The "Corpus Anchise 320" and the analysis of conversations between healthcare workers and people with dementia

Nicola Benvenuti Università di Torino

Andrea Bolioli CELI

Alessio Bosca CELI

Alessandro Mazzei Università di Torino

Pietro Vigorelli Gruppo Anchise

Abstract

The aim of this research was to create the first Italian corpus of free conversations between healthcare professionals and people with dementia, in order to investigate specific linguistic phenomena from a computational point of view. Most of the previous researches on speech disorders of people with dementia have been based on qualitative analysis, or on the study of a few dozen cases executed in laboratory conditions, and not in spontaneous speech (in particular for the Italian language). The creation of the Corpus Anchise 320 aims to investigate Dementia language by providing a broader number of dialogues collected in ecological conditions. Automatic linguistic analysis can help healthcare professionals to understand some characteristics of the language used by patients and to implement effective dialogue strategies.[1]

Introduction

In this paper we will present the construction of the first annotated corpus of conversations between healthcare workers and people with dementia for Italian, called "Corpus Anchise 320", and the quantitative linguistic analysis we carried out. The aim of the project is twofold. On the one hand, we created a dataset of spoken dialogue transcriptions that is useful for research on the language of people with dementia. On the other hand, techniques typical of computational linguistics are applied to help doctors in assessing the state of the disease and implement effective dialogue strategies. Focusing attention on verbal exchanges between speakers is one of the cornerstones of the approach developed by the Anchise Group to support people with dementia and their caregivers, i.e. the "Enabling Approach" (Vigorelli 2018).

The paper is divided in 4 sections. Section 1 introduces the topic of Alzheimer's language Section 2 presents the recent researches and related works. In Section 3, the creation of the Corpus Anchise 320 will be discussed, which collects the transcripts and annotations of a set of dialogues between healthcare professionals and dementia patients carried out by the Anchise Group from 2007 until today, in Italian language. Section 4 will report the results of the computational linguistic analysis with the StanfordNLP library for Italian. The results obtained will be discussed to outline some of the peculiarities of the Dementia language. Section 5 concludes this paper with some final considerations.

1 The Alzheimer's language

Dementia refers to a series of symptoms that manifest in "*difficulties with memory, language, problem solving and other cognitive skills that affect a persons ability to perform everyday activities.*" (Alzheimer's-Association 2018, 368). These symptoms change over time and reflect the degree of neuronal damage in different parts of the brain. Alzheimer's disease (AD), a neurodegenerative brain disease, is the most common form of dementia. One of the most popular neuropsychological tests for assessing a patient's neurocognitive and functional status is still the *Mini-Mental Test*, designed by Folstein et al. (1975).

The first symptoms are memory loss or a state of frequent confusion. Alzheimer's disease, semantic dementia, aphasia and amnesia all share

a close link with lexical memory and therefore declarative memory, while they would leave grammar and procedural memory intact. Language would thus move between a structural component, formed by grammatical rules that are stabilized over the course of life and are preserved longer as a crystallized function; and a semantic component that would collapse more quickly because it requires a mnemonic and contextualized effort that makes the cognitive activity of the individual more complex. This dissociation is confirmed by studies on Alzheimer's language (Almor 1999), (Kempler 2008), (Bucks 2000) in which it has been amply demonstrated that one of the first symptoms is anomia, or the difficulty in finding the lexical target; as opposed to a good ability to construct the sentence up to the advanced stages of the disease. These deficits would then be compensated through linguistic strategies, such as the high use of pronouns, circumlocutions and passepartout words present in the speech of Alzheimer's patients: "*empty words ("things", "do", "he", "it", etc.) are successfully and relatively easily activated precisely because they are high in frequency and allow the patients to produce fluent and grammatical sentences in the presence of debilitating semantic deficits*" (Almor 1999, 205). In the more advanced stages of the disease, communication becomes increasingly problematic as Alzheimer's patients experience difficulties in understanding and constructing a coherent discourse: "*their narratives are often ripetitive with topic changes, unclear references, and lack of coherence and informativeness*" (Kempler 2008, 76).

