JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

The FRA Corpus from the DISease and AGEing Project

Francesca M. Dovetto[1]* & Francesca Marra[1]

[1]Università degli Studi di Napoli Federico II, Italy

Abstract. The FRA corpus, from the DISAGE project (University of Naples Federico II), is an Italian speech corpus of elderly people with and without neurodegenerative pathologies. The project's main aim is to provide an Italian tool useful to observe language in pathological aging. Therefore, the project enrolled 20 patients diagnosed with Alzheimer's Disease (AD), 20 individuals suffering from Mild Cognitive Impairment (MCI), and 20 Healthy Controls (HC) balanced with respect to sex, age, and education. Participants underwent neuro-psychological testing (MMSE; Raven's Matrices) and only mild AD and amnestic MCI were recruited. The speech corpus is semi-spontaneous and dialogic: speech samples were collected via a three-stage interview designed to encourage the production of descriptive, conversational and narrative speech. Currently, 18 out of the 60 recordings collected have been manually and orthographically transcribed. Data is available on request.

Keywords. speech corpora, pathological speech, aging, Alzheimer's Disease, Mild Cognitive Impairment

A peer-reviewed contribution to *Journal for Language and Aging Research* (JLAR).
Submitted: 2024-08-31
Accepted: 2024-10-08
Published: 2024-10-21
DOI: 10.15460/jlar.2024.2.2.1522

1 Introduction

FRA is an Italian semi-spontaneous corpus of speech in pathological aging. It is the outcome of the *DISease and AGEing* Project[2] (*Language Corpora for the Assessment of Spontaneous Speech in Alzheimer's Disease and Mild Cognitive Impairment*), which has as its main aim to develop an Italian speech dataset of elderly people with and

*. Corresponding author, dovetto@unina.it

Please cite as Dovetto, Francesca M. and Francesca Marra. 2024. "The FRA Corpus from the DISease and AGEing Project." *Journal of Language and Aging Research* 2(2): 64–78. 10.15460/JLAR.2024.2.2.1522.

2. DISAGE is a three-year (2023–2025) project of the University Federico II of Naples, co-funded by the University of Naples Federico II and Banca Intesa San Paolo.

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

without neurodegenerative pathologies useful to support clinicians in pinpointing reliable Italian linguistic markers for the identification of those individuals with Mild Cognitive Impairment (MCI) who will likely develop Alzheimer's Disease (AD). Therefore, DISAGE moves towards a multidisciplinary and interdisciplinary approach and involves both humanities scholars, namely linguists, and those with medical, clinical, and psychopathological expertise.

Speech samples of patients diagnosed with AD, individuals with MCI, and healthy controls have been collected via a three-stage interview. All participants underwent neuro-psychological screening (MMSE, Folstein, Folstein, and McHugh 1975; Raven Colorated Matrices, Raven 1938). Only mild AD (MMSE cut-off 19 / 30) and amnestic MCI were recruited. Interviews were audiotaped through an H4nPro Zoom recorder (44 kHz / 16 bit), transcribed following CLIPS transcribing norms (Savy 2006), and analyzed in PRAAT (Boersma and Weenink 2024). Materials (i. e., audiotapes .wav, TextGrid .txt, transcriptions .txt, subject information including MMSE value, age, group, etc.) are available on request.

1.1 Approach to Language and Aging Research

Aging, whether healthy or pathological, results in language changes due to both physiology and cognition: reduced informativeness of utterances (Pistono et al. 2017; Hilviu et al. 2022), difficulty in lexical retrieval (Wright 2016; Folia et al. 2022), and phonetic parameter fluctuations (Brückl and Sendlemeier 2003; Pistono et al. 2016) represent only a few instances of age-related variations. As expected, such language features are also found in elderly people suffering from pathological conditions. It is then fair to ask whether language weakening paths differ between healthy and pathological aging.

It is particularly worth observing what happens in neurodegenerative pathologies, such as (AD) and (MCI), as they have a growing impact on public health in terms of incidence, prevalence, mortality rates, and costs of care (Alzheimer's Association 2024). Because of AD's massive side effects, there is an urgent need for reliable ways to formulate its early and accurate diagnosis: language assessment, by being relatively low-effort, inexpensive, and non-invasive, can play a crucial role in the differential diagnosis of this pathology (Lemos et al. 2016). Unfortunately, language-based AD diagnosis is still limited because of the lack of available language data from AD.

