associated with cognito-behavioral impairment in AD and most notably can be measured through SVF impairment (Papp et al., 2016). Moreover, it has been shown that SVF (on animals) is especially sensitive to  $A\beta+$  (Mirandez, Aprahamian, Talib, Forlenza, & Radanovic, 2017). The fact that recent AD biomarker research and this thesis' behavioral findings overlap in the SVF represents an interesting link for future work. As both biomarkers and behavioral qualitative SVF-markers overlap on their conceptual links to neurocognitive function impairments, future studies should confirm the cognito-behavioral correlates of AD-related biomarkers as measured by the here-presented qualitative analysis of the SVF.

In summary, the first part of this thesis shows that computer-supported qualitative analysis of the SVF has the potential to model side-by-side AD-specific semantic memory and executive function impairments across the clinical stages of AD. Findings point towards an early-on structural semantic memory impairment that cannot be compensated through available spared executive functioning at a prodromal level and is gravely worsened through an additional executive function impairment that commences at an acute dementia stage. Via the SVF these findings can be directly connected to recent advances in the field of biomarker-related neurological research on AD disease trajectories. Hence, early AD-related cognitive impairment could be modelled through the SVF in line with state-of-the-art research using gold-standard neurological AD markers. This raises the question, whether the computer-supported qualitative SVF analysis could be used as a light-weight diagnosis (screening) approach for early-stage prodromal AD cases that would traditionally require invasive and costly biomarker testing.

## 5.2 THE SVF AS A COST-EFFECTIVE AND SCALABLE ASSESSMENT FOR AD

The second part of this thesis shows how afore-established theoretical insights on semantic memory and executive function impairment, as measured by the SVF across the clinical trajectory of AD (MCI & Dementia), can be harnessed to more efficiently screen for AD at a population level. Automatic Speech Recognition (ASR) based automatic SVF analysis pipelines are not resulting in a reduced overall screening performance but perform on par with a human annotation-based scheme. Moreover, the automated analysis pipeline is not particularly vulnerable to telephone-based down-sampling of participants' speech samples. This proves that an overall scalable telephone-based AD screening is feasible using the SVF alone.

As long as the pharmaceutical industry has not brought disease modifying agents to the market, the societal challenge of AD is a diagnostic one with two strategical components: screening for AD and AD disease monitoring (Dubois et al., 2016). Both strategies try to achieve a timely—as early as possible—insight into the progression between different disease stages. However, AD screening tries to identify the initial progression from healthy to pathological ageing, whereas monitoring identifies stage-related progression within pathological trajectories of AD. For both strategies the same imperatives apply: cost-effectiveness and scalability. In other words, screening as well as monitoring for AD should be effective but come at minimum resources and a maximum outreach.

While the first part of this thesis has shown that dedicated neurocognitive function signs of prodromal and acute clinical AD can be extracted with the help of computer-supported SVF analysis, the challenge remains to embed this into a cost-effective and scalable technical setup. The SVF does not only qualify through excellent psychometric properties but also through high ecological validity as the test situation is relatively close to everyday use of language. Moreover, from a patient-centric point of view, the SVF comes with considerably low patient burden as it typically takes only 60s and no additional materials nor lengthy test-instructions are required.

Taking this into consideration, chapter 4 provides answers to cost-effectiveness and scalability of the SVF as an early AD screening and monitoring solution. Cost effectiveness can be achieved through automatic processing of SVF speech input and extraction of afore-established qualitative computational measures for cognitive function profiling in AD.

Scalability on the other hand can be achieved through the concept of a telephone-based SVF assessment detaching the assessment from medical and clinical infrastructure. In both cases a diagnostic screening/ monitoring decision is simulated through a machine learning classifier which is common practice in the field of AD diagnosis support literature (König et al., 2015; Pellegrini et al., 2018).

### *5.2.1 Automatic Qualitative Analysis of the SVF for Cost-effective AD Screening*

Automatic processing of SVF speech data could potentially lower the human manual processing resources involved (i.e. speech transcription and annotation of qualitative SVF markers) in a qualitative, more fine-grained analysis that tells about semantic memory and executive function impairment in AD (compare Section 4.1). Automatic Speech Recognition (ASR) and downstream automatic classification (machine learning based diagnostic decision making) are two essential building blocks in this effort. To allow for automatic processing of the SVF in terms of qualitative feature extraction, speech has to be initially transformed into text initially. This, typically, is also the most resource intensive step if done manually. To overcome this challenge, Automatic Speech Recognition (ASR) can be used to automatically transcribe speech reducing the time and cost to generate speech transcripts. However, there is a common belief that ASR transcripts are not fit for medical-grade speech processing introducing transcription errors into the signal and thereby hampering the diagnostic interpretation.

Therefore, comparing the performance of an ASR-based decision support system against a process based on human transcripts of speech is naturally the first step to take (see section 4.1). Results showed that while automatically generated transcripts implied a relatively high task-specific verbal fluency error rate of 20%, this outcome represents a systematic artifact across all tested groups (AD, MCI and healthy elderlies). Therefore, the relatively high error rate did not affect the overall psychometric properties of the automatically evaluated SVF assessment and yielded on-par diagnostic screening results with the manual annotation (AUC over .90 for AD vs. healthy controls). Importantly, qualitative measures of the SVF, such as semantic clustering and switching behavior indicative of underlying semantic memory and executive function impairments, were found to be comparable as calculated from automatic and manual transcriptions.

In order to achieve those results, two technical components play a major role: ASR and downstream machine learning classification.

*Automatic Speech Recognition in Diagnostic Speech Analysis*

Automatic Speech Recognition (ASR) is one of the most promising tools for this setup but also poses a major concern in terms of quality transcriptions. This is due to multiple aspects: (1) ASR performance is a moving target continuously improving as AI applications for NLP or computer science in general improves, (2) ASR is based on learned statistical or neural models that will most likely never be perfect and maintain some level of error but can achieve reliable performance at the present, (3) ASR performance is vulnerable to a complex interplay of surrounding factors that humans might not explicitly think of or see as obvious. The following text will elaborate on the three mentioned aspects.

First, ASR is often considered to be the main limiting factor (Pakhomov et al., 2015; Tóth et al., 2015) in such speech-based diagnostic scenarios, including those that are purely based on the SVF (Pakhomov et al., 2015). However, since its experimental stages, ASR is a consistently improving tool and in the last decade reached industrial standard performance levels (Juang & Rabiner, 2005). Following the rapid technical development in the field of computer science and natural language processing, the performance of ASR will continue to improve (D. Yu & Deng, 2015). Since the key to training ASR systems is excessively large amounts of curated data, the most accurate results from ASR systems are achieved using industrially maintained systems, such as the Google Speech API<sup>6</sup> or Amazon AWS<sup>7</sup> ASR (Amazon Web Services), rather than training models independently. This in itself bears additional risks such as depending on remote cloud-computing systems or regulatory challenges when it comes to including a non-static performance component into a potential future diagnostic application. Especially in the medical device domain or clinical trials environment, assessment methods have to fulfill rigorous regulatory requirements that can exclude a performance component that might improve over a web-based dynamic interface (or even worsen) after initial regulatory clearance of the technology. Even with these considerations, the results of section 4.1 show a low overall error rate of 20% for the

---

<sup>6</sup> <https://cloud.google.com/speech/>

<sup>7</sup> <https://aws.amazon.com/transcribe/>

automated system (compared to the manual transcripts). Which is above reported ASR performances in previous literature for evaluating the SVF tasks (Lehr et al., 2012; Pakhomov et al., 2015); this illustrates well the influence of technical improvement over time on ASR system performance, with this thesis' section 4.1 published in 2018, benefitting from three or five years of additional ASR industrial advances.

Second, ASR has reached industrial standard performance since 2010 as major smartphone manufacturers include speech recognition in their products for personal voice-based assistants and other applications. Still, it remains and probably always will remain an imperfect system due to its nature of underlying statistical methods. This remains to be a major paradox especially for clinical research; human annotators and the clinical scientific community tend to perceive human-annotations as perfect or at least not error-prone in their speech recognition abilities which in fact is not true, as it has been shown that in certain tasks ASR systems plainly outperform human speech recognition performance (e.g. Cooke, Hershey, & Rennie, 2010). However, ASR outperforming humans is a rare case and is mentioned here purely to illustrate the divide between common opinion and scientific reality.

Third, ASR systems having many situational parameters that determine performance for specific situations remain an inaccessible black box for people that lack technical expertise in the field. This is a potential breeding ground for doubts and sheer rejection of the clinical research community to systematically consider ASR-based evaluation of traditional speech assessments. Major influencing factors for ASR performance are often not obvious or explicitly considered. These include domain of application, language and its underlying resources or physical surrounding conditions during speech recordings (e.g. background noise) reducing robustness of ASR (S. Watanabe, Delcroix, Metze, & Hershey, 2017). One way of improving ASR performance to a clinical task is by providing hints to the system by constraining the domain of the output. For a certain speech input, ASR systems typically give multiple transcript options with a probability rating. Based on the possible outcomes, performance could be improved by helping the system with 'hints' of the domain. For example, In the SVF-on-animals case, performance could be improved by constraining the output of the ASR system to animals or restricting it to nouns in general. 'Ant' for example could be recognized as 'aunt' or 'and' but could be ruled out as non-animal applying clever domain-based performance

optimization techniques, possibly reducing the overall error rate of the automatic transcription.

Recent applied diagnostic NLP research also sometimes uses a work-around that exploits exactly the fact of an imperfect ASR system in the domain of speech-based dementia assessment. The confusion of an ASR language model in understanding the speech and eventually transcribing it can be measured through the perplexity index, which is provided by standard web-based ASR tools (e.g. Google Cloud Speech API) alongside the list of suggestions for transcription of the given speech input. This perplexity score can be used in an inverse application to detect AD-related speech deterioration (T. Cohen & Pakhomov, 2020; Weiner, Engelbart, & Schultz, 2017). In fact, this approach has been proven to be well performing, as the ASR tends to systematically misrecognize AD patients' speech input due to multiple confounding factors such as lower pitch or bad articulation but also due to unexpected linguistic errors. In line with this, a significantly increased number of errors made by ASR for AD patients was found, compared to healthy subjects. However, if improved ASR systems will, in the future, be able to correctly recognize speech of AD patients, ASR perplexity might not be a helpful feature anymore.

