Claudia Frankenberg*, Jochen Weiner, Tanja Schultz, Maren Knebel, Christina Degen, Hans-W. Wahl and Johannes Schroeder

Perplexity – a new predictor of cognitive changes in spoken language? – results of the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE)

https://doi.org/10.1515/lingvan-2018-0026
Received May 15, 2018; accepted January 7, 2019

Abstract: In addition to memory loss, progressive deterioration of speech and language skills is among the main symptoms at the onset of Alzheimer's disease (AD) as well as in mild cognitive impairment (MCI). Detailed interview analyses demonstrated early symptoms years before the onset of AD/MCI. Automatic speech processing could be a promising approach to identifying underlying mechanisms in larger studies or even support diagnostics. Perplexity as a measure of predictability of text could be a sensitive indicator of cognitive deterioration. Therefore, voice recordings from the Interdisciplinary Longitudinal Study on Adult Development and Aging were analyzed with regard to neuropsychological parameters in participants that develop MCI/AD or remain cognitively healthy. Preliminary results indicate that perplexity predicts severity of cognitive deficits and information processing speed obtained 10–12 years later in participants who developed MCI/AD in contrast to those who stayed healthy. Findings support the heuristic value of research on the diagnostic potential of automatic speech processing.

Keywords: language and aging; speech processing; Alzheimer's disease; mild cognitive impairment.

1 Introduction

Linguistic changes in early Alzheimer's disease (AD) can be observed on different linguistic levels: On the linguistic surface reduced word fluency, especially category fluency, and prominent word finding deficits were frequently described (Lukatela et al. 1998; Barth et al. 2005; Dos Santos et al. 2011; Schröder and Pantel 2011). In addition, a reduced lexical diversity can be determined, which may arise from a higher rate of immediate repetition of nouns and verbs (De Lira et al. 2011) or alternatively, from a disproportional frequency of miscellaneous parts of spoken language as in lower rates of nouns and higher rates of verbs, adverbs and adjectives (Blanken et al. 1987). Linguistic changes also extend to mild cognitive impairment (MCI) i.e. the preclinical state of AD. Mild cognitive impairment (MCI) is associated with an increased risk of developing dementia (Schröder and Pantel 2011) and is characterized by neuropsychological deficits exceeding those losses which typically develop during physiological aging but (still) do not compare with the more severe deficits characteristic of early AD. Additional changes in MCI through to later stages of the disease comprise reduced syntactic complexity and impairments in semantic content (Ahmed et al. 2013). These changes lead to a simplification of spoken language with progression of the disease. However, in the German-speaking countries instruments to detect changes in the content of spoken language in natural situations, which are suitable for patients with AD or even MCI, are rare (Knebel et al. 2015). Therefore, we sought to examine perplexity as a potential

***Corresponding author: Claudia Frankenberg,** Heidelberg University, Section of Geriatric Psychiatry, Heidelberg, Germany, E-mail: Claudia.Frankenberg@med.uni-heidelberg.de
Jochen Weiner and Tanja Schultz: Bremen University, Cognitive Systems Lab, Bremen, Germany
Maren Knebel: Goethe Univerisity Frankfurt, Frankfurt Forum for Interdisciplinary Ageing Research, Frankfurt, Germany; and Heidelberg University, Section of Geriatric Psychiatry, Heidelberg, Germany
Christina Degen and Johannes Schroeder: Heidelberg University, Section of Geriatric Psychiatry, Heidelberg, Germany
Hans-W. Wahl: Heidelberg University, Institute of Psychology, Network Aging Research, Heidelberg, Germany

measure for determining content complexity in spoken language in physiological aging and at the onset of MCI or AD on basis of the interviews taken in the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE).

1.1 The Interdisciplinary Longitudinal Study on Adult Development and Aging

ILSE is a population-based follow-up study to investigate individual, societal, and socio-structural preconditions for mental and physical aging (Martin and Martin 2000). One thousand two subjects from two birth cohorts of 1930–1932 and 1950–1952 were recruited from the Heidelberg and Leipzig regions (Schröder et al. 1998; Schönknecht et al. 2005) on the basis of local community registers (which are compulsory in Germany). Four examination waves with an observation interval of more than 20 years were completed (1993–1996: t1, 1997–1999: t2, 2005–2007: t3, 2013–2016: t4). Each included expert geriatric, psychiatric and psychological assessments and semi-standardized biographical interviews which reached approximately 10,000 hours of interviews to date.

On the basis of a selection out of 145 manually transcribed interviews, Wendelstein (2016) demonstrated a tendency towards lower lexical richness (shown by diverging type-token-relations in group comparisons and Brunét-Index changes especially in the preclinical stage and in the beginning of AD; Brunét 1978) as well as an overproduction of pronouns and incomplete syntactic phrases in participants with preclinical AD/AD compared with healthy controls. Furthermore, the spoken language of subjects in the preclinical stage was characterized by a lower propositional content than the spoken language of the healthy controls (Wendelstein 2016).