Moreover, since not all individuals with MCI develop AD (Jicha, Parisi, and Dickson 2006), it would be worthwhile to establish reliable markers for the identification of those who will likely do so; that is why more studies are needed on the prodromal stage of AD, namely MCI. A few previous attempts have been made to collect linguistic corpora of people with MCI and AD (Fleming 2014; Fraser, Meltzer, and Rudzicz 2016; Beltrami et al. 2016; Duboisdindien et al. 2020; López-de-Ipiña, Martinez-de-Lizarduy, and Calvo 2020). However, such corpora are usually not freely available, they do not use homogeneous methods of data collection and annotation, nor do they consider the potential impact of spontaneous speech typical phenomena on language production. Furthermore, there is a lack of speech corpora in healthy aging, which, on the other hand, is necessary to accurately identify pathological linguistic features.

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

2 Methods

60 informants (20 AD; 20 MCI; 20 healthy controls(HC)), aged between 60 and 85 years old, all native speakers of Italian with a right-hand preference and at least 5 years of formal education, were recruited. 18 out of the 60 interviews collected (6 AD, 6 MCI and 6 HC; sex ratio 1:1) have been manually transcribed in a .txt format. The mean age of the sample transcribed is 74 y.o. (M.A. AD 73 years and 8 months; M.A. MCI 73 years and 5 months; M.A. HC 74 years and 4 months). Participants were recruited by a Research Unit of the Humanities Department with the aid of the Dementia and Cognitive Disturbances Center (Neurology Department, Federico II University Hospital, Naples) and Hematology Unit (Federico II University Hospital, Naples).

Inclusion criteria for HC are normal general cognition, as indicated by Mini-Mental State Examination scores above the normality cut-off (>23.8) (Folia et al. 2022), normal performance on the Raven's Coloured Progressive Matrices (Raven 1938), no history of psychiatric or neurological illness, and absence of hearing or visual loss. MCI and AD patients have all been submitted to formal clinical, neuropsychological, behavioural, functional, and linguistic assessment. Diagnosis has been performed according to established international criteria (Petersen 2004)(Petersen 2004; Matthews et al. 2008; Petersen et al. 2014). AD, MCI, and HC are balanced for age, sex, and origin and a minimum of 5 years of education was obtained by all.

The corpus covers the Italian language and Neapolitan dialect. Data were collected via a three-stage interview designed to elicit descriptive, conversational, and narrative speech; in particular, the ENPA picture description task was employed (Capasso and Miceli 2001), a conversation about daily topics of everyday life was engaged in (i./,e., daily routines, hobbies, recipes), and the fairy tale Little Red Riding Hood was elicited.

After the three-task procedure was developed, participants were recruited from among the patients of the Neurology and Hematology Departments (Federico II University, Naples). Their caregivers have been involved as well. Both patients and caregivers granted their written consent in order to take part in the study. Interviews were carried out inside the Frederico II University Hospital (Naples) in a windowed, non-soundproof room and they were audiotaped with an H4nPro Zoom recorder (44 kHz / 16 bit). The recordings were then uploaded to a 2017 MacBook Pro to be annotated via PRAAT (Boersma and Weenink 2024), transcribed following CLIPS transcribing norms (Savy 2006), and analyzed in PRAAT. Data were pseudonymized in both text and media files by removing personally identifiable information from the data set. The technique employed was that of data masking, which consists in hiding personally identifiable information via the <name> tag in .txt files (Table 1) and by adding noise in .wav files. Each transcription was first performed by an expert transcriber and then revised by a second expert transcriber.

In order to allow data comparison, the transcription method chosen follows that of analogous Italian corpora of neurotypical spontaneous speech. In particular, FRA takes into consideration the CLIPS ([Corpora and Lexicon of Written and Spoken Italian) (Albano Leoni, Sobrero, and Paoloni 2007) and the CIPPS (Italian Corpus of Patients Affected with Schizophrenia) (Dovetto and Gemelli 2013) orthographic transcription guidelines (Savy 2006; Dovetto and Gemelli 2013), which provide useful tools for observing verbal dysfluencies (e./,g., false starts, repetitions, clippings, slips, primary interjections) as well as non-verbal vocalizations (e./,g., throat clearing or tut-tuts, inspirations, laughs). Filled and silent pauses, interjections, and vocaliza-

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

tions are also annotated. Indeed, the project aims at focusing on elements defined as perilinguistic, i./,e., phonemically and morphologically brief and irrelevant items, and therefore weakly or not-at-all framed within the language system, as well as on those defined as paralinguistic, i./,e., contextual elements such as gestures and voice volume[3] (De Mauro 2008).

ID	Group	Sex	Age	MMSE	Education (yrs)	Origin
maB	AD	F	70;02	25	5	Campania
maC	AD	M	78;01	23	13	Campania
maE	AD	F	67;09	19	5	Campania
maH	AD	M	79	20	8	Campania
maL	AD	M	74;08	26	13	Campania
maQ	AD	F	73;02	23	13	Campania
mciE	MCI	F	78;04	23	5	Campania
mciF	MCI	M	72;05	24	13	Campania
mciG	MCI	M	75;10	28	8	Campania
mciL	MCI	F	72	29	5	Campania
mciQ	MCI	F	68	30	13	Campania
mciR	MCI	M	74;07	27	8	Campania
hcA	HC	M	78	29	13	Campania
hcB	HC	F	65;02	28	5	Campania
hcE	HC	M	71;08	28	13	Campania
hcH	HC	F	76;04	29	13	Campania
hcS	HC	F	79;02	25	5	Campania
hcT	HC	M	78;09	28	8	Campania

Table 1: Sociolinguistic data

3 Corpus description

For open access to FRA corpus, researchers can visit the site offered by the LiSa Lab (Lingua e Salute) of LUPT (Federico II University, Naples). At present, the CIPPS corpus's orthographic transcription and the projects' presentation of Italian speech corpora to be published (e./,g., language corpora of spontaneous speech in healthy and pathological aging) are available for consultation at that location. The FRA corpus will also be uploaded as soon as the transcriptions are completed.

The FRA corpus, as part of the *DISAGE* project of the University of Naples Federico II, is co-funded by the University Federico II of Naples and Banca Intesa San Paolo. The data collection was carried out under the scientific responsibility of PI Francesca M. Dovetto, and performed by Francesca Marra of the Humanities Department of the Federico II University, in collaboration with the Neurology and Hematology Departments of the Federico II University Hospital (Naples). In particular, AD and MCI

3. The CLIPS transcription system (Savy 2006) makes it possible to specify whether an utterance is spoken aloud or in a whisper and, in the notes field, to indicate any gestures that accompany speech.

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

patients and controls were recruited under the responsibility of Elena Salvatore (Advanced Biomedical Sciences Department, Federico II University) and HC recruitment was carried out in collaboration with the Hematology Department (Federico II University).

Currently, the corpus consists of the interviews of 18 participants with at least 5 years of formal education, divided into three groups (6 AD, 6 MCI and 6 HC) and balanced with respect to sex, age and origin; the attendees' sociolinguistic data are summarized in Table 1.

Each interview comprises 3 parts: picture description, conversation, and narration (see Section 3.2). At this point, 54 files, including 18 text files (.txt), 18 audio files (.wav), and 18 textgrids (.txt) are available on request. The total length of the transcribed sample is 3 hours and 36 minutes of speech, corresponding to 25,787 tokens and 3,801 types. Altogether, the total amount of conversational turns of the corpus is 1,491 (Table 2).

ID	Group	Turns
maB	AD	113
maC	AD	127
maE	AD	102
maH	AD	84
maL	AD	57
maQ	AD	48
mciE	MCI	152
mciF	MCI	190
mciG	MCI	138
mciL	MCI	39
mciQ	MCI	43
mciR	MCI	53
hcA	HC	46
hcB	HC	71
hcE	HC	39
hcH	HC	64
hcS	HC	45
hcT	HC	80

Table 2: Conversational turns per participant

Tags frequently used in the orthographic transcription are summarized in Table 3.

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

Tag	**Phenomenon**
<name>	data masking tag
<unclear>	unintellible word
<vocal>	vocalization
<time>	a pause due to an overlap in dialogic turns
<NOISE>	presence of sudden noise
# . . . #	overlap
{<dialect> . . . </dialect>}	dialectal expression between curly brackets
{<repetition> . . . <>}	fragment, word or syntagm repetition between curly brackets
{<foreign word> . . . <word>}	foreign word between curly brackets
{<NOISE> . . . <>}	fragments, syllables, words and/or utterances surrounded by noise between curly brackets

Table 3: Tags employed in the transcriptions

3.1 Transcription and annotation type

Orthographic and manual transcription was performed in a .txt format; dialogical turns plus inter-turn and intra-turn pauses were annotated in PRAAT on dedicated tiers: G/F tier for Giver's and Follower's dialogical turns, G/F pauses tier for Giver's and Follower's silent pauses, and Inter-turn tier for silent pauses between different speakers' turns. The transcription employed, following the CLIPS method, is faithful to real produced speech: it includes semi-lexical, non-lexical, and non-verbal phenomena, and tags dialectal expressions and reports overlaps. Silent pauses' duration is also mentioned. Furthermore, the speech flow is segmented into dialogic turns, a notion based on the semantic-pragmatic coherence within the production of the speaker.[4] Tags for semi-lexical linguistic elements, as well as verbal non-lexical and non-verbal vocal phenomena, are reported in Table 5.

4. Turn segmentation is based on the definition of a turn used in CLIPS, i./,e., the turn-taking by one of the two interlocutors, whether that person interrupts the turn of the other speaker or overlaps with it without necessarily constituting an interruption. The turn thus represents the gaining of ground by one of the two interlocutors, which can either interrupt the turn of the other locutor or overlap the latter, and which continues as long as there is semantic-pragmatic coherence (see Savy 2006, 11).

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

Category	**Symbol / Tag**	**Description**
Semi-lexical symbol	+	Follows a fragmented word
Semi-lexical symbol	_	Designates an inside-word interruption
Semi-lexical symbol	*	Precedes a non-word and / or neologisms
Semi-lexical symbol	/	Indicates a false start
Non-lexical / non-verbal tag	<pause dur="... s/>	Silent pause duration in seconds
Non-lexical / non-verbal tag	<eeh>, <ehm>	Filled pause
Non-lexical / non-verbal tag	<vv>, <cc>	A vowel / consonant prolongation
Non-lexical / non-verbal tag	<ah>, <eh>, <beh>, ...	Interjections
Non-lexical / non-verbal tag	<laugh>, <cough>, <breath>, ...	Vocal non-verbal phenomena

Table 4: Symbols and tags employed in the transcriptions

Sound-text alignment and POS tagging of the FRA corpus will also be provided. Available scripts in PRAAT can be used to extract data from the TextGrids, such as the duration logger Script (DiCanio 2011), which is useful for drawing out the duration of all annotated silent pauses.

Each transcription is preceded by an informational table reporting data on the material, speakers, recording, and transcription (Table 5). The label of each recording and of the associated audio file reflects the main information contained in Table 6.[5]

5. As an example, DGmaB01N designates recording 1 of the AD (Italian ma) dialogic corpus patient B, regional variety Neapolitan.

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

Category	**Label**	**Reference**
text_inf	MAP	The speaker (maA, mciB, hcC, ...)
text_inf	NdD	Dialogue number (01, 02, 03, ...)
text_inf	REG	Regional variety (N)
speakers_inf	INp1	Name, Surname (initials), sex, age, place of birth of the Giver
speakers_inf	INp2	Name, Surname (initials), sex, age, place of birth of the Follower
speakers_inf	INp3	Other locutors, if applicable)
recording_inf	TYP	Type of recording (DAT, video tape, audio tape, acquisition with PC, etc.)
recording_inf	LOC	Place of recording
recording_inf	DAT	Date of recording
recording_inf	DUR	Length of recording
recording_inf	CON	General conditions during recording
transcription_inf	CMT	Comments
transcription_inf	Nst	Number of turns

Table 5: Labels of recordings and associated audio files, modified by Savy (2006)

3.2 Stimulus material

- Phase 1: Picture description task
  Observe this picture carefully and then describe it to me. Additional questions:
  Who are the men in the background? What are they doing?
  What is/are the man on the armchair/the kids on the floor doing?
  Which objects are in the image?

- