Contrastive conversational analysis of language production by Alzheimer's and control people

Maïté BOYÉ[a] and Natalia GRABAR[b] and Mai THI TRAN[ab1]
[a] *Institut d'Orthophonie, Université Lille 2, Lille, France*
[b] *CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France*

Abstract. Alzheimer's disease is characterized by memory disorders, although it affects also other cognitive functions (executive functions, attention, gnosia, praxia and language) and the communication ability. Our objective is to study verbal communication of people affected by the Alzheimer's disease at early to moderate stages. One particularity of our approach is that we are working in ecological conversational situations: people are interviewed by persons they know and in non-artificial environment. We propose a contrastive study of verbal productions of five people affected by the Alzheimer's disease and of five control people, both over 80 years old. To obtain quantitative and qualitative results, the oral corpora are transcribed and processed with the NLP methods and tools, and manually. Our results indicate that the Alzheimer's patients present lexical and semantic deficit and that their conversation is reduced comparing to the control people.

Keywords. Conversation, Linguistics, Natural Language Processing, Language of Patients, Alzheimer's disease, Loss of Language, Speech and Language Therapy, France

Introduction

It is estimated that currently over 850,000 people are affected by the Alzheimer's disease (AD) in France, and that the number may reach up to 1,200,000 in 2020 [1]. Most of the time, AD patients present a satisfactory communication skills at early stages of the disease: disorders concern language production, although its understanding is preserved [2]. The main disorder, relative to the lexical production, shows lexical and semantic deficiency due to progressive loss of semantic knowledge [3]. Phonological and morpho-syntactic skills are preserved at this step [2]. While pathology is developing, lexical and semantic troubles become more important, with the increasing of semantic paraphasia, loss of lexical understanding and morpho-syntactic disorders. Disorders gain then both language production and understanding, which makes the communication increasingly difficult [2]. Gnosic disorders (such as prosopagnosia) [1], as well as other cognitive and functional troubles [4,5] can appear. Decrease in language competence and in cognitive functions lead to muteness [2] and

1 Corresponding Author.

progressive social isolation. Maintaining the communication is the major challenge of speech and language therapists.

Studies of healthy old people indicate that: language performance depends on their education and age [6]; they show more frequently approximate and incoherent discourse, ambiguous references, periphrases and redundancies [7]; they do grammatical errors more frequently than younger persons [8,9]; they often speak about past events to maintain the conversation; they become egocentric and are not attentive to what other people are saying, which can lead to monologues; discussed topics are respected but with very frequent digressions and loss of the main idea [10]. Some existing work also proposed analysis of semi-supervised or supervised conversations of old people with AD. Concerning the comparative studies of language communication of healthy and AD persons, this existing work reveal that: AD persons produce syntactically poorer sentences [8,9], mention lesser number of ideas and words [11], produce redundant, less precise and informative discourse [12,13,14,15], rare use of the modalizers [12], pronouns miss the reference, which implies a loss of the semantic cohesion [14,15].

Our objective is to propose a contrastive study of verbal conversation of the AD and healthy patients. We want to propose additional diagnostic and therapeutic elements: define specific criteria of the conversational AD language and distinguish deficient and preserved elements. We work with transcribed spoken corpus of AD and healthy persons collected in ecological and non-artificial context (known interlocutor, non-supervised conversation). This is one of the specificities of our study. The other one is that the corpus is processed with Natural Language Processing (NLP) tools and through manual analysis.

Methods

The corpus is collected with ten people: five AD patients (89 to 99 years) and five healthy people (83 to 102 years). The average age is 90 years. These people are chosen in order to build as homogeneous as possible cohort concerning their age, gender (females), social and cultural level, and residence place. Concerning the AD patients, they have early to moderate disease stage and they can communicate (both, speak and understand). All of them have gone through a neurological test at most six months before the recruitment. They do not present other neurological or psychiatric history. To start the conversation, two pictures are proposed (train departure for vacations and bicycle ride). First the preferred picture and then the other one are discussed. The conversations can last for 20 to 40 minutes. The conversation are then transcribed with the Transcriber software [16]. Both text transcription and annotations *(eg,* disfluencies, hesitations, pauses) can be encoded. The transcription process needed over 100 hours. The transcribed corpus is processed with the NLP tools: POS tagging and lemmatization with TreeTagger [17] and Flemm [18]. The lemmas are morphologically analyzed Dérif [19]. These tools are used in their usual settings: no adaptations have been done.