Methodological problems prevented comprehensive analyses of the interviews, as manual transcription required at least eight times the duration of each interview. Hence, we sought to develop an automatic speech recognition (ASR) system to transcribe the interviews for linguistic analysis. The ASR is part of a fully automatic speech and language processing pipeline (Weiner et al. 2016b, Weiner et al. 2017) which is developed to provide information for individual diagnostics. For this purpose, both acoustic and linguistic features were extracted from the interview recordings and their transcriptions. A classifier was trained on these features. Based on these the classifier would indicate the diagnosis of the speakers. The results appear to be very promising for both acoustic and linguistic features including a combination of these (Weiner et al. 2016b, Weiner et al. 2017). One of the promising concepts that originally stems from ASR and could also be used for the observation of changes in spontaneous speech is perplexity (Weiner et al. 2017). Perplexity is a measurement introduced from information theory to quantify how well a model predicts a sample. Projected to the present study, it indicates how well an utterance spoken by an ILSE participant can be predicted by a language model.

Based on these findings, the aim of the present study is to critically examine perplexity as an additional (linguistic) marker for speech changes and cognitive deficits. Since propositional density as an aspect of content complexity is known to change up to 12 years before the diagnoses of AD (Wendelstein 2016), the main focus is on the potential predictive value of perplexity. Associations of perplexity are examined with neuropsychological performances in a healthy control group (HC) and MCI/AD, which may add to its potential relevance for diagnostic processes.

2 Methods

2.1 Sample

The ILSE, with its four examination waves, including semi-standardized biographical interviews, which comprised three different parts, allows a focus on longitudinal analyses on speech. In the interviews, open narrative-generating questions were followed by an explicit questioning for defined events and circumstance of life such as remembrance of elementary school, the entrance into professional life, the moving out of the children, or transition to retirement. The third part of the interview considered future demands with prospective questions about general desires, fears, ideas, and plans for the future. The duration of

the interviews amounted to 2.5–6 hours at the first examination wave and subsequently between 0.5 and 2.5 hours at the following waves. The interviews of the first and second examination waves were recorded with analogue recording devices on tape. For the third and fourth examination wave the interviews were recorded digitally in mp3 and later in PCM format.

The subsample used in the present study consists of randomly chosen participants from the ILSE who fulfilled the inclusion criteria. To avoid bias, only participants from the older birth cohort (born between 1930 and 1932) were included. Subjects had to be diagnosed with either MCI (according to the criteria of Aging-associated cognitive decline; Levy 1994; Schönknecht et al. 2005) or early AD (according to ICD 10 and NINCDS/ADRDA criteria; McKhann et al. 1984) or had to be HC. Cognitive diagnoses were established in diagnostic conferences under the direction of an experienced geriatric psychiatrist as described previously (Toro et al. 2014). Patients whose cognitive deficits were attributable to primary physical disorders, e.g. tumors or cardiovascular disease (mild cognitive disorder, ICD-10: F06.7), were excluded (n = 2). Accordingly, a sample of 51 participants was available. Of these, 48 participants were cognitively healthy, while three participants were diagnosed with MCI at t1. For the present study, data from HC (n = 31) and from participants who had developed MCI (n = 15) and AD (n = 5) at t3 were included. As the latter group comprised five subjects only, participants with MCI and AD were merged into a cognitively impaired group (MCI/AD) following previous studies of our group (Degen et al. 2016).

2.2 Neuropsychological assessment

A detailed neuropsychological examination with tests on memory performance and learning, attention and concentration, processing speed, language, visuospatial functions and abstract thinking was conducted at each examination wave. The following language-related neuropsychological tests are considered sensitive and useful in the diagnosis of dementia, and were therefore included to investigate the relationship between perplexity as a computer-linguistic measure and cognitive abilities relevant in aging and dementia (t3): the word fluency task *word finding* (Leistungsprüfsystem, Horn 1983; also used at t1) and from the Consortium to Establish a Registry for Alzheimer's Disease (CERAD; Morris et al. 1989; Welsh et al. 1994; Aebi 2002), the 15-item version of the *Boston-Naming Test* (CERAD, Morris et al. 1989; BNT, Kaplan et al. 1983 Engl. orig.). As a less language-oriented task, the *Trail Making Test* (TMT; Reitan 1992) was used to depict information processing speed and executive functioning (scores indicate time needed for the task, therefore higher scores indicate lower performance in the TMT and lower scores indicate better performance). Memory was assessed by the subtest *logical memory* (Wechsler Memory Scale, German version by Petermann and Lepach 2012) and by the recall of a word list, that was memorized beforehand (CERAD; also at t1). Additionally, the *Mini Mental State Examination* (MMSE; Folstein et al. 1975) as one of the most widely known dementia screening instruments (from the German version of the CERAD; Morris et al. 1989; Welsh et al. 1994; Aebi 2002) was included. Besides the TMT, higher scores indicate better performance in the described neuropsychological tests.