There is a significant body of research on setups using ASR systems and features based on those ASR transcripts to classify between AD and healthy subjects in a similar way as this thesis presents in chapter 4. Studies report similar, but also comparably worse performances to results presented here, using an ASR based pipeline for cognitive assessment in AD. This partially depends on the cognitive task being evaluated but also on the type of extracted markers/features for the final machine learning classification of pathological subjects. Regarding the task, results have been reported from relatively unconstrained speech assessments such as retelling a story (Wechsler Logic Memory WLM; Lehr et al., 2012), spontaneous free speech initiated by questions (Mirheidari, Blackburn, Reuber, Walker, & Christensen, 2016) or picture descriptions (Sadeghian, David Schaffer, & Zahorian, 2017; Zhou, Fraser, & Rudzicz, 2016). However, there is also evidence on ASR-based AD classification from a similar setup on the SVF (Pakhomov et al., 2015). A common way to evaluate the performance of an ASR system is by looking at its Word Error Rate (WER), which represents the percentage of wrongly transcribed words as compared to the manual ground truth transcription. Overall, comparable studies report higher WER than chapter 4 of this thesis:

27% (Lehr et al., 2012), 41% (Mirheidari et al., 2016), 31% (Sadeghian et al., 2017), 38% (Zhou et al., 2016) and 30% (Pakhomov et al., 2015). All studies used various methods to increase ASR performance such as speaker adaptation or semantic checks on the suggested transcript options (similar to the presented approach on animals). The relatively high WER may be correlated with the quality of the speech data; picture descriptions from the DementiaBank<sup>8</sup> corpus, for example, are of relatively low audio quality and very noisy (Sadeghian et al., 2017; Zhou et al., 2016). On the other hand task-specific WER is low if the transcript can be screened for on-task words leaving only a fraction of the overall transcript and thereby drastically reducing the WER; this is the case for studies that rely on predefined semantic content that is clinically relevant such as to-be-remembered story elements in the Wechsler Logical Memory (WLM; Lehr et al., 2012) or the list of possible animals in the SVF (Pakhomov et al., 2015).

Regardless of the ASR WER, extracting state-of-the-art NLP features from an ASR-transcribed text input will most probably result in classification algorithms working with ASR-resilient features that help to reach state-of-the-art classification performance in AD vs. healthy subjects (Zhou et al., 2016). Similarly, Clark and colleagues (D. G. Clark et al., 2016) found that automatic SVF scores—or especially the NLP-based multitude of features—yielded higher accuracy for the prediction of conversion from healthy to dementia as compared to manual scores. However, this relates to an ongoing discussion about explainability and clinically valid features used in ‘diagnostic’ ML approaches for AD. In a nutshell: not every well-performing ML feature represents clinically valid variance of the disease’s cognitive profile. Some features are just exploiting non-cognitive systematic variance correlated with the disease (e.g. age, affective comorbidity, etc.). To prevent the introduction of these confounding factors, chapter 4 of this thesis restricted the analysis to qualitative SVF features that have been clinically validated in Chapter 3.

#### *Downstream Diagnostic Decision Support Using ASR-Based Input*

ASR and downstream automatic ML-based classification are two essential building blocks on the way to a concept for automatic screening of prodromal and acute AD based on qualitative SVF analysis. After ASR-based pre-processing of SVF speech recordings, qualitative features can be extracted from the SVF that are indicative of semantic memory and executive

---

<sup>8</sup> <https://talkbank.org/DementiaBank/>

function impairment and sensitive to AD-related cognitive decline as shown in chapter 3. Those features are then used in an additional downstream classification of AD vs. healthy subjects to simulate practical clinical decision support. Chapter 4 presents encouraging results showing improved classification results for AD when using additional qualitative SVF markers eventually performing on par with human SVF annotation-based decisions.

For this additional downstream classification of AD vs. healthy subjects, ASR-based systems often achieve similar results as the manual annotation-based ones (see chapter 4)(Lehr et al., 2012; Sadeghian et al., 2017) or initially worse but then improved through a more liberal feature selection process (e.g. Mirheidari et al., 2016; Zhou et al., 2016). The downstream classification results of the ASR-based diagnostic system presented in chapter 4 are only comparable to some of the previous literature (Lehr et al., 2012; Sadeghian et al., 2017). Similarly, chapter 4 relied on features/variables based on the semantic content of the speech (i.e. named animals) which have an explicit equivalent in clinical research tradition (i.e. computationally implemented clustering and switching measures, following Troyer et al., 1997). This is different to other studies that initially report low classification performance when relying on the word error rate (WER)-prone semantics for clinical variables but then switch through the means of machine learning feature selection to a different type of features (mostly para-linguistic or acoustic properties of speech). These features circumvent the problems associated with WER as transcriptions are not needed for para-linguistic/ acoustic measures. A classifier relying on those features might also be discriminative in this very scenario, hence the WER-resilient features boost classification performance but have no clinical feature equivalent (Mirheidari et al., 2016; Zhou et al., 2016).

The results presented in chapter 4 of this thesis rely on semantic information despite the ASR-related disadvantages and use only features that are recognized in (experimental) clinical research. This is important for the intended clinical application: supporting the human expert decision. Therefore, the downstream automatic decision support application (i.e. the ML classifier differentiating AD vs. healthy) has to be based upon clinically accepted and validated features from the very pathological domain (i.e. neuropsychology on AD). By using clinically trusted measures in the downstream classification it can be argued that the results are more transparent, explainable and generalizable. Another major advantage is that human experts can comprehend these features, making them more likely to be adopted into clinical

practice (Chandler et al., 2020). Letting a classifier switch from semantic features to more 'robust' WER-resistant features (acoustic or non-linguistic properties of speech) might overfit to the specific assessment situation and context. This often results in good classification performance but doesn't create the same quality of clinically actionable insights needed for real clinical decision support in the form of screening or monitoring for AD.

### *5.2.2 Feasibility of an SVF-Based Scalable AD Screening Approach*

Automatic processing and screening for prodromal AD based on the SVF is based on ASR and the downstream machine learning classification. After providing evidence for both components separately, section 4.2 presents a combined proof of concept for a scalable AD screening scenario and first feasibility results. In this concept the SVF would be administered and recorded over an ordinary landline telephone which is a technical prerequisite nearly every household in a developed country fulfills. Then, the recording would be automatically transcribed via an ASR system and adjacently fed into an automatic feature extraction pipeline extracting afore-mentioned qualitative SVF features. Those features would finally be evaluated by an ML classifier for a screening decision. Beyond introducing this technical concept of an AD screening application, this section also proved that a telephone-quality speech signal of the SVF can be fed into an adjacent automatic qualitative analysis pipeline without dramatically worsening the ASR performance. Compared to the high-quality recorded samples automatically analyzed in the first paper (task-relevant WER ~ 20%) results from section 4.2 are worse but still represent an acceptable task-relevant WER of around 30%. This task-relevant WER of around 30% is amongst the better results of WERs reported in the field on similar screening approaches (i.e. Lehr et al., 2012; Pakhomov et al., 2015); however comparable work does not use telephone-quality down sampling of the audio file. To simulate a downstream screening application again an ML-based classification scenario was used which showed good sensitivity of 0.85 and decent specificity of 0.65 for impaired cognition (MCI + Dementia) versus Healthy Controls (HC).

These results make an important contribution to an overall multidisciplinary field of work targeting AD disease interception through timely screening and early intervention at a prodromal non-dementia AD stage. As introduced earlier, the societal challenge of AD essentially is a diagnostic one: as long as there is no effective pharmacological treatment for acute AD clinical stages on the market, early detection is the imperative mid-term solution.

This is not only important for non-pharmaceutical interventions but also for near-future pharmaceutical interventions that might only intercept the disease trajectory at a prodromal early stage. Particularly in a highly developed industrialized society (e.g. Europe or US) where only 50% of all AD cases are diagnosed (Prince et al., 2016) putting cost-effective frontline screening solutions at the center of the challenge. The presented fully automatic analysis of an SVF recorded over ordinary telephone achieves 85% sensitivity for detecting clinical stages of AD. Thereby this thesis proposes a powerful concept for cost-effective and scalable screening of AD addressing exactly this diagnostic problem.

AD dementia screening and monitoring over the phone employing easy-to-administer speech-based assessments have been previously investigated (e.g. the Telephone Interview for Cognitive Status—TICS; Brandt, Spencer, & Folstein, 1988). Early work using the TELE interview reported very good results on the identification of dementia cases either as sampled from a registered database with already established diagnosis (Gatz et al., 1995) or with undiagnosed samples with diagnosis being established in the call follow-up call (Gatz et al., 2002). In both studies, sensitivity was well above 80%, comparable to the results presented in section 4.2, but specificity was significantly higher (~ 90%). Focusing more on classic psychometric testing, Lipton and colleagues (2003) explicitly compared SVF and a particular memory test (four items cued associative recall) procedure to the interview-based TICS. They found the memory assessment to have better sensitivity and specificity than the other two procedures. They reported a lower sensitivity (78%) for the SVF as compared to the results reported here (85%). There are also more recent studies that leverage telemedical scenarios for screening requiring physical testing material with the patient (Vestal, Smith-Olinde, Hicks, Hutton, & Hart, 2006). Such a procedure however proved a major disadvantage due to a complex technical setup. However, all of those mentioned phone-based screening studies embrace the concept of scalability through telephone as a means of administration. Still, they require manual evaluation and therefore at max provide a remote and scalable option but fall short in terms of cost-effectiveness. Such an approach is scalable but requires the same amount of human resources for evaluation as a classical testing scenario. Although this definitely has value in some scenarios (e.g. remote diagnostics has become a huge need in 2020-2021 COVID-19 pandemic), it is not fit for the purpose of population-wide AD screening with low-cost tests of high sensitivity and lower specificity (Dubois et al., 2016).

Overall, there have been approaches reported earlier that address the same challenge in a similar fashion compared to this thesis' chapter 4 but fall short in one or more requirements to serve as a feasible solution for cost-effective broad early AD screening. Therefore, I would like to sum up three main emerging factors that determine the success of such an approach: (1) the difficulty of the psychometric task has to be sensitive to all stages of AD dementia, (2) the implementation of the downstream decision logic should be state-of-the-art and (3) time is always a factor. The following paragraphs will explain those conclusions in more detail.

First, many of the related studies focus on the clinical dementia stage, using tasks that are arguably not suited for detecting MCI patients with relatively preserved cognitive functioning (e.g. memory tests using only four items or simple time orientation tasks). Although those studies report excellent sensitivity on a demented population, at an MCI stage sensitivity would probably drop significantly due to the low difficulty level of the psychometric assessment. In the here-presented studies from chapter 4, performance has been evaluated across both dementia and MCI stages reaching decent performance. More importantly the SVF is an open-ended performance task that possesses excellent discriminatory power across the full spectrum of cognitive performance. This is a major benefit as compared to a predefined memory task.