Several aspects, covering intra and inter-individual communication, are studied. The objective is to provide comparative analysis of the conversation of AD and healthy people as exhaustively as possible. Among the studied aspects, we have the following: individual and overlapped time of speech (participation of person in the conversation), turns of speaking (dynamics of language exchange), language disfluencies (such as

empty and non-empty (*heu*, *hmm*) pauses, primes, repetitions and stutters, self-corrections, incomplete sentences) [20,21], breath groups, average length of sentences, informativity of sentences (use of discourse elements such as *oh*, *ah*, *bon (well)*, *là (here)*), sentences containing only *yes* and *no* statements (up to ten occurrences), use of modalizers (cover non referential elements such as *je pense que (I think that)* and indicates that the person has some perspective and is detached from what he is saying [12]); personal pronouns (number of occurrences and their percentage comparing to the number of nouns); verbs (we distinguish the tenses of verbs and compute the percentage of a given tense among all the verbs); distribution of lemmas according to POS tags; reported speech and interpolated clauses (both allow to observe the detachment of the person from what he is saying); morphological complexity (observed further to the morphological analysis of words into bases and affixes); lexical diversity (computes the number of different words within the POS categories); lexical frequency (frequency of a given lemma for a given speaker).

Results and Discussion

Table 1. Quantitative results and salience of the aspects studied.

Aspects studies	AD patients	Healthy people
Time of speech	*11.26 min*	*14..20 min*
Overlaps of speech	*0.59 min*	*1.01 min*
Empty pauses	*65*	*31*
Non-empty pauses	*13*	*34.60*
Disfluencies	*2.80*	*8.40*
Number of words per breath group	3.90 words	4.31 words
Average length of sentences	*4.79 words*	*7.26 words*
Sentences containing only yes and no statements	*88.80*	*52.40*
Personal pronouns (total percentage)	*1%*	*0.65%*
Personal pronouns (1st person singular percentage)	*36%*	*23%*
Verbs	*377*	*553*
Distribution of lemmas according to POS tags	10	14
Lexical diversity	*168.60*	*302.40*

In Table 1 we indicate the average quantitative results for AD and healthy people groups. In italic and larger characters we indicate those aspects that show salient differences and may be involved in the diagnosis of the disease at early stages.

The results obtained describe extensively the discourse produced by AD patients. We can observe the following differences by comparison with the healthy people: AD patients produce a lesser number of words for the same speaking time; they show a higher number of turns of speaking; they speak with lesser speed; they have a greater number of stutters, self-corrections and incomplete sentences; they produce shorter sentences; they have an important number of *yes* and *no* sentences; they have a greater number of empty pauses and lesser number of non-empty pauses; they seldom use reported speech and interpolated clauses; they have poorer lexicon; they show a very high percentage of personal pronouns, especially of pronouns like *je (I)*. In general, we can notice that the discourse produced by the AD patients is less fluent and rich than the discourse of healthy people. It misses several aspects that mark the spontaneous and

natural speech (disfluencies, reported speech and interpolated clauses). Nevertheless, several elements typical of the spoken language are present, which means that the AD discourse begins to show structural and conceptual disorders at this level. We can see that several salient aspects are related to the NLP features (*eg*, lexicon, syntax). Notice that in an existing study [22], dedicated to the analysis of aphasia discourse, the authors observed that classical features (and not NLP features) are relevant to the diagnosis. The fact that the NLP tools can efficiently help the analysis and diagnosis of the Alzheimer's disease at early stages is a very positive result of our study.

Yet, the results obtained are to be taken with precaution because the variability within each group is very high. This is due to the personality of each participant, their habits and ways of life. In future work, it is necessary to study a larger group of people to reduce the influence of the inter-personal variability.

Conclusion and Perspectives

We proposed a comparative analysis of spoken corpora produced by Alzheimer's disease patients and by healthy people, all over 80 years old. The data allowed to perform an extensive study of conversation of these two categories of people and to compare them. Up to now, such analysis was difficult due to several aspects involved and to the fact that the existing studies seldom addressed the conversation analysis of people with AD: these studies usually analyze discourse produced in supervised and artificial contexts (supervised description of images in artificial settings). Our study mends this situation. During our study, we observed that lexical and semantic deficits are indeed specific to AD patients. They can be observed through the poor lexical diversity. Even if these people are considered to have the communication skills, the language loss has already started (decreasing number of sentences, of words, of speaking time...). The analysis we propose also shows that AD patients is less detached from what they are saying. The speech becomes non-natural and one can wonder whether the AD patients can still manage their discourse or whether they just exploit the speech reflexes they learned previously in their lives. We think that the obtained results are interesting and that they can help the speech care of such people. Such care should start as soon as possible in order to maintain the communication skills as long as possible. Since the communication loss of the AD patients is going increasing with the evolution of the disease and causes their isolation, our study allows to better understand, at the level of language production, what are the specific language deficiencies of AD patients at early to moderate stages. Besides, the comparison of AD patients discourse with healthy people discourse provides a better analysis and evaluation, and also it provides a basis for the diagnosis and a better speech care. The conclusions of our study help to better understand the development of the disease. In future work, we plan to study larger corpora collected with additional participants within similar ecological context. The use of the NLP tools allows to systematically process larger amounts of data and to observe several specificities of the AD patients speech. Nevertheless, it is necessary to adapt some NLP tools to the processing of the data we use (to complete the morphological analysis with other morphological rules) and to use additional tools (*eg*, syntactic analysis). Complementary study of verbal and non-verbal (gesture) communication is an interesting issue. If similar data are available in other languages, the approach proposed can be applied to process them. Finally, the automatic categorization of speakers as AD-suffering or not and the automatic

diagnosis of this disease as early stages is another important perspective of the current work.