2 Related works

The recent workshop on the creation of medical dialogues corpora (Bhatia et al. 2020) is a consequence of an increasing interest on this specific application field. The main reason of this interest is on the possibility of design and realize software applications which can assist professionals in medicine in their daily work in order to avoid errors: "*It is imperative to find a solution to minimize causes of such errors, via better tooling and visualization or by providing automated decision support assistants to medical practitioners.*". With this final aim, the creation of medical dialogues corpora can be seen as a first step toward the creation of a virtual medical assistant that can assist, speed-up, improve the capacities of medical practitioners.

As stated in (de la Fuente Garcia 2020) "*datasets containing both clinical information and spontaneous speech suitable for statistical learning are relatively scarce. In addition, speech data are often collected under different conditions, such as monologue and dialogue recording protocols.*" A notable example is the Carolina Conversations Collection (CCC), that is amongst the few spontaneous dialogue datasets available in English in the context of AD research. It is hosted and distributed by the Medical University of South Carolina (Pope 2011).

The study of AD language with computational methods is fairly recent, but a number of work showed the applicability of symbolic and statistic algorithms for the prediction of dementia and similar diseases (Karlekar et al. 2018, Mirheidari et al. 2019, Kong et al. 2019).

In (Karlekar et al. 2018) neural networks have been used on the publicly available DementiaBank dataset in order to predict Alzheimer's dementia of a patient starting from the language produced and annotated with the POS feature. They reached precision result between 80-85%. Interestingly, they showed that there is no significative difference between the prediction results by considering the gender.

In (Mirheidari et al. 2019) an automatic dementia detection system was presented, including a diarisation unit, an automatic speech recogniser, conversation analysis (CA) based acoustic and lexical feature extraction module and a machine learning classifier, in order to facilitate and improve screening procedures for dementia. They showed that using these features, they can obtain a high value of precision in detecting dementia for both a neurologist-patient and VirtualAgent-patient conversations.

In (Kong et al. 2019) neural networks on DementiaBank dataset have been used too. They reached precision results close to the state of art (80-85%), but they pointed out on the scalability of their neural methods that need less data. Moreover, they showed that "*the attention mechanism of the model manages to capture similar key concepts as the information unit features specified by human experts.*".

As for Italian language, in (Beltrami 2018) the participants (both healthy and cognitively impaired) were asked to answer to three specific tasks, i.e. the description of a drawing, details of a last dream and the description of a working day. The researchers investigated whether the analysis performed by Natural Language Processing

techniques could reveal alterations of the language performance in early cognitive decline.

3 The Corpus Anchise 320

Corpus Anchise 320 collects the transcripts of dialogues between healthcare professionals and patients carried out over the period from 2007 until today by the Anchise Group, an association of experts (doctors, psychologists, nurses, trainers) for the research, training and care of the elderly with dementia. The corpus consists of an unselected series of people diagnosed with dementia and not only those with an established diagnosis with specific criteria for Alzheimer's were included. For probabilistic reasons, most patients are affected by AD. The corpus contains 320 individual conversations resulting from transcription of about 15 minutes of dialogue for each patient in which the patient can speak freely with the health worker. This peculiarity is of considerable importance in a field of investigation that was mainly based on "*formal medical-psychological situations of the anamnestic investigation and the collection of tests*" (Lai 2000).

The corpus contains 20,588 turns of conversation, consisting of 10,193 turns of patients with dementia and 10,381 turns of health workers. The total number of tokens is 222,856 and the total number of types (different words) is 14,513. In the table below we present a small portion of one conversation.

7	**P**	*Eh ma mia figlia… è dura, è dura.* [Eh but my daughter... it's hard, it's hard.]
8	**O**	*Lo sa Marta che da quando abbiamo iniziato a vederci lei parla molto, molto meglio?* [You know, Marta, that since we started seeing each other you talk much, much better?]
9	**P**	*Ma a casa mia no! Loro non capiscono. Han detto che non capiscono niente…* [But not at home! They don't understand. They said they don't understand anything…]
10	**O**	*Forse sono loro che non capiscono.* [Maybe it is they who do not understand.]

Table 1: An excerpt from a conversation between a patient (P) and an health worker (O), with English translation (turn 7 to 10).