Secondly, the screening decision logic often is implemented as a simple cut-off score. Often multiple tasks are scored and eventually all are summed up in one overall score encompassing different kinds of assessments: simple questions about the date as well as complex memory recalls. This method is similar to the traditional Mini-Mental State Exam (MMSE; Folstein et al., 1975) which stages dementia cognitive syndrome severity over a total score of 30. Recent advances in statistics, especially in ML, un-arguably provide more sophisticated methods to establish a screening decision. Results from section 4.1 show the superiority of an ML-based decision logic over a classic standard-deviation cutoff-based one.

Finally, when considering population-wide frontline screening, time is the most sensitive factor. The SVF in the presented studies needs at most four minutes to be administered, including instructions. Other screening studies though report administration times well over 15 minutes. When scaling up for broad screening coverage across a certain population this becomes a decisive factor.

### 5.3 OVERALL CONCLUSION

To conclude, this thesis advances current AD research in a two-fold manner: (1) improving the understanding of the decline of executive function and semantic memory in AD as measured through computational qualitative analysis of the SVF and (2) embedding these theoretical advances into practical clinical decision support concepts that help cost-effectively screen population-wide for early-stage AD.

First, recent advances in computer-supported qualitative analysis of the SVF help measuring executive function and semantic memory using temporal as well as semantic modalities. This thesis shows that aMCI and AD dementia are marked through both semantic memory as well as executive function impairments. Semantic memory is structurally affected from an early aMCI stage and particularly worsened through the inability to compensate by engaging executive function. This effect prevails even when repeatedly confronted with the same task. Hence, over the course of the disease, hampered executive functioning is found to be the main driver of later-stage AD patients' notably poor performance in the SVF and probably poor cognition overall.

Second, harnessing the computer-supported qualitative analysis of the SVF, this thesis shows that automatic processing of SVF speech data represents a cost-effective as well as scalable solution to screen and monitor population-wide for AD-related cognitive decline. This thesis showed that ASR-based automatic SVF analysis pipelines are not resulting in a reduced overall screening performance but perform on par with a human annotation-based scheme. In addition, this thesis also proved the feasibility of an overall scalable telephone-based SVF assessment concept, showing that the automated analysis pipeline is not particularly vulnerable to telephone-based down-sampling of participants' speech samples.

This is very encouraging, as it shows how the qualitative analysis of the SVF detecting subtle and early semantic memory as well as executive function impairments in clinical stages of AD dementia can be transferred into a cost-effective scalable screening solution. This might be a pivotal element within the AD mid-term strategy which demands early intervention for better therapeutic success (Dubois et al., 2016). It has been shown that interventions at early AD stages show promising results and are more likely to be effective (Sindi et al., 2015). Thereby, this work could substantially contribute to the mid-term strategy in fighting AD on a societal level by enabling population-wide screening at low cost and low patient burden.

From a clinical perspective, the SVF is, and always has been, one of the most sensitive tasks to efficiently detect cognitive impairment especially within AD. However, its simultaneous demands in semantic memory and executive function makes the SVF task difficult for a more differential diagnosis of impaired cognitive functions (Shao et al., 2014). This thesis provides the qualitative insights into the AD-related SVF impairment to better separate both neurocognitive functions. Thereby, this thesis helps to establish AD stage-related impairment profiles of executive function and semantic memory (Guarino et al., 2019; Verma & Howard, 2012) through one of the most-adopted tasks in clinical practice, the SVF.

Alzheimer's Disease has a huge impact on an ever-ageing society of highly developed and industrialized countries such as EU member states including Germany. The socio-economic impact on our *global* society is estimated to grow up to more than \$1 trillion US dollars by 2030 (Wimo et al., 2017). Understanding different neurocognitive function impairments and their specific patterns along the AD clinical trajectory remains an important target and results can be translated into more specific diagnostic tools. Combined with recent advances in computer science, powerful screening applications emerge at this interdisciplinary juncture. In that sense, this thesis makes an important contribution to the overall societal mid-term strategy in the fight against AD.

## 6 OUTLOOK

This thesis makes important contributions to better understand the semantic memory and executive function impairment in both prodromal and acute clinical AD stages through harnessing computer-supported qualitative analysis of the SVF. Additionally, this thesis shows how this improved understanding can be translated in more cost-effective and scalable screening of early clinical AD. Three main future research topics emerge from these results: (1) Establishing construct-validity of novel computational qualitative SVF markers, (2) longitudinal SVF-based monitoring of cognition and profiling of preclinical AD for preventative pharmaceutical trials and (3) improving AD diagnosis in primary health-care for more efficient transition between sectors.

### *Construct Validity of Novel Computational Qualitative SVF Markers*

This thesis established the neuropsychological importance of novel computational qualitative SVF markers for AD. Future research should help to further confirm those markers' psychometric properties also beyond the clinical case of AD. As discussed in chapter 3 the SVF's psychometric properties and involved neurocognitive domain constructs have been subject to a significant body of research. The involvement of lexico-semantic memory, and executive function, related to strategic search and retrieval processes has been well established (Amunts et al., 2020; Shao et al., 2014). However, these studies mainly investigate what neurocognitive domains and subdomains predict the overall traditional quantitative performance marker for SVF: the word count. It has been discussed and shown multiple times though, that the SVF draws resources from multiple neurocognitive domains. Therefore, solely regarding underlying neurocognitive constructs of the overall word count will not be conclusive (Shao et al., 2014). This is why qualitative markers of an SVF performance (i.e., clustering and switching) have been introduced in the first place (A. K. Troyer et al., 1997). But there is today very little literature using the same construct validity approaches on the qualitative measures as have been used on the overall quantitative performance metric (i.e. the overall word count). A notable exception is the work of Unsworth, Spillers and Brewer (2011) who found different latent constructs to be correlated with clustering (Working Memory & Vocabulary) and switching (Working Memory & Processing Speed) as initially established by Troyer (A. K. Troyer et al., 1997).

Clinical research though has been providing evidence about what traditional qualitative SVF markers signify. These works draw evidence from different pathological groups that are associated with certain neurocognitive impairments. There is evidence from Schizophrenia patients being impaired in clustering during the SVF arguably as a result of their executive function impairment which manifests in disorganized thinking (Robert et al., 1997). Additional evidence stems from focal lesions (A. K. Troyer, Moscovitch, Winocur, Alexander, & Stuss, 1998), Parkinson's patients (A. K. Troyer et al., 1998), fronto-temporal dementia and Primary Progressive Aphasia (PPA) (Van Den Berg et al., 2017). Within the scope of this thesis, there is a lot of evidence from studies with AD patients (March & Pattison, 2006; Peter et al., 2016; Price et al., 2012; A. K. Troyer et al., 1998; Weakley & Schmitter-Edgecombe, 2014). However, as already discussed in chapter 0, this body of evidence is contradictory in itself drawing no clear picture whether and how clustering and switching capture the main neurocognitive impairments in AD. This is partially why this thesis introduced a novel computational methodology for qualitatively investigating SVF performance in AD.

Recent novel computational qualitative markers hence have been investigated similarly on dedicated pathological groups such as AD (D. G. Clark et al., 2016; Linz, Tröger, Alexandersson, & König, 2017; Linz, Tröger, Alexandersson, Wolters, et al., 2017; Pakhomov et al., 2016; Pakhomov & Hemmy, 2014) or traumatic brain injuries (Woods et al., 2016). However, those studies vary significantly in their methodological implementation of the automatic semantic measurements. This is mainly because of the usage of different semantic embeddings to model inter-word semantic distance as has been discussed in chapter 0.

Future research should therefore try to unify or at least compare underlying computational approaches for qualitatively modelling semantic organization within the SVF performance. This means directly comparing different semantic embedding approaches for clustering and semantic coherence measures in the SVF. At the same time future work should also apply a more rigorous construct validation approach to novel computational measures in healthy ageing populations like it has been used on the traditional qualitative Troyer markers earlier (Unsworth et al., 2011).

#### *Automatic SVF Analysis for Longitudinal Monitoring in AD Trials*

This thesis shows how the SVF could be used to measure both executive function and semantic memory in an efficient manner in experimental cross-sectional settings. In future,

these results should be leveraged in longitudinal SVF-based monitoring of cognition and profiling of pre-clinical AD for preventative pharmaceutical trials. Today there is a common understanding that the ideal effective long-term strategy to fight AD is by developing disease modifying drugs that prevent cognitive decline into clinical dementia stage (secondary prevention of AD; Mortamais et al., 2017; Ritchie et al., 2016). After decades of pharmaceutical research focusing on clinical AD stages, the majority of clinical trials failed and a new—secondary prevention—era of clinical trials has been proclaimed (R. A. Sperling et al., 2014). Evidence from longitudinal studies focusing on preclinical AD as well as AR-AD (Asymptomatic @ risk AD; Dubois et al., 2016) indicate that subtle cognitive symptoms can be found although patients are not classified as being in a clinical stage of AD due to current diagnostic standards (Donohue et al., 2014; Mortamais et al., 2017). The rationale behind latest secondary preventative trials in AD is the following: Subjects with subtle within-person cognitive decline that do not show clinical cognitive symptoms yet (neither aMCI nor dementia) and that at the same time fulfill biomarker indications for an at-risk state (AR-AD) are very likely to convert to clinical AD in the future. These subjects therefore represent the prime target for novel preventative pharmaceutical and non-pharmaceutical therapies.

The main challenge in such a scheme is to efficiently identify AR-AD subjects, as those subjects have not yet been officially registered with cognitive problems within the health-care system. The cost-effective yet scalable automatic SVF-based screening concept presented in this thesis can play an important role for answering this challenge. Additionally, longitudinal models about subtle intra-personal cognitive change with respect to different neurocognitive functions will be key. There is evidence that classic cognitive tests (especially the SVF) are able to measure amyloid-related cognitive decline at a pre-clinical stage. This only becomes evident though with advanced statistics measures (e.g. machine learning) and longitudinal modeling (Donohue et al., 2014; Papp et al., 2016). Consequently, it has been shown that SVF is sensitive to AR-AD biomarker status (Mirandez et al., 2017; Terrera, Harrison, Ritchie, & Ritchie, 2020) and therefore represents an ideal brief monitoring assessment for preclinical AD (Jutten et al., 2020; Papp et al., 2016).