2.3 Perplexity

Modern applications like automatic speech recognition and machine translation are based on statistical modeling. In these applications, language models are used to model the probability of word sequences. Models are trained on a corpus of texts and their performance is measured by calculating their perplexity on a test text. Perplexity measures how well the model predicts the data. The lower the perplexity the better the match between the model and the test text, i.e. the better the predictability of the test text. Perplexity is a concept from information theory closely related to entropy (Jelinek 1985). The perplexity of a model given a text is defined as

$$perplexity = 10^{\frac{-logprob}{words-OOV}}$$

where *logprob* is the logarithm of the probability that the model assigns to the text. The negative logprob describes the number of bits needed to encode the test text using an optimal code based on the language

model. The variable *words* is the number of words in the text and *OOV* (out of vocabulary) is the number of words in the text unknown to the model. Thus, the logprob is normalized for text length and the exponent of the equation describes the average number of bits per word needed to encode the test text under the use of the language model. The better a model is able to predict the test text, the fewer bits it will require to encode that text and the resulting perplexity will be lower. Intuitively, perplexity can be interpreted as the average number of possible continuations of a sentence that the model would deem likely.

The perplexity scores were calculated using the following procedure: The transcripts of a person's speech were split into ten parts of equal length. Then a model was built using nine parts and was evaluated on the tenth. This was repeated ten times so that each part was used exactly once as the evaluation text. Then the arithmetic mean of the ten perplexities was computed as the perplexity feature. SRILM (Stolcke 2002) was used to train and evaluate the models. The resulting score measures how well, on average, a model which is trained on 90 percent of a person's speech can predict the remaining 10 percent of the person's speech. A lower score indicates a better predictability, i.e. for persons with very little variability in their speech the perplexity feature will have a very low value.

In the present analysis, perplexities of 1-gram and 2-gram language models are investigated. These two models differ in the context that they consider when modeling the probability of a text: While the 1-gram language model takes into account one word at a time, the 2-gram language model considers each word in the context of its predecessor. In the calculation of the perplexity, this difference is contained in the calculation of the probability that the model assigns to the text (logprob).

2.4 Statistical analyses

To examine the association between neuropsychological test performance and perplexity scores Pearson's correlation coefficients were calculated for each group separately using 1-gram and 2-gram perplexity scores. The significance level was set at 95 percent, therefore values of $p < 0.05$ were considered significant and tests were two-sided. Additionally, determination coefficients (squared R) and confidence intervals after Fisher's z-transformation of significant perplexity-neuropsychology at t3 correlations are reported, to allow comparison between the two groups. For descriptive analyses, chi-square and independent t-tests were used. All analyses were calculated using Microsoft Excel 2011 or IBM SPSS 24.

3 Results

As demonstrated in Table 1, diagnostic groups showed only minor, non-significant differences with respect to sex, age and years of education. As was expected, the patients group presented significantly lower MMSE scores than the HC.

Table 2 shows the correlations between neuropsychological test scores and perplexity scores for HC. While perplexity scores were significantly intercorrelated, only minor, non-significant correlations between 1-gram and 2-gram perplexity scores obtained at t1 and t3 (n's $>= 8$) and neuropsychological performance

Table 1: Sample Description.

	Control group (n = 31)	MCI/AD (n = 20)	X^2 / t (df)	p[a]
Sex (male/female)	16/15	15/5	X^2 (1) = 0.095	n. sign.
Age in years (t1) M (SD)	62.81 (0.94)	63.00 (0.97)	t (49) = −0.70	n. sign.
Education in years M (SD)	13.81 (3.07)	13.00 (2.51)	t (49) = 0.98	n. sign.
MMSE (t3) M (SD)	28.93 (1.17)	27.38 (1.93)	t (44) = 2.96	p = 0.008
min. – max. values	26–30	24–30		

SD: standard deviation, M: Mean, df: degrees of freedom, X^2: chi-square, min.: minimum, max.: maximum; n: sample, AD: early Alzheimer's disease, MCI: Mild cognitive impairment, MMSE: Mini Mental State Examination, t1: first examination wave, 1993–1996; t3: third examination wave, 2005–2007; [a]alpha set at 0.05.

Table 2: Pearson's correlation of perplexity and neuropsychological measures at t1 and t3 in HC.

