How Word Choice Affects Cognitive Impairment Detection by Handwriting Analysis: A Preliminary Study

Nicole Dalia Cilia[(✉)], Claudio De Stefano, Francesco Fontanella,
and Alessandra Scotto di Freca

Department of Electrical and Information Engineering (DIEI),
University of Cassino and Southern Lazio,
Via G. Di Biasio 43, 03043 Cassino, FR, Italy
{nicoledalia.cilia,destefano,fontanella,a.scotto}@unicas.it

Abstract. In this paper we present the results of a preliminary study in which we considered two copy tasks of regular words and non-words, collecting the handwriting data produced by 99 subjects by using a graphic tablet. The rationale of our approach is to analyze kinematic and pressure properties of handwriting by extracting some standard features proposed in the literature for testing the discriminative power of non-words task to distinguish patients from healthy controls. To this aim, we considered two classification methods, namely Random Forest and Decision Tree, and a standard statistical ANOVA analysis. The obtained results are very encouraging and seem to confirm the hypothesis that machine learning-based analysis of handwriting on the difference of Word/Non-Word tasks can be profitably used to support the Cognitive Impairment diagnosis.

Keywords: Word and Non-Word · Handwriting · Classification algorithm · Cognitive Impairment

1 Introduction

Traditional theories have postulated that cognitive processing and the motor system were functionally independent so that one movement was the end result of cognitive processing. However, it is increasingly evident that the two systems and their relations are much more complex than previously imagined. The existence of such motor traces of the mind represents a great leap forward in research; researchers have often been forced to look at the "black box" of cognition through indirect, off-line observations, such as reaction times or errors. A serious drawback of this approach is that behavioral results provide little information on how the cognitive process evolves over time and how multiple processes can converge and guide final responses. The motor activities have been

investigated with increasing frequency in recent decades. In this context, the knowledge within cognitive sciences has specialized in many application fields. One of the most prolific is certainly the medical one in which both predictive and rehabilitative research converge. For example, hand movements offer continuous flows of output that can reveal the dynamics being developed and, according to a growing body of literature, can be argued that hand movements can provide, with a high degree of fidelity, traces of the mind.

Furthermore, recently, researchers have shown that some motor activities can represent good indices of neurodegenerative disease. For example, patients affected by Cognitive Impairment (CI) exhibit alterations in the spatial organization and poor control of fine movements. This implies that, at least in principle, some diagnostic signs of CI should be detectable by motor tasks. In this context, alterations in the ability of writing are considered very significant, since writing skill is the result of complex interactions between the biomechanical parts (arm, wrist, hand, etc.) and the control and memorization part of the elementary motor sequences used by each individual to produce the handwritten traces [10]. For example, in the clinical course of CI, dysgraphia occurs both during the initial phase, and in the progression of the disorder [11]. The alterations in the form and in the characteristics of handwriting can, therefore, be indicative of the onset of neurodegenerative disorders, helping physicians to make an early diagnosis. Of course, in this case, these signs are not easily visible and need a measure tools ad hoc to recognize movements and to analyze them.

Moving from these considerations, in this paper we present the results of a preliminary study in which we have considered two copy tasks of regular words and non-words, collecting the data produced by the handwriting of 99 subjects recorded by means of a graphic tablet. The rationale of our approach is to use the kinematic and pressure properties of handwritings by using some standard features proposed in the literature for testing the discriminative power of non-words to distinguish patients from healthy controls. For this aim, we considered two effective and widely used classification methods, namely Random Forest and decision trees, and a standard statistical ANOVA analysis. In particular, the paper is organized as follows: Sect. 2 presents the related work. Section 3 describes the protocol developed to collect traits of patients. Section 4 shows the structure of the dataset and feature extraction method. Section 5 displays the experiments and presents the results obtained. We conclude our paper in Sect. 6 with some future work perspectives.

