doi:10.3233/978-1-61499-852-5-705

Automated Syntactic Analysis of Language Abilities in Persons with Mild and Subjective Cognitive Impairment

Kristina LUNDHOLM FORS[1], Kathleen FRASER and Dimitrios KOKKINAKIS
Department of Swedish Language, University of Gothenburg, Sweden

Abstract. In this work we analyze the syntactic complexity of transcribed Swedish-language picture descriptions using a variety of automated syntactic features, and investigate the features' predictive power in classifying narratives from people with subjective and mild cognitive impairment and healthy controls. Our results indicate that while there are no statistically significant differences, syntactic features can still be moderately successful at distinguishing the participant groups when used in a machine learning framework.

Keywords. cognitive impairment, syntactic analysis, natural language processing.

1. Introduction

Dementia is characterized by a decline in cognitive skills, memory, language, and executive function which has far-reaching effects on peoples' everyday activities. Early diagnosis of dementia has important clinical significance and impact on society, considering the fact that the total estimated worldwide cost of dementia in 2015 was 818 billion US$, while it is estimated that by 2018, dementia will become a "trillion dollar disease" [1]. The research presented here is part of a larger project which investigates how multimodal data resources and language related measures can be used for the development and evaluation of classification algorithms for differentiating between healthy adults and persons in various stages of cognitive decline. In this study, we use automated methods to evaluate the syntactic complexity of spoken narratives to investigate whether syntactic ability is affected in the preclinical stages of dementia, and whether automated measures of syntactic ability may be useful in detecting early cognitive impairment.

2. Background

Although progressive memory impairment is generally considered the primary cognitive feature for neurodegenerative diseases such as Alzheimer's disease (AD), it is known that language deficits also occur in AD patients as a primary symptom early in the disease [2].These language disturbances seem to be an intrinsic part of AD and appear to be among the earliest of symptoms. Mild cognitive impairment (MCI) is often considered a prodoromal state of dementia [3], and is characterized by minor problems with cognition

[1] Corresponding Author: Kristina Lundholm Fors, Department of Swedish, University of Gothenburg, PO Box 200, SE 405 30, Gothenburg, Sweden; E-mail: kristina.lundholmfors@svenska.gu.se.

that are not severe enough to interfere significantly with daily life, and that do not warrant a dementia diagnosis. A person diagnosed with subjective cognitive impairment (SCI) is perceiving a decline in cognitive function, but does not differ from healthy persons on standardized tests. SCI is a risk factor for MCI but the majority of persons with SCI do not develop dementia, and SCI can have other underlying causes such as depression [4].

There have been conflicting findings about whether the syntactic complexity of spoken language is affected in MCI and AD. Ahmed et al. [5] found that syntactic complexity in persons with MCI was impaired compared to healthy controls, and that there was a degradation in syntactic complexity with disease progression. Roark et al. [6] reported shorter clauses and fewer left-branching structures in MCI speakers on a story-retelling task. In their review, Boschi et al. [7] found that a greater number of inflectional errors in persons with AD is the only relatively consistent result on the morpho-syntactic level, but that several studies report a simplification of syntax. However, other research shows that the assessment of syntax in people with possible AD results in both grammatical and coherent structures [8]. There appears to be some effect of the speech task on the resulting complexity. For instance, Sajjadi et al. [9] found that the picture description task was more sensitive to semantic deficits, while directed interviews were more sensitive to morpho-syntactic deficits in semantic dementia and AD. However, previous work analyzing picture description narratives in English has uncovered some signs of syntactic simplification in AD speakers on this task, including reduced number of prepositional phrases and subordinate clauses [10]. Automated analyses of picture descriptions also reported some evidence for syntactic simplification in AD, although not to the extent or severity seen in other types of dementia, such as primary progressive aphasia [11,12].