	t1 perplexity 1	t1 perplexity 2	t3 perplexity 1	t3 perplexity 2	t1 free reprod. word list	t1 verbal fluency	t3 free reprod. word list	t3 Boston Naming Test	t3 verbal fluency	t3 lm imm. recall	t3 lm delayed	t3 MMSE	t3 TMT A	t3 TMT B
t1 perplexity 1	–													
t1 perplexity 2	0.94** (n = 28)	–												
t3 perplexity 1	0.86** (n = 8)	−0.75* (n = 8)	–											
t3 perplexity 2	0.88** (n = 8)	0.38* (n = 8)	0.98** (n = 8)	–										
t1 free reprod. word list	−0.13 (n = 28)	−0.06 (n = 28)	0.30 (n = 8)	0.34 (n = 8)	–									
t1 verbal fluency	−0.33 (n = 28)	−0.32 (n = 28)	0.08 (n = 8)	0.06 (n = 8)	0.30 (n = 31)	–								
t3 free reprod. word list	0.03 (n = 26)	0.06 (n = 26)	−0.29 (n = 8)	−0.26 (n = 8)	0.04 (n = 28)	0.01 (n = 28)	–							
t3 Boston Naming Test	−0.10 (n = 26)	0.02 (n = 26)	0.54 (n = 8)	0.65 (n = 8)	−0.04 (n = 28)	0.30 (n = 28)	0.23 (n = 28)	–						
t3 verbal fluency	−0.27 (n = 26)	−0.20 (n = 26)	−0.53 (n = 8)	−0.53 (n = 8)	0.34 (n = 28)	0.62** (n = 28)	0.45* (n = 28)	0.25 (n = 28)	–					
t3 lm imm. recall	−0.18 (n = 27)	−0.08 (n = 27)	0.12 (n = 8)	0.09 (n = 8)	−0.00 (n = 30)	0.30 (n = 30)	0.20 (n = 28)	0.14 (n = 28)	0.06 (n = 28)	–				
t3 lm delayed recall	−0.12 (n = 27)	−0.12 (n = 27)	0.25 (n = 8)	0.16 (n = 8)	−0.04 (n = 30)	0.12 (n = 30)	0.31 (n = 28)	0.08 (n = 28)	−0.05 (n = 28)	0.72** (n = 30)	–			
t3 MMSE	0.16 (n = 27)	0.23 (n = 27)	0.46 (n = 8)	0.50 (n = 8)	0.14 (n = 30)	−0.06 (n = 30)	0.27 (n = 28)	−0.01 (n = 28)	0.02 (n = 28)	0.03 (n = 30)	0.02 (n = 30)	–		
t3 TMT A	−0.07 (n = 26)	−0.05 (n = 26)	0.33 (n = 8)	0.28 (n = 8)	0.09 (n = 28)	0.44* (n = 28)	−0.12 (n = 28)	0.07 (n = 28)	0.06 (n = 28)	−0.34 (n = 28)	−0.36 (n = 28)	0.01 (n = 28)	–	
t3 TMT B	−0.08 (n = 26)	−0.04 (n = 26)	−0.38 (n = 8)	−0.48 (n = 8)	−0.26 (n = 28)	−0.55** (n = 28)	−0.01 (n = 28)	−0.29 (n = 28)	−0.37 (n = 28)	−0.19 (n = 28)	−0.02 (n = 28)	−0.05 (n = 28)	−0.08 (n = 28)	–

imm. = immediate; lm = logical memory; MMSE = Mini Mental State Examination; perplexity 1 = perplexity (1-gram); perplexity 2 = perplexity (2-gram); reprod. = reproduction; t1= first examination wave (1993–1996); t3 = third examination wave (2005–2007); TMT = Trail Making Test. *$p \leq 0.05$; **$p \leq 0.01$.

Table 3: Pearson's correlation of perplexity and neuropsychological measures at t1 and t3 in MCI/AD.

	t1 perplexity 1	t1 perplexity 2	t3 perplexity 1	t3 perplexity 2	t1 free reprod. word list	t1 verbal fluency	t3 free reprod. word list	t3 Boston Naming Test	t3 verbal fluency	t3 lm imm. recall	t3 lm delayed	t3 MMSE	t3 TMT A	t3 TMT B
t1 perplexity 1	–													
t1 perplexity 2	0.95** (n = 20)	–												
t3 perplexity 1	0.58 (n = 6)	−0.26 (n = 6)	–											
t3 perplexity 2	0.15 (n = 6)	−0.15 (n = 6)	0.68 (n = 6)	–										
t1 free reprod. word list	0.24 (n = 20)	0.27 (n = 20)	−0.06 (n = 6)	0.00 (n = 6)	–									
t1 verbal fluency	0.13 (n = 20)	0.18 (n = 20)	−0.76 (n = 6)	−0.21 (n = 6)	0.23 (n = 20)	–								
t3 free reprod. word list	0.22 (n = 15)	0.17 (n = 15)	0.10 (n = 6)	0.46 (n = 6)	0.45 (n = 15)	0.13 (n = 15)	–							
t3 Boston Naming Test	0.40 (n = 15)	0.47 (n = 15)	−0.37 (n = 6)	0.27 (n = 6)	0.44 (n = 15)	0.12 (n = 15)	0.23 (n = 15)	–						
t3 verbal fluency	0.28 (n = 15)	0.36 (n = 15)	−0.61 (n = 6)	−0.42 (n = 6)	0.69** (n = 15)	0.45 (n = 15)	0.13 (n = 15)	0.54* (n = 15)	–					
t3 lm imm. recall	0.16 (n = 16)	0.05 (n = 16)	0.72 (n = 6)	0.39 (n = 6)	0.29 (n = 16)	0.04 (n = 16)	0.33 (n = 15)	0.29 (n = 15)	−0.04 (n = 15)	–				
t3 lm delayed recall	0.44 (n = 16)	0.32 (n = 16)	0.70 (n = 6)	0.28 (n = 6)	0.47 (n = 16)	0.26 (n = 16)	0.24 (n = 15)	0.39 (n = 15)	0.15 (n = 15)	0.85** (n = 16)	–			
t3 MMSE	0.58* (n = 16)	0.48 (n = 16)	−0.03 (n = 6)	0.09 (n = 6)	0.75** (n = 16)	0.42 (n = 16)	0.46 (n = 15)	0.46 (n = 15)	0.59* (n = 15)	0.38 (n = 16)	0.59* (n = 16)	–		
t3 TMT A	−0.55* (n = 15)	−0.53* (n = 15)	0.03 (n = 6)	−0.45 (n = 6)	−0.71** (n = 15)	−0.44 (n = 15)	−0.58* (n = 15)	−0.68** (n = 15)	−0.60* (n = 15)	−0.26 (n = 15)	−0.50 (n = 15)	−0.80** (n = 15)	–	
t3 TMT B	−0.36 (n = 15)	−0.32 (n = 15)	0.04 (n = 6)	−0.46 (n = 6)	−0.44 (n = 15)	−0.21 (n = 15)	−0.61* (n = 15)	−0.56* (n = 15)	−0.42 (n = 15)	−0.22 (n = 15)	−0.36 (n = 15)	−0.62* (n = 15)	0.82** (n = 15)	–