2   Related Work

To date, there are many studies that investigate how variations in handwriting are prodromal indices of neurodegenerative diseases. An exhaustive review on the subject has been proposed in [5]. In [3] and in [4] instead, an experimental protocol was proposed that included various tasks which investigate the possible impaired motor cognitive functions in the AD. However, in literature, there is a difference between the type of task that seems promising to reveal some typical

characteristics of the patient suffering from dementia, which seem to deviate from purely mnemonic functions. One of these types of tasks has been investigated in [1] and in [2], in which authors studied the copy tasks to support the diagnose of Alzheimer's Disease. However, the choice of the type of task has a very specific meaning. In fact, in copying tasks, unlike those of freewriting, stimuli are constantly present, and subjects can have online feedback without having recourse to memory. The cognitive impact could have consequences on the motor aspects visible from the graphic traits. In the literature, hypotheses have been made on the possibility of using words without semantic content to investigate cognitive impairment. However, many of these studies do not use online measurement tools but usually measure the final result of cognitive processing by investigating the mistakes made in terms of substitutions or inversions of letters. For example, [7] presented a literature review of the research investigating the nature of writing impairment associated with AD. They reported that in most studies words are usually categorized in regular, irregular, and non-words. Orthographically regular words have a predictable phoneme-grapheme correspondence (e.g., cat), whereas irregular words have atypical phoneme-grapheme correspondences (e.g., laugh). Non-words or pseudo-words, instead, are non-meaningful pronounceable letter strings that conform to phoneme-grapheme conversion rules and are often used to assess phonological spelling. In [9] authors proposed a writing test from dictation to 22 patients twice, with an interval of 9–12 months between the tests. They found that agraphic impairment evolved through three phases in patients with AD. The first one is a phase of mild impairment (with a few possible phonologically plausible errors). In the second phase, non-phonological spelling errors predominate, phonologically plausible errors are fewer and the errors mostly involve irregular words and non-words. The study in [8] investigated handwriting performance on a written and oral spelling task. The authors selected thirty-two words from the English language: twelve regular words, twelve irregular words and eight non-words. The study aims to find logical patterns in spelling deterioration with disease progression. The results suggested that spelling in individuals with AD was impaired relative to Healthy Control (HC). Finally, [6] used a written spelling test made up of regular words, non-words and words with unpredictable orthography. The purpose of the study was to test the cognitive deterioration from mild to moderate AD. The authors found little correlation between dysgraphia and dementia severity.

3   Acquisition Protocol

In the following subsections, the protocol designed for collecting handwriting samples and the dataset collection procedure are detailed. The 99 subjects who participated in the experiments, namely 59 CI patients and 40 healthy controls, were recruited with the support of the geriatric ward, Alzheimer unit, of the "Federico II" hospital in Naples. As concerns the recruiting criteria, we took into account clinical tests (such as PET, TAC and enzymatic analyses) and standard cognitive tests (such as MMSE). Finally, for both patients and controls, it was

necessary to check whether they were on therapy or not, excluding those who used psychotropic drugs or any other drug that could influence their cognitive abilities. As previously said, the aim of the protocol is to record the dynamics of the handwriting, in order to investigate whether there are specific features that allow us to distinguish subjects affected by the above-mentioned diseases from healthy ones. The two tasks considered for this study, namely the "word" (W) task composed of two words and the "Non-Word" (NW) task composed of the other two words, require to copy four words in the appropriate box. The words chosen, as suggested in the literature, [6,9] are the two regular words of the Italian language "pane" and "mela" ("bread" and "apple" in English), and the two non-words "taganaccio" and "lonfo", i.e. nonsense words. This task aims to compare the features extracted from handwriting movements of these different types of words. The criteria according to which the structure of protocols was chosen concern:

(i) The copy tasks of NW allow us to compare the variations of the writing respect to the reorganization of the motor plan.
(ii) Tasks need to involve different graphic arrangements, e.g. words with ascenders and/or descendants, allow testing fine motor control capabilities. Indeed the regular words (W) have different descender (the "p" of the first word) and ascender traits (the "l" in the second word). The NWs propose the same structure: The first word have descender traits (the "g" in "taganaccio") and the ascender traits in the second one (the "l" and "f" in "lonfo"). Note that the NWs have to be built following the syntactic rules of language chosen.
(iii) Tasks need to involve different pen-ups that allow the analysis of air movements, which is known to be altered in the CI patients.
(iv) We have chosen to present the tasks asking the subjects to copy each word in the appropriate box. Indeed, according to the literature, the box allows the assessment of the spatial organization skills of the patient.

The two W and the two NW chosen for this study, with different ascender and descender traits, are shown in Fig. 1.