The corpus has been created in two phases. In the first phase, health professionals of Anchise Group created the audio recordings, transcribed portions of dialogues and annotated each transcription with a series of metadata with the aim of investigating the relationship between the language, age, sex and stage of dementia (MMSE[2] score). In the second phase, we collected the 320 transcriptions, we removed pragmalinguistic comments of health professionals, such as "[Touch the recorder]", "[Silence]", "[Laughs]", etc., and we analyzed and annotated the corpus as described in the following sections.

Corpus Anchise 320 has been built and archived according to EU General Data Protection Regulation (GDPR). Audio recording and transcriptions were made with the consent of the speaker, as far as possible, of the family member and of the head of the facility or department. Personal data have been anonymized. The dataset is not publicly available but it can be requested to the authors for research purposes.

4 Computational linguistic analysis of Corpus Anchise 320

In this Section we will discuss the results of the lexical analysis (3.1) and of the morphosyntactic analysis (3.2) carried out on the Corpus Anchise 320.

3.1 Lexical analysis

The Corpus Anchise 320 contains 222,856 tokens and 14,513 types. The relationship between types and tokens constitutes the *Types-Token Ratio* (TTR), which represents a type of index to calculate the lexical richness of a text (Torruella 2013). The number of tokens and types were subsequently calculated for the patient's total turns and the health worker's total turns. The results are shown in Table 2.

	Token	***Types***	***TTR***
Corpus Anchise	222.856	14.513	0,07
Patients	144.405	8.499	0,06
Health workers	78.451	6.014	0,08

Table 2: Token/types data.

[2] *Mini-Mental State Examination.*

TTR is low for both speakers. As for the patient, this trend is closely linked to Alzheimer's disease, in which "*the production of high-frequency words is relatively preserved while the production of low-frequency is impaired*" (Almor 1999, 204). As for the health professional, this trend reflects the Grice Principle of Cooperation between speakers in which it is necessary to conform the conversational contribution to what is required, when it occurs, by the accepted common intent or by the direction of the verbal exchange. Finally, if look at the number of tokens, we get that the patients speak more but with a poorer vocabulary relative to the lower lexical richness index than the sample of the health workers.

A frequency list was then created on the corpus sample of patients with dementia. The table 3 is the result of a pre-processing phase where 4 types of function words have been removed from the frequency list, i.e. adpositions, determiners, conjunctions and auxiliaries.

From the analysis of the data it emerges that the first 50 words in order of frequency cover 32% of the entire Corpus Anchise 320 and 49.4% of the patients' speech; the first 100 words cover 40.0% of the entire corpus and 61.8% of that of patients with dementia; the first 200 words cover 46.7% of the entire corpus and 72.1% of the words used by patients. This means that, on an expressive level, patients with the use of 200 words cover almost three quarters of all the vocabulary used in these conversations.

	Words	*Frequency*		*Words*	*Frequency*
1	non	3.954	10	adesso	687
2	sì	3.123	11	lì	666
3	mi	2.272	12	mia	639
4	io	2.063	13	qui	639
5	no	1.546	14	casa	637
6	eh	1.343	15	me	612
7	anche	1032	16	fare	577
8	bene	1007	17	lei	552
9	cosa	768	18	so	535

Table 3: Words frequency of patients with dementia.

The analysis of the words most used by patients diagnosed with dementia present in the Corpus Anchises 320 shows a high percentage of deictics, such as "io" ("me"), "qui" ("here"), "lì" ("there"), and the presence of semantically empty words, such as "cosa" ("thing") and "cose" ("things"). This attitude confirms the scientific research carried out so far on Alzheimer's language regarding word finding deficits: "*the earliest language deficits observed in DAT is anomia. (...) Semantically empty words are scattered throughout the DAT patient's utterances in place of content words, thereby maintaining fluency and sacrificing informational content.*" (Kempler 1991, 98). From the analysis of the first 100 words used, we note the presence of the words "casa" ("home") with 637 occurrences, "mamma" ("mother") with 394 occurrences, "marito" ("husband") with 190 occurrences, "figli" ("children") with 162 occurrences. As the corpus contains spontaneous speech, we can note that the most common topic is the patient's family.