3. Data Set

The participants in this study are all part of the longitudinal ongoing Gothenburg MCI study [13]. All patients in the study undergo baseline investigations, such as neurological examination, psychiatric evaluation, and brain imaging. The demographic information for the participants is presented in Table 1. All participants gave informed written consent. Both the Gothenburg MCI-study and the current one are approved by the local ethical committee review board (ref. nr: L09199, 1999; T479-11, 2011 & 206-16, 2016). The difference in age between the groups is not statistically significant, but there is a significant difference in years of education. A post-hoc LSD test reveals that the SCI group is significantly more educated than both the HC group ($p = 0.001$) and the MCI group ($p = 0.026$). The MMSE score differs significantly between the three groups, and post-hoc LSD tests show that scores in the MCI group are significantly lower than in the HC group ($p < 0.0001$) and the SCI group ($p < 0.0001$), whereas there is no difference between the SCI group and the HC group.

The Cookie-Theft picture from the Boston Diagnostic Aphasia Examination [14] was used to elicit spontaneous speech. The participants were presented with the picture and were asked to describe everything that they could see and everything that was happening in the picture. They were told that they would not be interrupted and could talk for as long as they liked. The narratives were recorded and manually transcribed. During speech transcription, special attention was paid to non-speech acoustic events including speech dysfluencies such as filled pauses, hesitations, false-starts, and repetitions.

Table 1. Demographic information: age, education, and Mini-Mental State Exam (MMSE) scores are given in the format: mean (standard dev.). The MMSE is a general test of cognitive status and has a max score of 30.

	HC (*n*=36)	**MCI (*n*=31)**	**SCI (*n*=23)**	**Group comparison**
Age (years)	67.9 (7.2)	70.1 (5.6)	66.3 (6.9)	F(2, 87) = 2.287, p = 0.108
Education (years)	13.2 (3.4)	14.1 (3.6)	16.1 (2.1)	F(2, 87) = 6.014, p = 0.004
MMSE	29.6 (0.61)	28.2 (1.43)	29.5 (0.90)	F(2, 86) = 16.275, p <0.0001
Sex (F/M)	23/13	16/15	14/9	

4. Methods: syntactic analysis

Here we examined the syntactic complexity of the Cookie-Theft transcriptions using tools from natural language processing. Most of the syntactic measures operate on parse trees, one for each sentence in a text. These parse trees were generated by two Swedish parsers, a dependency one [15] and a constituent-based one [16]. The following syntactic features were extracted from the text, in addition to sentence length and number of false starts and interrupted or incomplete sentences:

- **Dependency distance:** Dependency distance is measured as the number of words between a given word and its dependency head, calculated for each word in the sentence. We compute average, maximum, and total dependency distance for each sentence, and then average these quantities over each sentence in the transcript.
- **Phrase type proportion:** Phrase type proportion and length (below) are derived from work on rating the fluency of machine translations [17]. The phrase type proportion is the total number of words belonging to a given phrase type (here prepositional phrases, noun phrases, and verb groups), divided by the total number of words in the narrative. We additionally extend this feature to apply to clauses; namely main finite and infinitive clauses, and subordinate clauses.
- **Phrase type length:** The phrase type length is the total number of words belonging to the given phrase or clause type, divided by the total number of occurrences of that phrase or clause type.

5. Results

The results of the syntactic analyses are displayed in Table 2. After correcting for multiple comparisons, our alpha value is 0.003. Using leave-one-out cross-validation, we train a random forest classifier [18] with 50 trees, and with the maximum number of features and maximum depth hyperparameters selected in a nested validation loop. In the task of distinguishing the MCI and HC groups, the classifier achieves a mean F-score of 0.68, and in the task of distinguishing between the MCI and SCI groups, an F-score of 0.66. However, when distinguishing between the SCI and HC groups, the F-score is only 0.54.

Table 2. Means and standard deviations of the syntactic measures in the three groups. * = distance; † = proportion; ‡= length. The fifth column displays the results of the ANOVA.