As concern the acquisition tool we have used a Graphic Tablet able to record the movements of the pen used by the examined subject. X, y, and z (pressure) coordinates are recorded for each task and saved in a txt file. The task was printed on A4 white sheet placed on the graphic tablet.

4   Structure of Dataset and Methods

The features extracted during the handwriting process have been exploited to investigate the presence of neurodegenerative diseases in the examined subjects. We used the MovAlyzer tool to process the handwritten trace, considering both on-paper and on-air traits. Their segmentation in elementary strokes was obtained by assuming as segmentation points both pen up and pen down, as well
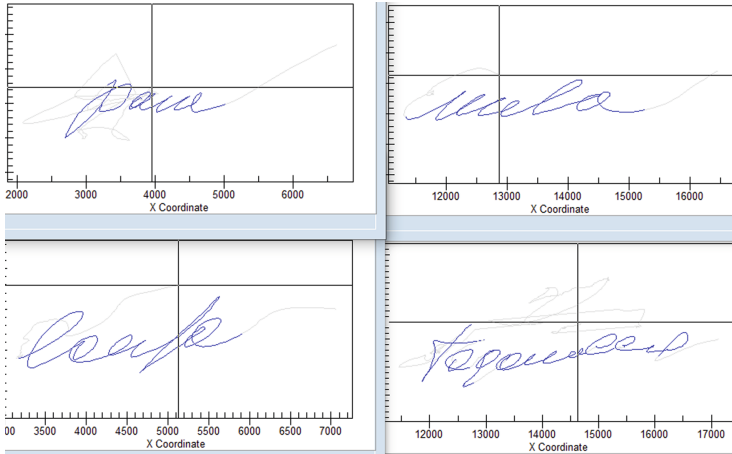
Fig. 1. Esemples of tasks. Above the regular words and below the non-words.

as the zero-crossing of the vertical velocity profile. The feature values were computed for each stroke and averaged over all the strokes relative to a single task. We also merged the W tasks between them and NW tasks, considering them as two repetitions of the same type of task. In our experiments we have included the following features for each considered type of task, separately computed for both on-paper and on-air traits: (i) Number Of Stroke; (ii) Absolute Velocity Mean; (iii) Size Mean; (iv) Loop Surface; (v) Slant Max; (vi) Horizontal Size Mean; (vii) Vertical Size mean; (viii) Total Duration; (ix) Duration Mean; (x) Absolute Size mean; (xi) Peak Vertical Velocity; (xii) Peak Vertical Acceleration; (xiii) Absolute Jerk; (xiv) Average Pen Pressure. We also included age and sex of the subject.

We have analyzed this dataset into two steps: the classification and the standard statistical analysis. For both procedures, we designed the experiments organizing the data in three groups: data obtained by extracting on paper features, data related to on-air features and data including both types of features. Thus, for each type of task, we generated three different datasets, each relative to one of the above groups and containing the samples derived for the 99 subjects. For the classification step, we used two different classification schemes, namely the Random Forest (RF) and the Decision Trees (DT) with C algorithm. For both of them, 500 iterations were performed and a 5 fold validation strategy was considered. For the statistical step, we have used the two-way ANOVA analysis and a Multiple Comparison with Holm-Sidak correction as Post Hoc analysis to understand which variables have a major effect on the results, following our 2×2 experimental design (Patient - HC; Word - NonWord). In particular, the two-factor ANOVA, used in these analyses, allows us to understand if there is a main effect on the interaction of the two factors (Label or Task), that is, in

other words, if the effects of the two factors are dependent or independent. The obtained results with two-way ANOVA are:

(i) On the first independent variable, that is "Label" which includes Patients and Healthy control that allow us to compare the mean of two independent groups.
(ii) On the second independent variable, that is "task" that allow us to understand if there is a significant difference in the mean of value of W and NW tasks.
(iii) The two-way ANOVA which examines the influence of two different categorical independent variables (Patient - HC; W - NW) on one continuous dependent variable (one feature) taken into account.

All of these analyses are calculated on the values of each feature, used as a dependent variable, separately. Furthermore, the values are calculated on the three groups of features shown above.