3.2 Morphosyntactic analysis

The Corpus Anchise 320 was analyzed morpho-syntactically by means of the StanfordNLP library in Python language (Qi 2018). The default pre-trained neural model for Italian was used. Specifically, tokenization, lemmatization, POS tagging and Dependency parsing were carried out. These annotations, i.e. ID, Form, Lemma, POS, FEATS, HEAD, DEPREL, were organized according to the CoNLL-U format (Zeman 2018, Bosco 2014). A linguist reviewed the automatic annotations.

ID	TOKEN	XPOS	LEMMA	FEATS	HEAD	DEPREL
3	che	PRON	che	...	4	nsubj
4	fa	VERB	fare	...	2	acl:relcl
5	un	DET	uno	...	6	det
6	po'	ADV	poco	...	7	advmod
7	fatica	NOUN	fatica	...	4	obj
8	a	ADP	a	...	9	mark
9	parlare	VERB	parlare	...	4	xcomp

Table 4: An excerpt from the annotated corpus. Features are not shown due to space constraints.

The analysis of the linguistic data of patients suffering from dementia was made using both the LIP[3] corpus (De Mauro 1993, 155) and the speech corpus of healthcare professionals as a reference.

[3] *Lessico di frequenza dell'italiano parlato*

The analysis of the percentages of occurrence of the parts of speech, in the patient corpus sample, reveals a superior use of pronouns and adverbs both with respect to LIP and with respect to the corpus of health workers. With reference to the LIP, the use of pronouns records 10.9% of occurrence, while the use of adverbs 10.1%. If we compare these data with the rates of occurrence in the patients' speech (Table 5), 13.9% frequency for pronouns and 14.2% for adverbs respectively, we notice a notable difference. Furthermore, these two indices, when added together, are 1.7 percentage points higher than the health workers' speech (ADV 13,2%, PRON 13,2 %). This trend would confirm what was said in the analysis relating to word frequency, i.e. the difficulty for patients to access the lexicon and therefore to compensate for this deficit with the use of deictics, closely linked to the context. If we cross these data with the rate of names used by patients (1.6 percentage points lower than the corpus of the health workers, NOUN 13,2%), we can deduce that the patient implements a real compensatory strategy linked to the impairment of access to semantic memory. A significant difference is also present with the LIP, which records a rate of names of 15.7% against 11.6% of the corpus relating to patients with dementia.

	Patients	*%*
ADJ	4.843	3,3
ADP	11.177	7,7
ADV	20.560	14,2
AUX	10.586	7,3
CCONJ	5.562	3,8
DET	14.200	9,8
INTJ	7.383	5,1
NOUN	16.787	11,6
NUM	1.145	0,7
PRON	20.118	13,9
PROPN	1.799	1,2
SCONJ	5.078	3,5
VERB	24.749	17,1
X	418	0,3
TOT.	144.405	

Table 5: Percentages of occurrence of the parts of speech.

At the morphosyntactic level, it is known in the literature that Alzheimer's patients do not suffer from serious deficits in the construction of the sentence: "*sentence production in DAT is characterized by intact morphosyntactic structure (i.e., subject verb agreement, well formed plural and tense markings)*", (Kempler 2008, 75). However, some linguistic phenomena that emerged from the analysis of the occurrences of the verbal system could be linked to a spatial-temporal disorientation characteristic of Alzheimer's disease (Macrì 2016). This disorientation is reflected in the massive use of the indicative mode present in 95.9% of cases (Table 6). The use of the subjunctive and conditional modes appears to be almost minimal with percentages that are around 1%. This tendency could be paraphrased in terms of cognitive work, since the two verbal modes require both the ability to imagine possible worlds and - at the level of sentence construction - of conjugation and temporal concordance.

Verb form	**Fin**	**Inf**	**Part**	**Ger**
	25.771	4.655	4.751	158
	(72,9%)	(13,2%)	(13,4%)	(0,4%)
Mood	**Ind**	**Sub**	**Imp**	**Cnd**
	24.702	452	326	285
	(95,9%)	(1,8%)	(1,2%)	(1,1%)
Tense	**Pres**	**Past**	**Imp**	**Fut**
	21.958	4.800	3.459	304
	(71,9%)	(15,7%)	(11,33%)	(0,9%)

Table 6:: Percentages of occurrence of the verbal system.