Feature	HC	MCI	SCI	Comparison
Mean length of sentence	14.05 (4.233)	15.11 (3.633)	16.44 (4.040)	2.524 (p=0.086)
False starts	0.005 (0.008)	0.009 (0.009)	0.003 (0.005)	4.243 (p=0.017)
Interruptions	0.006 (0.006)	0.01 (0.012)	0.008 (0.007)	3.064 (p=0.052)
Average dependency *	1.85 (0.206)	1.90 (0.229)	1.92 (0.209)	0.686 (p=0.506)
Total dependency *	27.82 (11.528)	31.19 (12.361)	34.71 (11.750)	2.395 (p= 0.097)
Maximum dependency *	5.94 (2.047)	6.58 (2.017)	6.81 (2.024)	1.515 (p=0.225)
PP type †	0.18 (0.056)	0.18 (0.048)	0.19 (0.033)	0.821 (p=0.443)
VG type †	0.26 (0.035)	0.26 (0.033)	0.25 (0.035)	0.264 (p=0.769)
NP type †	0.45 (0.050)	0.45 (0.052)	0.44 (0.051)	0.611 (p=0.545)
PP ‡	2.43 (0.248)	2.53 (0.354)	2.48 (0.181)	1.101 (p=0.337)
VG ‡	1.46 (0.186)	1.44 (0.151)	1.48 (0.186)	0.327 (p=0.722)
NP ‡	1.41 (0.154)	1.42 (0.135)	1.43 (0.135)	0.090 (p=0.914)
Main clause; finite verb †	0.52 (0.092)	0.47 (0.102)	0.47 (0.061)	4.146 (p=0.019)
Main clause; non-finite verb †	0.08 (0.058)	0.13 (0.059)	0.09 (0.059)	5.669 (p=0.005)
Subordinate clause †	0.29 (0.102)	0.28 (0.110)	0.33 (0.096)	1.569 (p=0.214)
Main clause; finite verb ‡	5.14 (0.772)	4.92 (0.666)	5.16 (0.753)	0.940 (p=0.395)
Main clause; non-finite verb ‡	5.87 (2.125)	6.20 (1.293)	6.19 (1.477)	0.381 (p=0.685)
Subordinate clause ‡	5.60 (1.20)	5.50 (0.908)	5.88 (0.935)	0.949 (p=0.391)

6. Discussion and Conclusion

The results of the statistical analysis indicate that there are no syntactic features which vary significantly between the groups after correcting for multiple comparisons. However, there are some trends which approach significance, namely an increase in the number of false starts, an increase in the proportion of main clauses where the main verb is nonfinite, and a reduction in the proportion of main clauses where the verb is finite in the MCI group. Since nonfinite verbs typically occur in auxiliary verb constructions, this result is somewhat unexpected given previous work reporting a decrease in auxiliary verb phrases and an increase in simple verb phrases in preclinical AD [19]. However, more work is needed to gain a clear interpretation of this effect, and to fully account for the effect of language differences (here Swedish, as opposed to Spanish in [19]). It may also be useful to evaluate verb use on a discourse level, as narrative style may have an effect on verb tense [20].

Despite a lack of statistical significance, when taken together the features were still moderately successful in training random forest classifiers to distinguish between the MCI group and the HC group, and the MCI group and SCI group. In both cases, output from the classifiers indicate that the most "important" features are false starts and proportion of main infinitive clauses, reinforcing the findings from the statistical analysis. However, the classifier was unable to distinguish between the HC and SCI groups, which is perhaps expected given that the SCI participants perform as well as the controls on all other cognitive and language tests in the battery.

The work presented here is a preliminary analysis of syntax in MCI and SCI. In future work we would like to increase the number of features to look at more specific syntactic structures, as well as to compare spoken and written Cookie-Theft descriptions. Additionally, we would like to compare and contrast with other elicited speech tasks, including conversational speech and story-telling. Finally, we plan to incorporate the syntactic features with measures of semantics, information content, discourse-level processing, and acoustic/phonetic production to gain a more complete picture of speech in mild cognitive impairment.