5   Experimental Results

As concerns the classification, in Fig. 2 we summarize the values of Accuracy for each group of tasks. The first column reports the type of task considered, the second one the classifier employed, while the following columns report, for each task, the value of Accuracy using all features, on-paper features and on-air features respectively.

From this table, we can point out that: firstly, in the large majority of cases, for each task the value of the Accuracy is over 80.00%, reaching the best value in the second group of tasks (NW) equal to 85.35 %. Secondly, we can observe that, in the same condition of considered feature, emerges a better classification using the RF classifier compared to DT. This is easily justifiable considering that Random Forest, unlike DT, is an ensemble of classifiers. However, as reported in the last row, the best-obtained result occurs with NW task. Indeed the classification accuracy of the second task is almost always 5% higher than the first type of tasks. This means that copying a word with no meaning can be more useful to diagnose CI than copying a regular word and that the same features better predict CI in NW task than W task.

In the Fig. 3 we show, instead, the value of standard statistical analysis, with two-way ANOVA, using R tool. In this table, we indicate, for each group of tasks, the statistical significance results of each feature taken into account. We report the value of the interaction of rows and columns (Label and Tasks) separately, and the value of their interaction (in grey), according to the translation criteria shown as footnote. Empty cells indicate that no statistical significance emerged for these factors.

We can claim that there is a high significance value for almost all features on the two factors separately. In particular:

Task	Classifier	Accuracy		
		All	OnPap	OnAir
Word	J48	78.79	81.52	78.79
	RF	82.32	83.15	82.83
Non-Word	J48	84.85	82.06	80.81
	RF	**84.85**	**83.69**	**85.35**

Fig. 2. Accuracy of Words and Non-Words tasks

(i) As regards all features we can note that there is an interaction on Number of Stroke, Absolute Velocity Mean, Total Duration and Jerk. As reported in Fig. 4 and in detail in Fig. 7:
  (i) the Variation of Number of Stroke is significant for each combination, except for W:P vs NW:HC. Patients produce almost double Number of Stroke of HC both for W and NW condition. Furthermore, HCs produce fewer strokes in NW than Patients in W condition.
  (ii) the Duration Tot is doubled both for P vs HC and W vs NW. But it is worth noticing that HCs employ less time to complete NW task than P to produce W task.
  (iii) for the Absolute Velocity Mean and for Jerk the value of HCs increases going from W to NW task but conversely, it decreases for patients.
(ii) As regards on-air features we can note that there is an interaction on Number of Stroke, Absolute Velocity Mean, Total Duration, Jerk and Loop Surface Mean. As reported in Fig. 5 and in detail in Fig. 7:
  (i) the variation of the Number of Stroke on air is even more evident than in all features. Patients produce almost double Number of Stroke of HCs both for W and NW condition. Moreover, HCs produce more stroke of NW than Patients in W condition.
  (ii) the Duration Tot is doubled both for P vs HC and W vs NW.
  (iii) for the Absolute Velocity Mean and for the Jerk, the values show the same relation compared to all feature but, in this case, the difference between HC and Patient in W task in Abs Velocity Mean is irrelevant.

(iii) Finally, as regards on paper features we can note that there is an interaction only on Slant and Peak Acceleration Mean. As reported in Fig. 6 and in detail in Fig. 7:

  (i) The variation of the Slant is significant for almost all comparisons and HC in NW is less than Slant in P in W condition.

  (ii) Peak Acceleration Mean shows an opposite trend between HCs and Patients in the comparison of W and NW tasks.

The second experimental setting has shown that some features can be particularly discriminative of the disease. For example, Absolute Velocity Mean and Jerk on all features and on-air features show an inverted trend between HCs and Ps in the passage from W tasks to NW tasks.

FEATURE	ALL		ON AIR		ON PAPER	
	LABEL	TASK	LABEL	TASK	LABEL	TASK
NumOfStroke		*		*		
	***	***	***	***	***	****
AbsoluteVelocityMean		.		*		
	***		*		***	
SizeMean					***	
LoopSurfaceMean				*		
	***			.	***	.
SlantMax						.
	***	***	***	***	***	**
HorizontalSizeMean	***	***	**		***	***
VerticalSizeMean			**		**	**
DurationTot		*		**		
	***	***	***	***	***	***
DurationMean						**
AbsoluteSizeMean	***		**			
PeakVerticalVelocityMean				**		
PeakVerticalAccelerationMean						.
		.				
JerkMean		*		**		
		.	.		***	
PenPressure Mean	***	**			***	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