Methodologically, it can be assumed that pre-clinical AD can only be detected through continuous monitoring of a person's cognitive abilities. This is because at such an early stage changes in cognitive performance that are actually indicative of future pathological cognitive

decline vary a lot between subjects and can only be identified through a within-person longitudinal approach (Jutten et al., 2020). Although episodic memory performance would be the natural marker for such longitudinal modelling, there is research suggesting that longitudinally monitored semantic memory is even superior to episodic memory as biomarker for pre-clinical AD (Venneri, Mitolo, & De Marco, 2016). At the same time, it can be assumed that classic cognitive assessment evaluations will probably not be sensitive enough to model subtle pre-clinical AD-related cognitive changes. Therefore, more comprehensive qualitative analysis schemes and more advanced statistical models build through ML on large cohorts should be engaged (Amieva et al., 2005).

However, when using repeated testing for continuous longitudinal monitoring of cognition, practice effects are likely to occur. Traditionally practice effects are regarded as a major confounder for clinical AD studies especially in longitudinal monitoring settings (Goldberg et al., 2015). This thesis however showed how repeated testing and the occurrence of practice effects can be a valuable signal itself for detecting AD-related neurocognitive impairments. Building upon this thesis' findings, the diagnostic power of practice effects should be especially considered because their absence could be a sign for preclinical AD (Hassenstab et al., 2015). There are multiple studies that consider PE as standalone diagnostic evidence for AD-related early cognitive impairment. Focusing on SVF, studies have shown no PE for MCI (Duff et al., 2008) or at least smaller PE than HC (Cooper et al., 2001, 2004). Therefore, future work should try to harness SVF-related PE for even better monitoring and screening of preclinical stage, probable AD.

#### *Transferring Results into Primary Health-care*

Last but not least, this thesis introduced a powerful concept for cost-effective and scalable AD screening. This perfectly stages research on improving AD diagnosis in primary health-care for more efficient transition between sectors. In Germany, as in the majority of European countries, the first point of contact for people with cognitive impairments is the family doctor—primary care (Winter, Maaz, & Kuhlmeier, 2006). A prospective study showed that in a sample of about 600 family doctor patients screened positive for dementia, only 40% of the patients were diagnosed with dementia (Eichler et al., 2014). This percentage is consistent with the rates of undiagnosed dementia from a meta-analysis of international studies (Lang et al., 2017). It can therefore be assumed that the current care system in

Germany does not usually diagnose dementia. An as-early-as-possible established diagnosis opens transfer into specialist care—secondary care—helps to optimize stays in the hospital—tertiary care—as well as helps to minimize side effects of existing pharmaceutical intervention plans for age-related comorbidities (Michalowsky et al., 2016). A diagnosis also opens up access to the assistance system, to support for relatives and to the organization of outpatient help to prevent premature institutionalization.

The above-mentioned positive consequences of an early diagnosis reflect the current care system which might be changing as soon as pharmaceutical research succeeds in the long-term AD strategy by getting a disease-modifying drug to the market. If a potentially disease-modifying treatment for prodromal and early AD dementia becomes available in the coming years, the healthcare system will face the challenge of providing targeted advice to many citizens concerned about their memory. The first point of contact for these people would again be the primary care sector—the family doctor. In the future, this will require even more low-threshold yet reliable instruments for early diagnosis and case identification. Such instruments should already identify persons at risk for dementia at the GP level, who would benefit from further diagnostics and at the same time spare persons who do not have an increased risk for cognitive disorders from further diagnostic assessment burden.

Here is where this thesis' computer-supported automatic analysis of the SVF for screening AD can be a potential starting point. Offering relatively comprehensive yet brief neuropsychological assessment at low cost and little patient burden would make upstream diagnosis or exclusion of cognitive impairment already affordable in primary care.

In summary, this thesis' results motivate three clear future avenues of AD research. Eventually, longitudinal SVF-based monitoring of cognition and profiling of preclinical AD could be an essential stepping stone for future preventative pharmaceutical AD trials. This could help to finally bring an effective treatment for AD to the market. From a methodological point of view, novel computational SVF markers have been accepted in their ability to indicate AD-related neurocognitive function impairments. However, there is still some lack in traditional cognitive psychology research on their construct-validity as compared to a broad range of psychometrics. Finally, at the very end of this AD research there should be implications for the primary health-care sector. This thesis' concept for cost-effective and

scalable SVF-based AD screening could therefore be the first step for preventative AD frontline diagnosis at the ground level of everyday health-care.

## BIBLIOGRAPHY

- Abbott, A. (2011). Dementia: A problem for our age. *Nature*, 475(7355 SUPPL.). doi:10.1038/475S2a
- Aisen, P., Touchon, J., Amariglio, R., Andrieu, S., Bateman, R., Breitner, J., ... Vellas, B. (2017). EU/US/CTAD Task Force: Lessons learned from recent and current Alzheimer's prevention trials. *The Journal of Prevention of Alzheimer's Disease*, 4(2), 116–124.
- Alderwick, H., & Dixon, J. (2019). The NHS long term plan. *BMJ (Clinical Research Ed.)*, 364, l84.
- Al-hameed, S., Benaissa, M., & Christensen, H. (2016). Simple and robust audio - based detection of biomarkers for Alzheimer' s disease. *7th Workshop on Speech and Language Processing for Assistive Technologies*, 32–36.
- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub.
- Amieva, H., Jacqmin-Gadda, H., Orgogozo, J. M., Le Carret, N., Helmer, C., Letenneur, L., ... Dartigues, J. F. (2005). The 9 year cognitive decline before dementia of the Alzheimer type: A prospective population-based study. *Brain: A Journal of Neurology*, 128(5), 1093–1101.
- Amunts, J., Camilleri, J. A., Eickhoff, S. B., Heim, S., & Weis, S. (2020). Executive functions predict verbal fluency scores in healthy participants. *Scientific Reports*, 10(1), 1–11.
- Auriacombe, S., Lechevallier, N., Amieva, H., Harston, S., Raoux, N., & Dartigues, J. F. (2006). A longitudinal study of quantitative and qualitative features of category verbal fluency in incident Alzheimer's disease subjects: Results from the PAQUID study. *Dementia and Geriatric Cognitive Disorders*, 21(4), 260–266.

- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226.
- Bateman, R. J., Xiong, C., Benzinger, T. L. S., Fagan, A. M., Goate, A., Fox, N. C., ... Dominantly Inherited Alzheimer Network. (2012). Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *The New England Journal of Medicine*, *367*(9), 795–804.
- Birn, R. M., Kenworthy, L., Case, L., Caravella, R., Jones, T. B., Bandettini, P. a., & Martin, A. (2010). Neural systems supporting lexical search guided by letter and semantic category cues: A self-paced overt response fMRI study of verbal fluency. *NeuroImage*, *49*(1), 1099–1107.
- Boersma, P., & Weenink, D. Praat: doing phonetics by computer (Version 6.1.42 ). Retrieved from <https://www.fon.hum.uva.nl/praat/>
- Braak, H., & Braak, E. (1991). Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica*, *82*(4), 239–259.
- Brando, E., Olmedo, R., & Solares Canal, C. (2017). *The application of technologies in dementia diagnosis and intervention: A literature review*. Retrieved from <https://repositorio.uc.cl/xmlui/bitstream/handle/11534/29999/The%20application%20of%20technologies%20in%20dementia%20diagnosis%20and%20intervention%20A%20literature%20review.pdf>
- Brandt, J., & Manning, K. J. (2009). Patterns of word-list generation in mild cognitive impairment and Alzheimer's disease. *The Clinical Neuropsychologist*, *23*(5), 870–879.
- Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *1*(2), 111–117.

- Buckner, R. L. (2004). Memory and executive function in aging and ad: Multiple factors that cause decline and reserve factors that compensate. *Neuron*, *44*(1), 195–208.
- Canning, S. J. D., Leach, L., Stuss, D., Ngo, L., & Black, S. E. (2004). Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. *Neurology*, *62*(4), 556–562.
- Carson, N., Leach, L., & Murphy, K. J. (2018). A re-examination of Montreal Cognitive Assessment (MoCA) cutoff scores. *International Journal of Geriatric Psychiatry*, *33*(2), 379–388.
- Castanho, T. C., Portugal-Nunes, C., Moreira, P. S., Amorim, L., Palha, J. A., Sousa, N., & Correia Santos, N. (2016). Applicability of the Telephone Interview for Cognitive Status (Modified) in a community sample with low education level: association with an extensive neuropsychological battery. *International Journal of Geriatric Psychiatry*, *31*(2), 128–136.
- Chandler, C., Foltz, P. W., Cohen, A. S., Holmlund, T. B., Cheng, J., Bernstein, J. C., ... Elvevåg, B. (2020). Machine learning for longitudinal applications of neuropsychological testing. *Intelligence-Based Medicine*, 100006.
- Chung, P.-C., Hsu, Y.-L., Wang, C.-Y., Lin, C.-W., Wang, J.-S., & Pai, M.-C. (2012, May). Gait analysis for patients with Alzheimer's disease using a triaxial accelerometer. *2012 IEEE International Symposium on Circuits and Systems*. Presented at the 2012 IEEE International Symposium on Circuits and Systems - ISCAS 2012, Seoul, Korea (South). doi:10.1109/iscas.2012.6271484
- Clark, D. G., McLaughlin, P. M., Woo, E., Hwang, K., Hurtz, S., Ramirez, L., ... Apostolova, L. G. (2016). Novel verbal fluency scores and structural brain imaging for prediction of

cognitive outcome in mild cognitive impairment. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 2, 113–122.

Clark, L. J., Gatz, M., Zheng, L., Chen, Y.-L., McCleary, C., & Mack, W. J. (2009). Longitudinal verbal fluency in normal aging, preclinical, and prevalent Alzheimer's disease. *American Journal of Alzheimer's Disease and Other Dementias*, 24(6), 461–468.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Cohen, T., & Pakhomov, S. (2020). *A Tale of Two Perplexities: Sensitivity of Neural Language Models to Lexical Retrieval Deficits in Dementia of the Alzheimer's Type*. Retrieved from <http://arxiv.org/abs/2005.03593>

Collie, A., & Maruff, P. (2000). The neuropsychology of preclinical Alzheimer's disease and mild cognitive impairment. *Neuroscience and Biobehavioral Reviews*, 24(3), 365–374.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2), 240–247.

Cooke, M., Hershey, J. R., & Rennie, S. J. (2010). Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24(1), 1–15.