5 Conclusion and further research

In this paper we presented the first Italian corpus of conversations between healthcare professionals and people with dementia, called "Corpus Anchise 320". The study of this corpus with computational linguistic analysis confirmed some characteristics of the language of people with dementia, such as the reduction in the rate of names and the increase in deictics. Corpus Anchise 320 has been built and archived according to GDPR. It is not publicly available but it can be requested to the authors for research purposes.

The large number of the sample (320 conversations) and the use of computational analysis will make it possible to identify indicators of pathological language to be used in the preclinical phase, to trace the change in the linguistic abilities of people with dementia as the disease progresses, to put in relation the characteristics of the pathological language with a

series of metalinguistic data such as age, sex and degree of dementia. The corpus will be increased in the coming months with the addition and annotation of other transcripts of dialogues of people with dementia.

References

Alzheimer's-Association. "2018 Alzheimer's disease facts and figures." *Alzheimer's and Dementia*, 2018: 367-429.

Almor, A., Kempler, D., MacDonald, M. C., Andersen, E. S., & Tyler, L. K. (1999). Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease. *Brain and language*, 67(3), 202-227.

Associazione Gruppo Anchise. http://www.formalzheimer.it/.

P. Bhatia, S. Lin, R. Gangadharaiah, B. Wallace, I. Shafran, C. Shivade, N. Du, and M. Diab, editors. Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, Online, July 2020. Association for Computational Linguistics

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., & Calzà, L. (2018). Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?. *Frontiers in aging neuroscience*, 10, 369.

Bucks, R. S., Singh, S., Cuerden, J. M., & Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, *14*(1), 71-91.

Bosco, C., Montemagni, S., Simi, M. (2013). Converting Italian Treebanks: Towards an Italian Stanford Treebanks.

Bosco, C., Dell'Orletta, F., Montemagni, S., Sanguinetti, M., & Simi, M. (2014). The EVALITA 2014 dependency parsing task. In *EVALITA 2014 Evaluation of NLP and Speech Tools for Italian* (pp. 1-8). Pisa University Press.

de la Fuente Garcia, S., Haider, F., & Luz, S. (2020). Cross-corpus Feature Learning between Spontaneous Monologue and Dialogue for Automatic Classification of Alzheimer's Dementia Speech. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 5851-5855). IEEE.

De Mauro, T., Mancini, F., Vedovelli, M., Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato.* Milano: Etaslibri.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, *12*(3), 189-198.

S. Karlekar, T. Niu, and M. Bansal. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 701–707, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Kempler, D. (1991). Language Changes in Dementia of Alzheimer Type. In *Dementia and Communication,* by Rosemary Lubinsky, 98-114. Philadelphia: B.C. Decker, Inc.

Kempler, D., & Goral, M. (2008). Language and dementia: Neuropsychological aspects. *Annual review of applied linguistics*, 28, 73.

W. Kong, H. Jang, G. Carenini, and T. Field. A neural model for predicting dementia from language. volume 106 of Proceedings of Machine Learning Research, pages 270–286, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR.

Lai, G. (2000). Conversazioni con l'Alzheimer. *Prospettive sociali e sanitarie*, 18, 2-5.

Macrì, A. (2016). La lingua della demenza di Alzheimer. Analisi linguistica del parlato spontaneo. In *Le lingue della malattia*, 329-424. Milano: Mimesis Edizioni.

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., & Christensen, H. (2019). Dementia

detection using automatic analysis of conversations. Computer Speech & Language, 53, 65-79.

Pope, C., & Davis, B. H. (2011). Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1), 143-161.

Qi, P., Dozat, T., Yuhao Zhang, Y., & Manning, C. D. (2018). Universal Dependency Parsing from Scratch In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 160-170.

Torruella, J.; Capsada, R. (2013). "Lexical Statistics and Tipological Structures: A Measure of Lexical Richness." *Procedia - Social and Behavioral Sciences*, pp. 447-454.

Vigorelli, P. (2018). *Alzheimer, come parlare e comunicare nella vita quotidiana nonostante la malattia.* Milano: Franco Angeli Editore.

Vigorelli, P. (2004). *La conversazione possibile con il malato Alzheimer*. Milano: Franco Angeli Editore.

Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., ... & Petrov, S. (2018, October). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies* (pp. 1-21).