imm. = immediate; lm = logical memory; MMSE = Mini Mental State Examination; perplexity 1 = perplexity (1-gram); perplexity 2 = perplexity (2-gram); reprod. = reproduction; t1= first examination wave (1993–1996); t3 = third examination wave (2005–2007); TMT = Trail Making Test. *$p \leq 0.05$; **$p \leq 0.01$.

arose. Neuropsychological parameters show significant associations as follows: verbal fluency at t3 is associated with verbal fluency at t1 and with the free recall of a word list. Additionally, immediate and delayed recall of two stories are highly associated (t3). Furthermore, t1 verbal fluency correlated significantly with measures of information processing speed (TMT A) and executive functions (TMT B) (Table 2).

In contrast, perplexity scores obtained at t1 in the MCI/AD patients (Table 3) were significantly correlated with neuropsychological performance at t3. 1-gram perplexity (t1) was significantly associated with t3 MMSE scores ($r = 0.58, p < 0.05$) and with t3 TMT A performance ($r = -0.55, p < 0.05$); 2-gram perplexity (t1) with t3 TMT A performance ($r = -0.53, p < 0.05$). No significant intercorrelations between perplexity measured at t1 and t3 were observed.

Furthermore, for participants with MCI/AD, t3 neuropsychological tests correlate with the reproduction of a word list at t1 as follows: The reproduction of a word list at t1 is associated with verbal fluency ($r = 0.69$, $p < 0.01$) and with the MMSE score ($r = 0.75, p < 0.01$) as well as with the TMT A ($r = -0.71, p < 0.01$; t3). Further associations concerning neuropsychology at t3 are shown in Table 3.

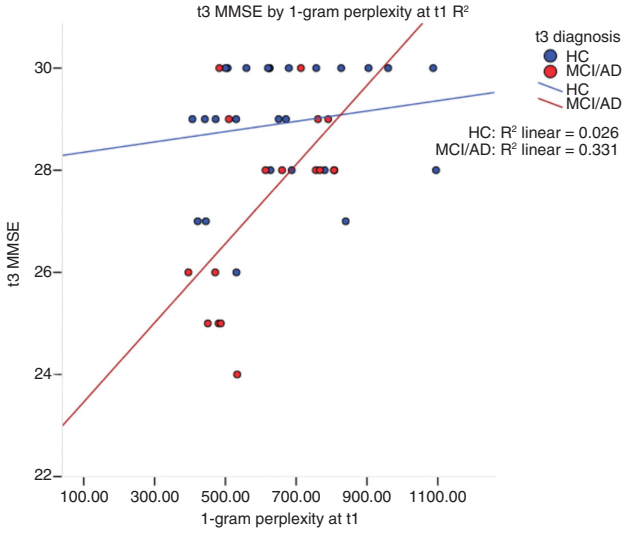

Figure 1: Corrected R^2 for 1-gram perplexity t1 and MMSE t3 correlations. Cave: x-axis cuts y-axis at MMSE-score 22 in this figure as MMSE in the sample ranges 24–30.

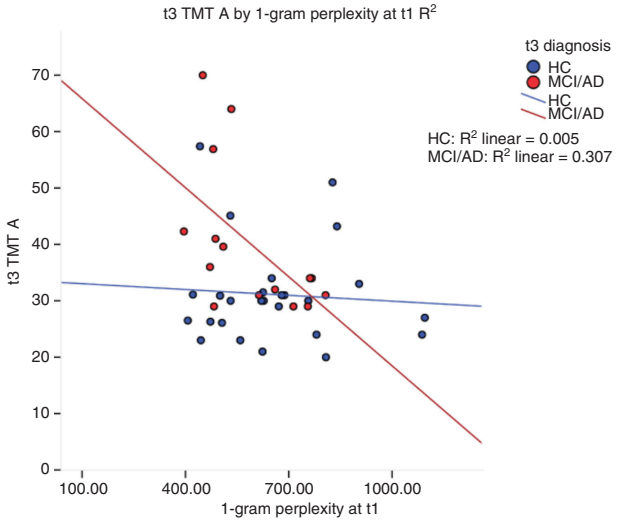

Figure 2: Corrected R^2 for 1-gram perplexity t1 and TMT A t3 correlations.