Cooper, D. B., Epker, M., Lacritz, L., Weine, M., Rosenberg, R. N., Honig, L., & Cullum, C. M. (2001). Effects of practice on category fluency in Alzheimer's disease. *The Clinical Neuropsychologist*, 15(1), 125–128.

Cooper, D. B., Lacritz, L. H., Weiner, M. F., Rosenberg, R. N., & Cullum, C. M. (2004). Category fluency in mild cognitive impairment. *Alzheimer Disease and Associated Disorders*, 18(3), 120–122.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

- Costa, A., Bak, T., Caffarra, P., Caltagirone, C., Ceccaldi, M., Collette, F., ... Cappa, S. F. (2017). The need for harmonisation and innovation of neuropsychological assessment in neurodegenerative dementias in Europe: consensus document of the Joint Program for Neurodegenerative Diseases Working Group. *Alzheimer's Research & Therapy*, 9(1), 27.
- Cowppli-Bony, P., Fabrigoule, C., Letenneur, L., Ritchie, K., Alperovitch, A., Dartigues, J. F., & Dubois, B. (2005). Le test des 5 mots : validité dans la détection de la maladie d'Alzheimer dans la population générale. *Revue neurologique*, 161(12), 1205–1212.
- Crawford, J. R., & Henry, J. D. (2005). Assessment of executive dysfunction. In *The Effectiveness of Rehabilitation for Cognitive Deficits* (pp. 233–246). Oxford University Press.
- Crowell, T. A., Luis, C. A., Vanderploeg, R. D., Schinka, J. A., & Mullan, M. (2002). Memory patterns and executive functioning in Mild Cognitive Impairment and Alzheimer's Disease. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 9(4), 288–297.
- de Jager, C. A., Schrijnemaekers, A.-C. M. C., Honey, T. E. M., & Budge, M. M. (2009). Detection of MCI in the clinic: evaluation of the sensitivity and specificity of a computerised test battery, the Hopkins Verbal Learning Test and the MMSE. *Age and Ageing*, 38(4), 455–460.
- Dodge, H. H., Mattek, N., Gregor, M., Bowman, M., Seelye, A., Ybarra, O., ... Kaye, J. A. (2015). Social Markers of Mild Cognitive Impairment: Proportion of Word Counts in Free Conversational Speech. *Current Alzheimer Research*, 12(6), 513–519.

- Donohue, M. C., Sperling, R. A., Salmon, D. P., Rentz, D. M., Raman, R., Thomas, R. G., ... Aisen, P. S. (2014). The preclinical Alzheimer cognitive composite: Measuring amyloid-related decline. *JAMA Neurology*, *71*(8), 961–970.
- Drachman, D. A. (2014). The amyloid hypothesis, time to move on: Amyloid is the downstream result, not cause, of Alzheimer's disease. *Alzheimer's and Dementia*, *10*(3), 372–380.
- Drummond, E., & Wisniewski, T. (2020). Using Proteomics to Understand Alzheimer's Disease Pathogenesis. In T. Wisniewski (Ed.), *Alzheimer's Disease* (pp. 37–51). Brisbane (AU): Codon Publications.
- Dubois, B., Hampel, H., Feldman, H. H., Scheltens, P., Aisen, P., Andrieu, S., ... Jack, C. R. (2016). *Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria* (Vol. 12, pp. 292–323).
- Dubois, B., Slachevsky, A., Litvan, I., & Pillon, B. (2000). The FAB. *Neurology*, *55*(11), 1621–1626.
- Duff, K., Beglinger, L. J., Van Der Heiden, S., Moser, D. J., Arndt, S., Schultz, S. K., & Paulsen, J. S. (2008). Short-term practice effects in amnesic mild cognitive impairment: Implications for diagnosis and treatment. *International Psychogeriatrics / IPA*, *20*(5), 986–999.
- Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., ... McCaffrey, R. J. (2011). Practice effects predict cognitive outcome in amnesic mild cognitive impairment. *The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry*, *19*(11), 932–939.

- Égerházi, A., Berecz, R., Bartók, E., & Degrell, I. (2007). Automated Neuropsychological Test Battery (CANTAB) in mild cognitive impairment and in Alzheimer's disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 31(3), 746–751.
- Eichler, T., Thyrian, J. R., Hertel, J., Köhler, L., Wucherer, D., Dreier, A., ... Hoffmann, W. (2014). Rates of formal diagnosis in people screened positive for dementia in primary care: results of the DelpHi-Trial. *Journal of Alzheimer's Disease: JAD*, 42(2), 451–458.
- El-Hayek, Y. H., Wiley, R. E., Khoury, C. P., Daya, R. P., Ballard, C., Evans, A. R., ... Atri, A. (2019). Tip of the Iceberg: Assessing the Global Socioeconomic Costs of Alzheimer's Disease and Related Dementias and Strategic Implications for Stakeholders. *Journal of Alzheimer's Disease: JAD*, 70(2), 321–339.
- Epelbaum, S., Genthon, R., Cavedo, E., Habert, M. O., Lamari, F., Gagliardi, G., ... Dubois, B. (2017). Preclinical Alzheimer's disease: A systematic review of the cohorts underlying the concept. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 13(4), 454–467.
- Fernaes, S. E., & Almkvist, O. (1998). Word Production: Dissociation of Two Retrieval Modes of Semantic Memory Across Time. *Journal of Clinical and Experimental Neuropsychology*, 20(2), 137–143.
- Fernaes, S. E., Östberg, P., Hellström, Å., & Wahlund, L. O. (2008). Cut the coda: Early fluency intervals predict diagnoses. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, 44(2), 161–169.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3), 189–198.

- Fowler, K. S., Saling, M. M., Conway, E. L., Semple, J. M., & Louis, W. J. (1997). Computerized neuropsychological tests in the early detection of dementia: prospective findings. *Journal of the International Neuropsychological Society: JINS*, 3(2), 139–146.
- Franco, C., Demongeot, J., Villemazet, C., & Vuillerme, N. (2010). Behavioral telemonitoring of the elderly at home: Detection of nycthemeral rhythms drifts from location data. *24th IEEE International Conference on Advanced Information Networking and Applications Workshops, WAINA 2010*, 759–766.
- Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2015). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease: JAD*, 49, 407–422.
- Fraser, K. C., Rudzicz, F., & Hirst, G. (2016). Detecting late-life depression in Alzheimer's disease through analysis of speech and language. *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 1–11.
- Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *The Journal of Artificial Intelligence Research*, 34, 443–498.
- Galton, C. J., Patterson, K., Xuereb, J. H., & Hodges, J. R. (2000). Atypical and typical presentations of Alzheimer's disease: a clinical, neuropsychological, neuroimaging and pathological study of 13 cases. *Brain: A Journal of Neurology*, 123 Pt 3, 484–498.
- Gatz, M., Reynolds, C. A., John, R., Johansson, B., Mortimer, J. A., Pedersen, N. L., ... John, R. (2002). Telephone Screening to Identify Potential Dementia Cases in a Population-Based Sample of Older Adults. *International Psychogeriatric Association*, 14(3), 273–289.

- Gatz, M., Reynolds, C., Nikolic, J., Lowe, B., Karel, M., & Pedersen, N. (1995). An Empirical Test of Telephone Screening to Identify Potential Dementia Cases. *International Psychogeriatrics / IPA*, 7(3).
- Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 1(1), 103–111.
- Gomez, R. G., & White, D. A. (2006). Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 21(8), 771–775.
- Goñi, J., Arrondo, G., Sepulcre, J., Martincorena, I., Vélez De Mendizábal, N., Corominas-Murtra, B., ... Villoslada, P. (2011). The semantic organization of the animal category: Evidence from semantic verbal fluency and network theory. *Cognitive Processing*, 12(2), 183–196.
- Gosztolya, G., Tóth, L., Grósz, T., Vincze, V., Hoffmann, I., Szatlóczki, G., ... Kálmán, J. (2016). Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 107–111.
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. Retrieved from <http://arxiv.org/abs/1802.06893>
- Grill, J. D., & Karlawish, J. (2010). Addressing the challenges to successful recruitment and retention in Alzheimer's disease clinical trials. *Alzheimer's Research & Therapy*, 2(6), 34.

- Grober, E., Ocepek-Welikson, K., & Teresi, J. a. (2009). The Free and Cued Selective Reminding Test : evidence of psychometric adequacy. *Psychology Science Quarterly*, 51(3), 266–282.
- Gruenewald, P. J., & Lockhead, G. R. (1980). The free recall of category examples. *Journal of Experimental Psychology. Human Learning and Memory*, 6(3), 225–240.
- Gualtieri, C. T., & Johnson, L. G. (2006). Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 21(7), 623–643.
- Guarino, A., Favieri, F., Boncompagni, I., Agostini, F., Cantone, M., & Casagrande, M. (2019). Executive functions in Alzheimer disease: A systematic review. *Frontiers in Aging Neuroscience*, 10(January). doi:10.3389/fnagi.2018.00437
- Hassenstab, J., Ruvolo, D., Jasiolec, M., Xiong, C., Grant, E., & Morris, J. C. (2015). *Absence of Practice Effects in Preclinical Alzheimer ' s Disease*. 29(6), 940–948.
- Hedden, T., & Gabrieli, J. D. E. (2004). Insights into the ageing mind: a view from cognitive neuroscience. *Nature Reviews. Neuroscience*, 5(2), 87–96.
- Henry, J. D., & Crawford, J. R. (2005). A meta-analytic review of verbal fluency deficits in depression. *Journal of Clinical and Experimental Neuropsychology*, 27(1), 78–101.
- Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia*, 42(9), 1212–1222.
- Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in Semantic Fields: How We Search Through Memory. *Topics in Cognitive Science*, 7(3), 513–534.