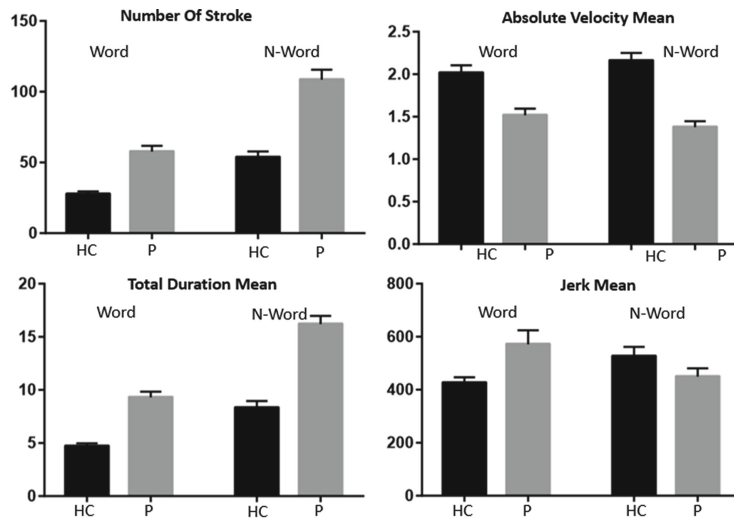**Fig. 3.** Statistical significance with ANOVA
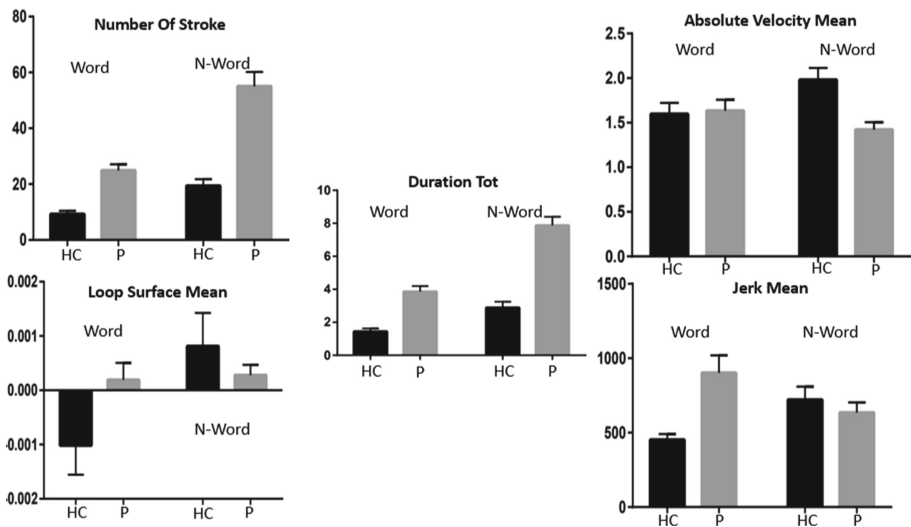
Fig. 4. Results of Post Hoc analysis on all features

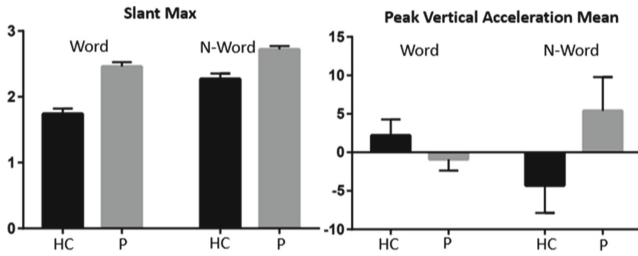**Fig. 5.** Results of Post Hoc analysis on air features