Additionally, corrected R^2 for 1-gram perplexity at t1 has been calculated and is presented in Figures 1 and 2 for the significant correlations contrasting MCI/AD and HC. For comparison, correlations were Fisher's z-transformed and confidence intervals reported: 1-gram perplexity t1 correlations for MCI/AD: $r_{MMSE} = 0.58$, CI [0.12, 1.21] vs. HC: $r_{MMSE} = 0.16$, CI [−0.24, 0.56]; MCI/AD: $r_{TMT\,A} = -0.55$, CI [−1.18, −0.05] vs. HC: $r_{TMT\,A} = -0.07$, CI [−0.48, 0.34]. While corrected R^2 is 0.331 for 1-gram perplexity at t1 and MMSE at t3 in MCI/AD, the corrected R^2 for HC is 0.026. Figure 2 shows R^2 for perplexity and the information processing speed (TMT A); which was corrected R^2 for MCI/AD is 0.307 and for HC 0.005.

4 Discussion

According to the findings of our study, perplexity measures can be derived from ASR on the basis of interviews in spoken language. We demonstrated that ASR and perplexity scores can be used in patient groups with MCI/AD. Perplexity measures at t1 are associated with cognitive decline in those prone to develop MCI or AD which did not apply for the HC.

ASR is used on a daily basis e.g. in mobile communication and in dictation software. At the same time, its use for the transcription of interview data is still in process and extremely difficult considering dialect or diverging quality of recordings (Weiner et al. 2016a). The model used for the present work was successfully used with patient groups (for more detail see Weiner et al. 2017).

The present study yielded two important findings regarding the potential of perplexity as a predictor of cognitive deterioration: (i) While in HC perplexity at t1 was not associated with neuropsychological measures at t3, correlations for perplexity with the score from the dementia screening instrument (MMSE) were significant in participants with MCI/AD. (ii) As would be expected, memory at t1 was associated with the degree of cognitive deterioration at t3 but perplexity showed associations, too. At the same time, perplexity itself was significantly correlated with information processing speed (TMT A) but not with memory scores in MCI/AD.

The findings support the idea of perplexity as a potential predictor of cognitive deterioration about a decade later. Looking at the pattern of correlations, one could assume perplexity scores might represent an additional cognitive domain ability, which is important for the content complexity of spoken language and deteriorates before the onset of MCI or AD. The absence of significant associations in HC could be due to smaller variances in MMSE- and TMT A-Scores. Confidence intervals for MCI/AD regarding the association between perplexity and the dementia screening instrument do not include 0 for MCI/AD. At the same time they overlap with HC correlations and therefore do not differ significantly between the two groups. This is in line with Wendelstein (2016), who found lower propositional density 12 years before AD was diagnosed, although impairments were not found in AD in Wendelstein's study. In this sample, AD and MCI were analyzed in one group, although group differences were not the main subject of the present work, but perplexity association might be a reflection of lower content density and a higher use of phrases before the onset of AD/MCI. And – while one must consider small sample sizes – in contrast to MCI/AD, the stability of the HC with respect to the 1-gram perplexity values over time is reflected in highly significant intercorrelations at t1 and t3.

However, potential differences in the spoken language can be masked by a small sample size or the heterogeneity of cognitive decline. Therefore, a more detailed analysis based on a larger data set and taking into account various confounding variables (e.g. education) is necessary to gain insight into the relationships and systematics behind the observations made. On the other hand associations between neuropsychological parameters seem plausible, – e.g. between verbal fluency and another language associated instrument (Boston Naming Test for confrontation naming) at t3, with the dementia screening and with a verbal memory score of t1 (free recall of a word list), which is assumed to be very sensitive. At the same time verbal fluency at t1 is not associated with any measure of cognitive deterioration at t3. Regarding the relevance of the differences between 1-gram and 2-gram prediction it can be assumed that the 1-gram language model correlates higher with the neuropsychological data, i.e. MMSE at t3 than the 2-gram language model. It is possible that the 2-gram model works less well for predictability, although these assumptions are speculative and future research with larger samples is needed to clarify feasible diverging predictability.

Moreover, subjects included in the present study were only followed up into their 8th decade of life. Hence, the present results do not necessarily apply to patients in whom cognitive deterioration develops later in life.

In summary, findings from our study demonstrate that perplexity could be a useful measure of early deterioration, years before the onset of MCI/AD. Despite the work on ASR used for interviews being still in progress, the results of the present work seem promising. Also, the findings reflect first analyses on a new and encouraging collaboration of clinical sciences and computer science with a comparatively easy-to-gain measure that is automatically calculated. This approach provides a chance of gaining a deeper understanding of the role of – maybe language associated – cognitive abilities besides memory in the course of preclinical AD. Further research could clarify the potential predictive value of perplexity for the understanding of the disease, and for physicians in individual diagnostic routines.