7. Acknowledgements

This work has received support from Riksbankens Jubileumsfond - The Swedish Foundation for Humanities & Social Sciences, through the grant agreement no: NHS 14-1761:1.

References

[1] M. J. Prince, World Alzheimer Report 2015. The global impact of dementia: An analysis of prevalence, incidence, cost and trends. Alzheimer's Disease International, 2015.
[2] V. O. B. Emery, "Language impairment in dementia of the Alzheimer type: a hierarchical decline?," J of Psychiatry in Medicine, vol. 30, no. 2, pp. 145–164, 2000.
[3] K. Ritchie and J. Touchon, "Mild cognitive impairment: conceptual basis and current nosological status," The Lancet, vol. 355, no. 9199, pp. 225–228, 2000.
[4] M. Eckerstrom, ¨ Subjective cognitive decline in memory clinic patients characteristics and clinical relevance. PhD thesis, Sahlgrenska Academy at University of Gothenburg, 2017.
[5] S. Ahmed, A.-M. F. Haigh, C. a. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease.," Brain, no. Pt 12, pp. 3727–37, 2013.
[6] B. Roark et al., "Spoken language derived measures for detecting mild cognitive impairment," IEEE transactions on audio, speech, and language processing, vol. 19, no. 7, pp. 2081–2090, 2011.
[7] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neu- ` rodegenerative language disorders: a review," Frontiers in Psychology, vol. 8, 2017.
[8] S. Kemper et al., "On the preservation of syntax in Alzheimer's disease: Evidence from written sentences," Arch of neurology, vol. 50, no. 1, pp. 81–86, 1993.
[9] S. A. Sajjadi, K. Patterson, M. Tomek, and P. J. Nestor, "Abnormalities of connected speech in semantic dementia vs Alzheimer's disease," Aphasiology, vol. 26, no. 6, pp. 847–866, 2012.
[10] B. Croisile et al., "Comparative study of oral and written picture description in patients with Alzheimer's disease," Brain and language, vol. 53, no. 1, pp. 1–19, 1996.
[11] M. Yancheva, K. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for Alzheimers disease and related dementias," in 6th SLPAT, pp. 134–139, sn, 2015.
[12] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimers disease in narrative speech," J of Alzheimer's Disease, vol. 49, no. 2, pp. 407–422, 2016.
[13] A. Wallin et al., "The Gothenburg MCI study: design and distribution of Alzheimers disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up," J of Cerebral Blood Flow & Metabolism, vol. 36, no. 1, pp. 114–131, 2016.
[14] P. Goodglass, B. Barresi, and E. Kaplan, "Boston Diagnostic Aphasia Examination," 1983. Philadelphia: Lippincott Williams and Willkins. A Wolters Kluwer Company.
[15] J. Nivre et al., "Maltparser: A language-independent system for data-driven dependency parsing," J of NLE, vol. 13, no. 2, pp. 95–135, 2007.
[16] D. Kokkinakis and S. J. Kokkinakis, "A cascaded finite-state parser for syntactic analysis of swedish," in 9th conf on Eur chapter of the Association for Computational Linguistics, pp. 245–248, 1999.
[17] J. Chae and A. Nenkova, "Predicting the fluency of text with shallow structural features: Case studies of machine translation and human-written text," in 12th EACL, pp. 139–147, 2009.
[18] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J Machine Learning Res, vol. 12, pp. 2825–2830, 2011. [19] F. Cuetos et al., "Linguistic changes in verbal expression: A preclinical marker of Alzheimer's disease," J Int Neuropsychol Soc, vol. 13, no. 3, pp. 433–439, 2007.
[19] F. Cuetos et al., "Linguistic changes in verbal expression: A preclinical marker of Alzheimer's disease," J Int Neuropsychol Soc, vol. 13, no. 3, pp. 433–439, 2007.
[20] C. Drummond et al., "Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment," Frontiers in Aging Neuroscience, vol. 7, pp. 1–11, 2015.