- Hoffmann, I., Nemeth, D., Dye, C. D., Pákási, M., Irinyi, T., & Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *International Journal of Speech-Language Pathology*, *12*(1), 29–34.
- Hsu, Y.-L., Chung, P.-C. J., Wang, W.-H., Pai, M.-C., Wang, C.-Y., Lin, C.-W., ... Wang, J.-S. (2014). Gait and balance analysis for patients with Alzheimer's disease using an inertial-sensor-based wearable instrument. *IEEE Journal of Biomedical and Health Informatics*, *18*(6), 1822–1830.
- Inoue, M., Jimbo, D., Taniguchi, M., & Urakami, K. (2011). Touch Panel-type Dementia Assessment Scale: a new computer-based rating scale for Alzheimer's disease. *Psychogeriatrics: The Official Journal of the Japanese Psychogeriatric Society*, *11*(1), 28–33.
- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models. Retrieved from <http://arxiv.org/abs/1612.03651>
- Juang, B.-H., & Rabiner, L. R. (2005). Automatic speech recognition--a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, *1*, 67.
- Jutten, R. J., Sikkes, S. A. M., Amariglio, R. E., Buckley, R. F., Properzi, M. J., Marshall, G. A., ... Papp, K. V. (2020). Identifying sensitive measures of cognitive decline at different clinical stages of alzheimer's disease. *Journal of the International Neuropsychological Society: JINS*, 1–13.
- Kalbe, E., Kessler, J., Calabrese, P., Smith, R., Passmore, a. P., Brand, M., & Bullock, R. (2004). DemTect: A new, sensitive cognitive screening test to support the diagnosis of mild

cognitive impairment and early dementia. *International Journal of Geriatric Psychiatry*, 19(2), 136–143.

Khodabakhsh, A., Yesil, F., Guner, E., & Demiroglu, C. (2015). Evaluation of Linguistic and Prosodic Features for Detection of Alzheimer's Disease in Turkish Conversational Speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 9, 1–15.

Klumpp, P., Janu, T., Arias-Vergara, T., Correa, J. C. V., Orozco-Arroyave, J. R., & Nöth, E. (2017). Apkinson - A mobile monitoring solution for Parkinson's disease. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2017-Augus*, 1839–1843.

König, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., & Robert, P. H. (2018). Use of Speech Analyses within a Mobile Application for the Assessment of Cognitive Impairment in Elderly People. *Current Alzheimer Research*, 15(2), 120–129.

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., ... David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1, 112–124.

Lambon Ralph, M. A., Patterson, K., Graham, N., Dawson, K., & Hodges, J. R. (2003). Homogeneity and heterogeneity in mild cognitive impairment and Alzheimer's disease: a cross-sectional and longitudinal study of 55 cases. *Brain: A Journal of Neurology*, 126(Pt 11), 2350–2362.

Lang, L., Clifford, A., Wei, L., Zhang, D., Leung, D., Augustine, G., ... Chen, R. (2017). Prevalence and determinants of undetected dementia in the community: a systematic literature review and a meta-analysis. *BMJ Open*, 7(2), e011146.

- Langbaum, J. B., Fleisher, A. S., Chen, K., Ayutyanont, N., Lopera, F., Quiroz, Y. T., ... Reiman, E. M. (2013). Ushering in the study and treatment of preclinical Alzheimer disease. *Nature Reviews. Neurology*, *9*(7), 371–381.
- Laske, C., Sohrabi, H. R., Frost, S. M., López-de-Ipiña, K., Garrard, P., Buscema, M., ... O'bryant, S. E. (2015). Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimer's and Dementia*, *11*(5), 561–578.
- Ledoux, K., Vannorsdall, T. D., Pickett, E. J., Bosley, L. V., Gordon, B., & Schretlen, D. J. (2014). Capturing additional information about the organization of entries in the lexicon from verbal fluency productions. *Journal of Clinical and Experimental Neuropsychology*, *36*(2), 205–220.
- Lehr, M., Prud'hommeaux, E., Shafran, I., & Roark, B. (2012). Fully automated neuropsychological assessment for detecting Mild Cognitive Impairment. *Proceedings of Interspeech 2012*, 1–4.
- Lehrl, S. (2005). *Mehrfachwahl-Wortschatz-Intelligenztest MWT-B*. Balingen, Germany: Spitta Verlag.
- Lerner, A. J., Ogrocki, P. K., & Thomas, P. J. (2009). Network graph analysis of category fluency testing. *Cognitive and Behavioral Neurology: Official Journal of the Society for Behavioral and Cognitive Neurology*, *22*(1), 45–52.
- Linz, N., Lundholm Fors, K., Lindsay, H., Eckerström, M., Alexandersson, J., & Kokkinakis, D. (2019). *Temporal Analysis of the Semantic Verbal Fluency Task in Persons with Subjective and Mild Cognitive Impairment*. *2*, 103–113.

- Linz, N., Tröger, J., Alexandersson, J., & König, A. (2017). Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. *IWCS 2017—12th International Conference on Computational Semantics—Short Papers.*, 1.
- Linz, N., Tröger, J., Alexandersson, J., Wolters, M., König, A., & Robert, P. (2017, November). Predicting Dementia Screening and Staging Scores from Semantic Verbal Fluency Performance. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 719–728. [ieeexplore.ieee.org](http://ieeexplore.ieee.org).
- Lipton, R. B., Katz, M. J., Kuslansky, G., Sliwinski, M. J., Stewart, W. F., Verghese, J., ... Buschke, H. (2003). Screening for dementia by telephone using the Memory Impairment Screen. *Journal of the American Geriatrics Society*, 51(10), 1382–1390.
- López-de-Ipiña, K., Alonso, J.-B., Travieso, C. M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., ... Martinez de Lizardui, U. (2013). On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, 13(5), 6730–6745.
- López-de-Ipiña, K., Martinez-de-Lizarduy, U., Calvo, P. M., Mekyska, J., Beitia, B., Barroso, N., ... Ecañ-Torres, M. (2018). Advances on Automatic Speech Analysis for Early Detection of Alzheimer Disease: A Non-linear Multi-task Approach. *Current Alzheimer Research*, 15(2), 139–148.
- March, E. G., & Pattison, P. (2006). Semantic verbal fluency in Alzheimer's disease: Approaches beyond the traditional scoring system. *Journal of Clinical and Experimental Neuropsychology*, 28(4), 549–566.

- Marczinski, C. A., & Kertesz, A. (2006). Category and letter fluency in semantic dementia, primary progressive aphasia, and Alzheimer's disease. *Brain and Language, 97*(3), 258–265.
- Mayr, U. (2002). On the dissociation between clustering and switching in verbal fluency: Comment on Troyer, Moscovitch, Winocur, Alexander and Stuss. *Neuropsychologia, 40*(5), 562–566.
- McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Jr, Kawas, C. H., ... Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 7*(3), 263–269.
- McWhinney, B. (1991). The CHILDES project. In *Tools for Analyzing Talk*. LEA New Jersey.
- Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., Carcavilla, N., & Ivanova, O. (2018). Voice Markers of Lexical Access in Mild Cognitive Impairment and Alzheimer's Disease. *Current Alzheimer Research, 15*(2), 111–119.
- Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., & Arana, J. M. (2014). Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders, 37*, 327–334.
- Michalowsky, B., Eichler, T., Thyrian, J. R., Hertel, J., Wucherer, D., Hoffmann, W., & Flessa, S. (2016). Healthcare resource utilization and cost in dementia: are there differences between patients screened positive for dementia with and those without a formal diagnosis of dementia in primary care in Germany? *International Psychogeriatrics, 28*(3), 359–369.

- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from <http://arxiv.org/abs/1301.3781>
- Mioshi, E., Dawson, K., Mitchell, J., Arnold, R., & Hodges, J. R. (2006). The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. *International Journal of Geriatric Psychiatry, 21*(11), 1078–1085.
- Mirandez, R. M., Aprahamian, I., Talib, L. L., Forlenza, O. V., & Radanovic, M. (2017). Multiple category verbal fluency in mild cognitive impairment and correlation with CSF biomarkers for Alzheimer's disease. *International Psychogeriatrics / IPA, 29*(6), 949–958.
- Mirheidari, B., Blackburn, D., Reuber, M., Walker, T., & Christensen, H. (2016). Diagnosing people with dementia using automatic Conversation Analysis. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 08-12-Sept, 1220–1224*. International Speech and Communication Association.
- Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., Fillenbaum, G., ... Clark, C. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology, 39*(9), 1159–1165.
- Morris, John C. (1997). Clinical Dementia Rating: A Reliable and Valid Diagnostic and Staging Measure for Dementia of the Alzheimer Type. *International Psychogeriatrics / IPA, 9*(Supplement S1), 173–176.

- Mortamais, M., Ash, J. A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., ... Ritchie, K. (2017). Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility. *Alzheimer's and Dementia*, *13*(4), 468–492.
- Mueller, K. D., Kosciak, R. L., LaRue, A., Clark, L. R., Hermann, B., Johnson, S. C., & Sager, M. A. (2015). Verbal fluency and early memory decline: Results from the Wisconsin registry for Alzheimer's prevention. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *30*(5), 448–457.
- Munro Cullum, C., Hynan, L. S., Grosch, M., Parikh, M., & Weiner, M. F. (2014). Teleneuropsychology: Evidence for video teleconference-based neuropsychological assessment. *Journal of the International Neuropsychological Society: JINS*, *20*(10), 1028–1033.
- Murphy, K. J., Rich, J. B., & Troyer, A. K. (2006). Verbal fluency patterns in amnesic mild cognitive impairment are characteristic of Alzheimer's type dementia. *Journal of the International Neuropsychological Society: JINS*, *12*(04), 570–574.
- Nasreddine, Z., Phillips, N., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., ... Chertkow, H. (2005). *Montreal Cognitive Assessment (MoCa) - Administration and Scoring Instructions* (pp. 1–11). pp. 1–11.
- Nikolai, T., Bezdicek, O., Markova, H., Stepankova, H., Michalec, J., Kopecek, M., ... Vyhnalek, M. (2017). Semantic verbal fluency impairment is detectable in patients with subjective cognitive decline. *Applied Neuropsychology. Adult*, 1–10.
- Nutter-Upham, K. E., Saykin, A. J., Rabin, L. A., Roth, R. M., Wishart, H. A., Pare, N., & Flashman, L. A. (2008). Verbal fluency performance in amnesic MCI and older adults with

cognitive complaints. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 23(3), 229–241.

O'Bryant, S. E., Lacritz, L. H., Hall, J., Waring, S. C., Chan, W., Khodr, Z. G., ... Cullum, C. M. (2010). Validation of the new interpretive guidelines for the clinical dementia rating scale sum of boxes score in the national Alzheimer's coordinating center database. *Archives of Neurology*, 67(6), 746–749.

O'Bryant, S. E., Waring, S. C., Cullum, C. M., Hall, J., Lacritz, L., Massman, P. J., ... Texas Alzheimer's Research Consortium. (2008). Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores: a Texas Alzheimer's research consortium study. *Archives of Neurology*, 65(8), 1091–1095.

Orimaye, S. O., Wong, J. S.-M., & Golden, K. J. (2014). Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 78–87.

Paavilainen, P., Korhonen, I., Cluitmans, L., Lötjönen, J., Särelä, A., & Partinen, M. (2003). Circadian Rhythm in Demented and Non-demented Nursing Home Residents Measured by the IST Vivago® WristCare Activity Signal. *Icadi*, 2–5.