Acknowledgments: The Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE) was funded for the first three measurements by the Federal Ministry for Family, Senior Citizen, Women, and Youth, Germany (Bundesministerium für Familie, Senioren, Frauen und Jugend – BMBFSJ) and the Ministry of Sciences, Research, and Arts Baden-Württemberg, Germany (Ministerium für Wissenschaft, Forschung und Kunst, Baden-Württemberg – MWK). The fourth examination wave was funded by the Dietmar Hopp Foundation, Germany (Dietmar Hopp Stiftung).

References

Aebi, Chantal. 2002. *Validierung der neuropsychologischen Testbatterie CERAD-NP*. Eine Multi-Center Studie. Switzerland: University of Basel dissertation. https://www.memoryclinic.ch/fileadmin/user_upload/Memory_Clinic/Literatur/2002/Aebi_2002.pdf (accessed 16 March 2018).

Ahmed, Samrah, Anne-Marie F. Haigh, Celeste A. de Jager & Peter Garrard. 2013. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain* 136(12). 3727–3737. https://doi.org/10.1093/brain/awt269 (accessed 15 December 2018).

Barth, Sonja, Peter Schönknecht, Johannes Pantel & Johannes Schröder. 2005. Neuropsychologische Profile in der Demenzdiagnostik: Eine Untersuchung mit der CERAD-NP-Testbatterie. Fortschritte der Neurologie. *Psychiatrie* 73(10). 568–576. https://doi.org/10.1055/s-2004-830249 (accessed 15 March 2018).

Blanken, Gerhard, Jürgen Dittmann, J.-Christian Haas & Claus-W. Wallesch. 1987. Spontaneous speech in senile dementia and aphasia: Implications for a neurolinguistic model of language production. *Cognition* 27(3). 247–274. https://doi.org/10.1016/S0010-0277(87)80011-2 (accessed 15 December 2018).

Brunét, Étienne. 1978. *Le Vocabulaire de Jean Giraudoux. Structure et Evolution. Statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la Langue Française*. Genève: Slatkine.

De Lira, Juliana O., Karin Z. Oritz, Aline C. Campanha, Paulo H. Ferreira Bertolucci & Thais S. Cianciarullo Minett. 2011. Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics* 23(3). 404–412. https://doi.org/10.1017/S1041610210001092 (accessed 15 December 2018).

Degen, Christina, Pablo, Toro, Peter Schönknecht, Christine Sattler & Johannes Schröder. 2016. Diabetes mellitus Type II and cognitive capacity in healthy aging, mild cognitive impairment and Alzheimer's disease. *Psychiatry Research* 30(240). 42–46. https://doi.org/10.1016/j.psychres.2016.04.009 (accessed 18 March 2018).

Dos Santos, Vasco, Philipp A. Thomann, Torsten Wüstenberg, Ulrich Seidl, Marco Essig & Johannes Schröder. 2011. Morphological cerebral correlates of CERAD test performance in mild cognitive impairment and Alzheimer's disease. *Journal of Alzheimer's Disease* 23(3). 411–420. https://doi.org/10.3233/JAD-2010-100156 (accessed 15 December 2018).

Folstein, Marshal F., Susan E. Folstein & Paul R. McHugh. 1975. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatry Research* 12(3). 189–198. https://doi.org/10.1016/0022-3956(75)90026-6 (accessed 15 December 2018).

Horn, Wolfgang. C. 1983. *Leistungsprüfsystem*. Göttingen: Hogrefe.

Jelinek, Frederick. 1985. The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE* 73(11). 1616–1624. https://doi.org/10.1109/PROC.1985.13343 (accessed 20 November 2018).

Kaplan, Edith, Harold Goodglass & Sandra Weintraub. 1983. *Boston naming test*. Philadelphia: Lea & Febiger.

Knebel, Maren, Julia Haberstroh, Anne Kümmel, Johannes Pantel & Johannes Schröder. 2015. CODEMamb – An observational communication behavior assessment tool for use in ambulatory dementia care. *Aging & Mental Health* 20(12). 1286–1296. https://doi.org/10.1080/13607863.2015.1075959 (accessed 15 December 2018).

Levy, Raymond. 1994. Aging-associated cognitive decline. *International Psychogeriatrics* 6(01). 63–68. https://doi.org/10.1017/S1041610294001626 (accessed 15 March 2018).

Lukatela, Katarina, Paul Malloy, Melissa Jenkins & Ronald Cohen. 1998. The naming deficit in early Alzheimer's and vascular dementia. *Neuropsychology* 12(4). 565–572. http://doi.org/10.1037/0894-4105.12.4.565 (accessed December 2018).