Fig. 6. Results of Post Hoc analysis on paper features

	Comparison	NumOfStroke	VelMean	Loop	Slant	DurTot	PeakAccel	JerkMean
ALL	W:HC vs. W:P	0.0009	0.0001			<0.0001		ns
	W:HC vs. N W:HC	0.0118	ns			0.0024		ns
	W:HC vs. N W:P	<0.0001	<0.0001			<0.0001		ns
	W:P vs. N W:HC	ns	<0.0001			ns		ns
	W:P vs. N W:P	<0.0001	ns			<0.0001		ns
	N W:HC vs. N W:P	<0.0001	<0.0001			<0.0001		ns
OnAir	W:HC vs. W:P	0.0166	ns	ns		0.0005		0.0090
	W:HC vs. N W:HC	ns	ns	0.0232		ns		ns
	W:HC vs. N W:P	<0.0001	ns	ns		<0.0001		ns
	W:P vs. N W:HC	ns	ns	ns		ns		ns
	W:P vs. N W:P	<0.0001	ns	ns		<0.0001		ns
	N W:HC vs. N W:P	<0.0001	0.0084	ns		<0.0001		ns
OnPap	W:HC vs. W:P				<0.0001		ns	
	W:HC vs. N W:HC				<0.0001		ns	
	W:HC vs. N W:P				<0.0001		ns	
	W:P vs. N W:HC				ns		ns	
	W:P vs. N W:P				0.0042		ns	
	N W:HC vs. N W:P				<0.0001		ns	

Fig. 7. Results of Multiple Comparison with Holm-Sidal correction- Post Hoc analysis

6   Conclusion and Open Issues

In this paper, we presented a novel solution to diagnose Cognitive impairment by means of only two writing tasks which included a copy of regular words and non-words. The preliminary results obtained are very encouraging and the work is in progress to increase general performance. As appeared from the experimental results there are some features particularly discriminative of the disease. Then, the next steps to be taken will include a classification experiment using just these features. We could also include a feature selection approach to improve overall performances.

References

1. Cilia, N.D., De Stefano, C., Fontanella, F., Molinara, M., Scotto di Freca, A.: Handwriting analysis to support alzheimer's disease diagnosis: a preliminary study. In: Vento, M., Percannella, G. (eds.) CAIP 2019. LNCS, vol. 11679, pp. 143–151. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29891-3_13
2. Cilia, N.D., De Stefano, C., Fontanella, F., Molinara, M., Scotto di Freca, A.: Using handwriting features to characterize cognitive impairment. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds.) ICIAP 2019. LNCS, vol. 11752, pp. 683–693. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30645-8_62
3. Cilia, N.D., De Stefano, C., Fontanella, F., Scotto di Freca, A.: An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis. Procedia Comput. Sci. **141**, 466–471 (2018)
4. Cilia, N.D., De Stefano, C., Fontanella, F., Di Freca, A.S.: Using genetic algorithms for the prediction of cognitive impairments. In: Castillo, P.A., Jiménez Laredo, J.L., Fernández de Vega, F. (eds.) EvoApplications 2020. LNCS, vol. 12104, pp. 479–493. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43722-0_31
5. De Stefano, C., Fontanella, F., Impedovo, D., Pirlo, G., Scotto di Freca, A.: Handwriting analysis to support neurodegenerative diseases diagnosis: a review. Pattern Recogn. Lett. **121**, 37–45 (2018)
6. Luzzatti, C., Laiacona, M., Agazzi, D.: Multiple patterns of writing disorders in dementia of the alzheimer-type and their evolution. Neuropsychologia **41**(7), 759–772 (2003)
7. Neils-Strunjas, J., Groves-Wright, K., Mashima, P., Harnish, S.: Dysgraphia in Alzheimer's disease: a review for clinical and research purposes. J. Speech Lang. Hear. Res. **49**(6), 1313–1330 (2006)
8. Pestell, S., Shanks, M.F., Warrington, J., Venneri, A.: Quality of spelling breakdown in alzheimer's disease is independent of disease progression. J. Clin. Exp. Neuropsychol. **22**(5), 599–612 (2000)
9. Platel, H., et al.: Characteristics and evolution of writing impairment in alzheimer's disease. Neuropsychologia **31**(11), 1147–1158 (1993)
10. Tseng, M.H., Cermak, S.A.: The influence of ergonomic factors and perceptual-motor abilities on handwriting performance. Am. J. Occup. Ther. **47**(10), 919–926 (1993)
11. Yan, J.H., Rountree, S., Massman, P., Doody, R.S., Li, H.: Alzheimer's disease and mild cognitive impairment deteriorate fine movement control. J. Psychiatr. Res. **42**(14), 1203–1212 (2008)