Pakhomov, S. V. S., Eberly, L., & Knopman, D. (2016). Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. *Neuropsychologia*, 89, 42–56.

Pakhomov, S. V. S., & Hemmy, L. S. (2014). *A Computational Linguistic Measure of Clustering Behavior on Semantic Verbal Fluency Task Predicts Risk of Future Dementia in the Nun Study*. 55(24). doi:10.1002/cncr.27633.Percutaneous

- Pakhomov, S. V. S., Marino, S. E., Banks, S., & Bernick, C. (2015). Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency. *Speech Communication, 75*, 14–26.
- Papp, K. V., Mormino, E. C., Amariglio, R. E., Munro, C., Dagley, A., Schultz, A. P., ... Rentz, D. M. (2016). Biomarker validation of a decline in semantic processing in preclinical alzheimer's disease. *Neuropsychology, 30*(5), 624–630.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. Retrieved from <http://arxiv.org/abs/1201.0490>
- Pellegrini, E., Ballerini, L., Hernandez, M. D. C. V., Chappell, F. M., González-Castro, V., Anblagan, D., ... Wardlaw, J. M. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer's & Dementia (Amsterdam, Netherlands), 10*(1), 519–535.
- Peter, J., Kaiser, J., Landerer, V., Köstering, L., Kaller, C. P., Heimbach, B., ... Klöppel, S. (2016). Category and design fluency in mild cognitive impairment: Performance, strategy use, and neural correlates. *Neuropsychologia, 93*(September), 21–29.
- Petersen, R. C. (2004). Mild cognitive impairment as a clinical entity and treatment target. *Archives of Neurology, 62*(7), 1160–1163; discussion 1167.
- Petersen, R. C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V., & Fratiglioni, L. (2014). Mild cognitive impairment: a concept in evolution. *Journal of Internal Medicine, 275*(3), 214–228.

- Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*, *56*(3), 303–308.
- Price, S. E., Kinsella, G. J., Ong, B., Storey, E., Mullaly, E., Phillips, M., ... Perre, D. (2012). Semantic verbal fluency strategies in amnesic mild cognitive impairment. *Neuropsychology*, *26*(4), 490–497.
- Prince, M., Comas-Herrera, A., Knapp, M., Guerchet, M., & Karagiannidou, M. (2016). *World Alzheimer Report 2016 Improving healthcare for people living with dementia. Coverage, Quality and costs now and in the future* (pp. 1–140). Retrieved from <https://www.alz.co.uk/research/world-report-2016>
- Randolph, C., Braun, A. R., Goldberg, T. E., & Chase, T. N. (1993). Semantic Fluency in Alzheimer's, Parkinson's, and Huntington's Disease: Dissociation of Storage and Retrieval Failures. *Neuropsychology*, *7*(1), 82–88.
- Raoux, N., Amieva, H., Le Goff, M., Auriacombe, S., Carcaillon, L., Letenneur, L., & Dartigues, J. F. (2008). Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID longitudinal study. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *44*(9), 1188–1196.
- Rentz, D. M., Locascio, J. J., Becker, J. A., Moran, E. K., Eng, E., Buckner, R. L., ... Johnson, K. A. (2010). Cognition, reserve, and amyloid deposition in normal aging. *Annals of Neurology*, *67*(3), 353–364.
- Ritchie, C. W., Molinuevo, J. L., Truyen, L., Satlin, A., Van der Geyten, S., & Lovestone, S. (2016). Development of interventions for the secondary prevention of Alzheimer's dementia:

- The European Prevention of Alzheimer's Dementia (EPAD) project. *The Lancet Psychiatry*, 3(2), 179–186.
- Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., & Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 2081–2090.
- Robert, P. H., Lafont, V., Medecin, I., Berthet, L., Thauby, S., Baudu, C., & Darcourt, G. (1998). Clustering and switching strategies in verbal fluency tasks: comparison between schizophrenics and healthy adults. *Journal of the International Neuropsychological Society: JINS*, 4(6), 539–546.
- Robert, P. H., Migneco, V., Marmod, D., Chaix, I., Thauby, S., Benoit, M., ... Darcourt, G. (1997). Verbal fluency in schizophrenia: The role of semantic clustering in category instance generation. *European Psychiatry: The Journal of the Association of European Psychiatrists*, 12(3), 124–129.
- Roe, C. M., Fagan, A. M., Grant, E. A., Marcus, D. S., Benzinger, T. L. S., Mintun, M. A., ... Morris, J. C. (2011). Cerebrospinal fluid biomarkers, education, brain volume, and future cognition. *Archives of Neurology*, 68(9), 1145–1151.
- Rohrer, D., Wixted, J., Salmon, D. P., & Butters, N. (1995). Retrieval From Semantic Memory and Its Implications for Alzheimer's Disease. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 21(5), 1127–1139.
- Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A new rating scale for Alzheimer's disease. *The American Journal of Psychiatry*, 141(11), 1356–1364.
- Sadeghian, R., David Schaffer, J., & Zahorian, S. A. (2017). Speech processing approach for diagnosing dementia in an early stage. *Proceedings of the Annual Conference of the*

*International Speech Communication Association, INTERSPEECH, 2017-Augus, 2705–2709.* International Speech Communication Association.

Satt, A., Hoory, R., König, A., Aalten, P., & Robert, P. H. (2014). Speech-based automatic and robust detection of very early dementia. *Fifteenth Annual Conference of the International Speech Communication Association*. Retrieved from [https://www.isca-speech.org/archive/interspeech\\_2014/i14\\_2538.html](https://www.isca-speech.org/archive/interspeech_2014/i14_2538.html)

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology, 5*(JUL), 1–10.

Sindi, S., Mangialasche, F., & Kivipelto, M. (2015). Advances in the prevention of Alzheimer's disease. *F1000prime Reports, 7*, 50.

Small, B. J., Fratiglioni, L., Viitanen, M., Winblad, B., & Bäckman, L. (2000). The course of cognitive impairment in preclinical Alzheimer disease: Three- and 6-year follow-up of a population-based sample. *Archives of Neurology, 57*(6), 839–844.

Small, J., & Sandhu, N. (2006). Picture naming in Alzheimer's disease: The role of episodic memory. *Brain and Language, 99*(1–2), 134–135.

Snyder, P. J., Kahle-Wroblewski, K., Brannan, S., Miller, D. S., Schindler, R. J., DeSanti, S., ... Carrillo, M. C. (2014). Assessing cognition and function in Alzheimer's disease clinical trials: do we have the right tools? *Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 10*(6), 853–860.

Speer, R., Chin, J., Lin, A., Jewett, S., & Nathan, L. (2018). *Luminosinsight/wordfreq: v2. 2*. October.

- Sperling, R. A., Rentz, D. M., Johnson, K. A., Karlawish, J., Donohue, M., Salmon, D. P., & Aisen, P. (2014). The A4 study: Stopping AD before symptoms begin? *Science Translational Medicine*, *6*(228), 228fs13-228fs13.
- Sperling, Reisa A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., ... Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *7*(3), 280–292.
- St-Hilaire, A., Hudon, C., Vallet, G. T., Bherer, L., Lussier, M., Gagnon, J.-F., ... Macoir, J. (2016). Normative data for phonemic and semantic verbal fluency test in the adult French-Quebec population and validation study in Alzheimer's disease and depression. *The Clinical Neuropsychologist*, *30*(7), 1126–1150.
- Suchy, Y., Kraybill, M. L., & Franchow, E. (2011). Practice effect and beyond: reaction to novelty as an independent predictor of cognitive decline among older adults. *Journal of the International Neuropsychological Society: JINS*, *17*(1), 101–111.
- Suzuki, T., Murase, S., Tanaka, T., & Okazawa, T. (2007). New approach for the early detection of dementia by recording in-house activities. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, *13*(1), 41–44.
- Szatlóczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in Aging Neuroscience*, *7*(195), 1–7.

- Taler, V., & Phillips, N. a. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Neuropsychology*, *30*(5), 501–556.
- Tarnanas, I., Schlee, W., Tsolaki, M., Müri, R., Mosimann, U., & Nef, T. (2013). Ecological validity of virtual reality daily living activities screening for early dementia: Longitudinal study. *Journal of Medical Internet Research*, *15*(8). doi:10.2196/games.2778
- Teng, E., Leone-Friedman, J., Lee, G. J., Woo, S., Apostolova, L. G., Harrell, S., ... Lu, P. H. (2013). Similar Verbal Fluency Patterns in Amnesic Mild Cognitive Impairment and Alzheimer's Disease. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *28*(5), 400–410.
- Terrera, G. M., Harrison, J. E., Ritchie, C. W., & Ritchie, K. (2020). Cognitive Functions as Predictors of Alzheimer's Disease Biomarker Status in the European Prevention of Alzheimer's Dementia Cohort. *Journal of Alzheimer's Disease: JAD*, *74*(4), 1203–1210.
- Thorgusen, S. R., Suchy, Y., Chelune, G. J., & Baucom, B. R. (2016). Neuropsychological Practice Effects in the Context of Cognitive Decline: Contributions from Learning and Task Novelty. *Journal of the International Neuropsychological Society: JINS*, *22*(4), 453–466.
- Tomaszewski Farias, S., Cahn-Weiner, D. A., Harvey, D. J., Reed, B. R., Mungas, D., Kramer, J. H., & Chui, H. (2009). Longitudinal changes in memory and executive functioning are associated with longitudinal change in instrumental activities of daily living in older adults. *The Clinical Neuropsychologist*, *23*(3), 446–461.
- Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical*

*Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*,  
14(2), 167–177.

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., ... Kálmán, J. (2015). Automatic detection of mild cognitive impairment from spontaneous speech using ASR. *Interspeech*, 1–5.

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Banreti, Z., ... Kalman, J. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2), 130–138.

Tröger, J., Linz, N., Alexandersson, J., König, A., & Robert, P. (2017, May 23). Automated speech-based screening for alzheimer's disease in a care service scenario. *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 292–297. Presented at the Barcelona, Spain. New York, NY, USA: Association for Computing Machinery.

Tröger, J., Linz, N., König, A., Robert, P., Alexandersson, J., Peter, J., & Kray, J. (2019). Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in Alzheimer's disease. *Neuropsychologia*. doi:10.1016/j.neuropsychologia.2019.05.007

Tröster, A. I., Salmon, D. P., McCullough, D., & Butters, N. (1989). A comparison of the category fluency deficits associated with Alzheimer's and Huntington's disease. *Brain and Language*, 37(3), 500–513.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). *Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults* (pp. 138–146). pp. 138–146.