Martin, Peter & Mike Martin. 2000. Design und Methodik der Interdisziplinären Längsschnittstudie des Erwachsenenalters. In Peter Martin, Klaus Udo Ettrich, Ursula Lehr, Dorothea Roether, Mike Martin & Antje Fischer-Cyrulies (eds.), *Aspekte der Entwicklung im mittleren und höheren Lebensalter. Ergebnisse der Interdisziplinären Längsschnittstudie des Erwachsenenalters (ILSE)*, 17–27. Darmstadt: Steinkopff.

McKhann, Guy, David Drachman, Marshall Folstein, Robert Katzman, Donald Price & Emanuel M. Stadlan. 1984. Clinical diagnosis of Alzheimer's disease. Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 34(7). 939–944. https://doi.org/10.1212/WNL.34.7.939 (accessed 15 December 2018).

Morris, John C., Albert Heyman, Richard C. Mohs, J. P. Hughes, Gerald van Belle, Gerda Fillenbaum, E. D. Mellits, C. Clark & the CERAD investigators. 1989. The Consortium to Establish a Registry for Alzheimer's disease (CERAD). Part 1. Clinical and neuropsychological assessment of Alzheimer's Disease. *Neurology* 39(9). 1159–1165. https://doi.org/10.1212/WNL.43.12.2457 (accessed 16 March 2018).

Petermann, Franz & Anja C. Lepach. 2012. *Wechsler Memory Scale – Fourth Edition, German Edition. Manual.* Frankfurt: Pearson Assessment.

Reitan, M. Ralph. 1992. *The Trail Making Test: Manual for administration and scoring*. Tucson: The Reitan Neuropsychological Laboratory.

Schröder, Johannes, Benita Kratz, Johannes Pantel, Elisabeth Minnemann, Ursula Lehr & Heinrich Sauer. 1998. Prevalence of mild cognitive impairment in an elderly community sample. In H. J. Gertz, T. Arendt (eds.), *Alzheimer's Disease — From Basic Research to Clinical Applications (Journal of Neural Transmission. Supplementa 54)*, 51–59. Vienna: Springer. https://doi.org/10.1007/978-3-7091-7508-8_5 (accessed 15 December 2018).

Schröder, Johannes & Johannes Pantel. 2011. *Die leichte kognitive Beeinträchtigung. Klinik, Diagnostik, Therapie und Prävention im Vorfeld der Alzheimer-Demenz*. Stuttgart: Schattauer.

Schönknecht, Peter, Johannes Pantel, Andreas Kruse & Johannes Schröder. 2005. Prevalence and natural course of aging-associated cognitive decline in a population-based sample of young-old subjects. *The American Journal of Psychiatry* 162(11). 2071–2077. https://doi.org/10.1176/appi.ajp.162.11.2071 (accessed 15 December 2018).

Stolcke, Andreas. 2002. SRILM – An extensible language modeling toolkit, in *International Conference on Spoken Language Processing (ICSLP 2002)*, vol. II, 901–904. Denver, Colorado.

Toro, Pablo, Christina Degen, Matthias Pierer, Deborah Gustafson, Johannes Schröder & Peter Schönknecht. 2014. Cholesterol in mild cognitive impairment and Alzheimer's disease in a birth cohort over 14 years. *European Archives of Psychiatry and Clinical Neuroscience* 264(6). 485–492. https://doi.org/10.1007/s00406-013-0468-2 (accessed 15 December 2018).

Weiner, Jochen, Claudia Frankenberg, Dominic Telaar, Britta Wendelstein, Johannes Schröder & Tanja Schultz. 2016a. Towards automatic transcription of ILSE – an Interdisciplinary longitudinal study of adult development and aging. Paper presented at the *Tenth International Conference on Language Resources and Evaluation, LREC'16*. Portorož, Slovenia, 23–28 May. http://www.lrec-conf.org/proceedings/lrec2016/pdf/12_Paper.pdf (accessed 16 March 2018).

Weiner, Jochen, Christian Herff & Tanja Schultz. 2016b. Speech-based detection of Alzheimer's disease in conversational German. Paper presented at the annual *Conference of the International Speech Communication Association, INTERSPEECH*. San Francisco, USA, 8–12 September. https://www.isca-speech.org/archive/Interspeech_2016/pdfs/0100.PDF (accessed 16 March 2018).

Weiner, Jochen, Mathis Engelbart & Tanja Schultz. 2017. Manual and automatic transcription in dementia detection from speech. Paper presented at the annual *Conference of the International Speech Communication Association, INTERSPEECH*. Stockholm, Sweden, 20–24 August. https://www.isca-speech.org/archive/Interspeech_2017/pdfs/0112.PDF (accessed 16 March 2018).

Welsh, Kathleen A., Nelson Butters, Richard C. Mohs, D. Beekly, S. Edland, G. Fillenbaum & A. Heyman. 1994. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). V. A normative study of the neuropsychological battery. *Neurology* 44(4). 609–614. https://doi.org/10.1212/WNL.44.4.609 (accessed 16 March 2018).

Wendelstein, Britta. 2016. *Gesprochene Sprache im Vorfeld der Alzheimer-Demenz. Linguistische Analysen im Verlauf von präklinischen Stadien bis zur leichten Demenz*. University of Heidelberg: Universitätsverlag Winter dissertation.