References

[1] Dubois, B. Actualités de la maladie d'alzheimer. *Annales Pharmaceutiques Françaises 2009*, **67**(2), 116-126.

[2] Joanette, Y., Kahlaoui, K., Champagne-Lavau, M. et Ska, B. *Troubles du langage et de la communication dans la maladie d'Alzheimer : description clinique et prise en charge 2006*, 223-241.

[3] Tran, T., Dasse, P., Letellier, L., Lubjinkowic, C., Thery, J. et Mackowiak, M. Les troubles du langage inauguraux et démence : étude des troubles lexicaux auprès de 28 patients au stade débutant de la maladie d'Alzheimer. *SHS Web of Conferences 2012*, **1,** 1659-1672.

[4] Traykov, L., Rigaud, A., Cesaro, P. et Boller, F. Le déficit neuropsychologique dans la maladie d'Alzheimer débutante. *L'Encéphale 2007*, **33**(1), 310-316.

[5] Fryer-Morand, M., Delsol, R., Nguyen, D. et Rabus, M. Le syndrome dysexécutif dans la maladie d'Alzheimer: à propos de 95 cas. *Neurologie-Psychiatrie-Gériatrie 2008*, 8, 23-29.

[6] Mackenzie, C. Adult spoken discourse: the influences of age and education. *International Journal of Language and Communication Disorders 2000*, **35**(2), 269-285.

[7] Feyereisen, P. et Hupet, M. *Parler et communiquer chez la personne âgée.* PUF, Paris 2002.

[8] Kynette, D. et Kemper, S. Aging and the loss of grammatical forms: A cross-sectional study of language performance. *Language and Communication 1986*, **6,** 65-72.

[9] Kemper, S., Rash, S., Kynette, D. et Norman, S. Telling stories: The structure of adults' narratives. *European Journal of Cognitive Psychology 1990*, **2** 205-228.

[10] Rousseau, T., De Saint-André, A. et Gatignol, P. Évaluation pragmatique de la communication des personnes âgées saines. *Neurologie Psychiatrie Gériatrie 2009*, **9,** 271-280.

[11] Croisile, B., Ska, B., Brabant, M., Duchene, A., Lepage, Y., Aimard, G. et Trillet, M. Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language 1996*, **53,** 1-19.

[12] Duong, A., Tardif, A. et Ska, B. Discourse about discourse: What is it and how does it progress in Alzheimer's disease? *Brain and cognition 2003*, **53,** 177-180.

[13] Berrewaerts, J., Hupet, M. et Feyereisen, P. Langage et démence : examen des capacités pragmatiques dans la maladie d'Alzheimer. *Revue de Neuropsychologie 2003*, **13**(2), 165-207.

[14] Ska, B. et Duong, A. Communication, discours et démence. *Psychologie et NeuroPsychiatrie du Vieillissement 2005*, **3**(2), 125-133.

[15] Lee, H. (2011). Vieillissement normal et maladie d'Alzheimer : analyse comparative de la narration semi-dirigée au niveau lexical. *In Colloque international sur les méthodes et analyses comparatives en sciences du langage*.

[16] Barras, C., Geoffrois, E., Wu, Z. et Liberman, M. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. *In Conference on Language Resources and Evaluation (LREC)*, 1373-1376.

[17] Schmid, H. Probabilistic part-of-speech tagging using decision trees. *In ICNMLP 1994*, 44-49, Manchester, UK.

[18] Namer, F. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues (TAL) 2000*, **41**(2), 523-547.

[19] Namer, F. *Morphologie, Lexique et TAL: l'analyseur DériF. TIC et Sciences cognitives*. Hermes Sciences Publishing, London 2009.

[20] Bove, R. *Analyse syntaxique automatique de l'oral: étude des disfluences*. Thèse de doctorat, Université d'Aix-Marseille I, Marseille 2008.

[21] Pallaud, B. Les amorces de mots comme faits antonymiques en langage oral. *Recherches sur le français parlé 2002*, **17,** 79-102.

[22] Gaspers, J., Thiele, K., Cimiano, P., Foltz, A., Stenneken, P. et Tscherepanow, M. An evaluation of measures to dissociate language and communication disorders from healthy controls using machine learning techniques. *In IHI 2012*, 209-218.