- Troyer, A. K., Moscovitch, M., Winocur, G., Alexander, M. P., & Stuss, D. (1998). Clustering and switching on verbal fluency: The effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia*, *36*(6), 499–504.
- Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., & Freedman, M. (1998). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society: JINS*, *4*(2), 137–143.
- Tulving, E. (1993). What is episodic memory? *Current Directions in Psychological Science*, *2*(3), 67–70.
- Unsworth, N., Spillers, G. J., & Brewer, G. A. (2011). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *The Quarterly Journal of Experimental Psychology*, *64*(3), 447–466.
- Van Den Berg, E., Jiskoot, L. C., Grosveld, M. J. H., Van Swieten, J. C., & Papma, J. M. (2017). Qualitative Assessment of Verbal Fluency Performance in Frontotemporal Dementia. *Dementia and Geriatric Cognitive Disorders*, *44*(1–2), 35–44.
- Venneri, A., Mitolo, M., & De Marco, M. (2016). Paradigm shift: semantic memory decline as a biomarker of preclinical Alzheimer's disease. *Biomarkers in Medicine*, *10*(1), 5–8.
- Verma, M., & Howard, R. J. (2012). Semantic memory and language dysfunction in early Alzheimer's disease: A review. *International Journal of Geriatric Psychiatry*, *27*(12), 1209–1217.
- Vestal, L., Smith-Olinde, L., Hicks, G., Hutton, T., & Hart, J. (2006). Efficacy of language assessment in Alzheimer's disease: comparing in-person examination and telemedicine. *Clinical Interventions in Aging*, *1*(4), 467–471.

- Vlassenko, A. G., Benzinger, T. L. S., & Morris, J. C. (2012). PET amyloid-beta imaging in preclinical Alzheimer's disease. *Biochimica et Biophysica Acta*, 1822(3), 370–379.
- Watanabe, S., Delcroix, M., Metze, F., & Hershey, J. R. (2017). *New Era for Robust Speech Recognition: Exploiting Deep Learning* (Shinji Watanabe, M. Delcroix, F. Metze, & J. R. Hershey, Eds.). Springer, Cham.
- Weakley, A., & Schmitter-Edgecombe, M. (2014). Analysis of verbal fluency ability in Alzheimer's disease: The Role of clustering, switching and semantic proximities. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 29(3), 256–268.
- Weiner, J., Engelbart, M., & Schultz, T. (2017). Manual and Automatic Transcriptions in Dementia Detection from Speech. *Interspeech2017*, 3117–3121.
- Whiteside, D. M., Kealey, T., Semla, M., Luu, H., Rice, L., Basso, M. R., & Roper, B. (2016). Verbal Fluency: Language or Executive Function Measure? *Applied Neuropsychology:Adult*, 23(1), 29–34.
- Wild, K., Howieson, D., Webbe, F., Seelye, A., & Kaye, J. (2008, November). Status of computerized cognitive testing in aging: A systematic review. *Alzheimer's and Dementia*, Vol. 4, pp. 428–437. doi:10.1016/j.jalz.2008.07.003
- Wimo, A., Guerchet, M., Ali, G. C., Wu, Y. T., Prina, A. M., Winblad, B., ... Prince, M. (2017). The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's and Dementia*, 13(1), 1–7.
- Winter, M. H.-J., Maaz, A., & Kuhlmeiy, A. (2006). Ambulante und stationäre medizinische Versorgung im Alter. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, 49(6), 575–582.

- Woods, D. L., Wyma, J. M., Herron, T. J., & Yund, E. W. (2016). Computerized Analysis of Verbal Fluency: Normative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury. *PloS One*, *11*(12), e0166439.
- World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization.
- Yee, E., Jones, M. N., & McRae, K. (2018). Semantic Memory. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–38). doi:10.1002/9781119170174.epcn309
- Yesavage, J. A., & Sheikh, J. I. (1986). Geriatric Depression Scale (GDS). *Clinical Gerontologist*, *5*(1–2), 165–173.
- Yu, B., Quatieri, T. F., Williamson, J. R., & Mundt, J. C. (2015). Cognitive impairment prediction in the elderly based on vocal biomarkers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3734–3738.
- Yu, D., & Deng, L. (2015). *Automatic Speech Recognition: A Deep Learning Approach*. Springer, London.
- Zhou, L., Fraser, K. C., & Rudzicz, F. (2016). Speech recognition in Alzheimer's disease and in its assessment. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 08-12-Sept, 1948–1952*. International Speech and Communication Association.

## CURRICULUM VITAE

Johannes Tröger (he/him/his) born 21<sup>st</sup> of August 1988 in Nürnberg, Germany, is a researcher in the department of cognitive assistants at the German Research Center for Artificial Intelligence (DFKI) in Saarbrücken. Since 2016 he has been working on medical applications of AI within the scope of various national and international projects. The focus is always on user experience and the added value of AI in real world applications. Johannes Tröger studied Educational Technology (M.Sc.) and Psychology (B.Sc.) at the University of Saarland. For the last three years he has been working on the application of AI research, i.e. methods of speech analysis, for diagnosis support of neurocognitive disorders: “We are working together with clinical partners across Europe to develop the next generation of objective speech-based bio markers: To detect psychiatric conditions such as Alzheimer's or Depression early on, population-wide and model them longitudinally as well as on an individual basis”.

**Selected Publications:****2021**

- Lindsay, H., & **Tröger, J.** (2021). Language Impairment in Alzheimer's Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech through Multilingual Machine Learning. *Frontiers in Aging Neuroscience*, *13*, 228.
- **Tröger, J.**, Lindsay, H., Mina, M., Linz, N., Klöppel, S., Kray, J., & Peter, J. (2021). Patients with amnesic MCI fail to adapt executive control when repeatedly tested with semantic verbal fluency tasks. *Journal of the International Neuropsychological Society*, accepted.
- Lindsay, H., Mueller, P., Linz, N., Mina, M., Zaghari, R., König, A., **Tröger, J.** (2021). Dissociating Semantic and Phonemic Search Strategies in the Phonemic Verbal Fluency Task in early Dementia. *In The Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, accepted.
- Baykara, E., Kuhn, C., Linz, N., **Tröger, J.**, Karch, J. (2021) Validation of a digital, tablet-based version of the Trail-Making-Test in the Delta platform. *Journal of the International Neuropsychological Society*, submitted.

**2020**

- Lindsay, H., **Tröger, J.**, Linz, N., Alexandersson, J., & König, A. (2020). What Difference Does it Make? Early Dementia Detection Using the Semantic and Phonemic Verbal Fluency Task. *RaPID-3, 2020*, (in press).

**2019**

- Lindsay, H., **Tröger, J.**, Linz, N., Alexandersson, J., & Prudlo, J. (2019). Automatic Detection of Language Impairment in Amyotrophic Lateral Sclerosis. *ExLing 2019*, 25, 133.
- Lindsay, H., Linz, N., **Tröger, J.**, & Alexandersson, J. (2019). Automatic Data-Driven Approaches for Evaluating the Phonemic Verbal Fluency Task with Healthy Adults. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing* (pp. 17-24).
- König, A., **Tröger, J.**, Linz, N., Zeghari, R., & Robert, P. (2019). Vers une digitalisation des tests cognitifs dans la pratique clinique grâce à l'aide de l'intelligence artificielle: l'application «Delta». *French Journal of Psychiatry*, 1, S50.
- König, A., Linz, N., Zeghari, R., Klinge, X., **Tröger, J.**, Alexandersson, J., & Robert, P. (2019). Detecting apathy in older adults with cognitive disorders using automatic speech analysis. *Journal of Alzheimer's Disease*, 69(4), 1183-1193.

## 2018

- **Tröger, J.**, Linz, N., König, A., Robert, P., & Alexandersson, J. (2018, May). Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare* (pp. 59-66). ACM.
- **Tröger, J.**, Linz, N., König, A., Robert, P., Alexandersson, J., Peter, J., & Kray, J. (2019). Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in Alzheimer's disease. *Neuropsychologia*, 131, 53-61.
- König, A., Linz, N., **Tröger, J.**, Wolters, M., Alexandersson, J., & Robert, P. (2018). Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task. *Dementia and Geriatric Cognitive Disorders*, 45(3-4), 198-210.
- Linz, N., Klinge, X., **Tröger, J.**, Alexandersson, J., Zeghari, R., Philippe, R., & König, A. (2018, July). Automatic detection of apathy using acoustic markers extracted from free emotional speech.
- Linz, N., **Tröger, J.**, Lindsay, H., König, A., Robert, P., Peter, J., & Alexandersson, J. (2018, May). Language modelling for the clinical semantic verbal fluency task. In *LREC 2018 Workshop RaPID-2: Resources and Processing of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments*.
- Linz, N., Klinge, X., **Tröger, J.**, Alexandersson, J., Zeghari, R., Philippe, R., & König, A. (2018, July). Automatic detection of apathy using acoustic markers extracted from free emotional speech. In *2ND WORKSHOP ON AI FOR AGING, REHABILITATION AND INDEPENDENT ASSISTED LIVING (ARIAL)@ IJCAI'18*.

## 2017

- **Tröger, J.**, Linz, N., Alexandersson, J., König, A., & Robert, P. (2017, May). Automated speech-based screening for alzheimer's disease in a care service scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (pp. 292-297). ACM.
- Linz, N., **Tröger, J.**, Alexandersson, J., Wolters, M., König, A., & Robert, P. (2017, November). Predicting dementia screening and staging scores from semantic verbal

fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 719-728). IEEE.

- Rohr, M., **Tröger, J.**, Michely, N., Uhde, A., & Wentura, D. (2017). Recognition memory for low-and high-frequency-filtered emotional faces: Low spatial frequencies drive emotional memory enhancement, whereas high spatial frequencies drive the emotion-induced recognition bias. *Memory & cognition*, 45(5), 699-715.
- Linz, N., **Tröger, J.**, Alexandersson, J., & König, A. (2017, September). Using neural word embeddings in the analysis of the clinical semantic verbal fluency task. In *IWCS 2017-12th International Conference on Computational Semantics* (pp. 1-7).

## **2014**

- Dörrenbächer, S., Müller, P. M., **Tröger, J.**, & Kray, J. (2014). Dissociable effects of game elements on motivation and cognition in a task-switching training in middle childhood. *Frontiers in psychology*, 5, 1275.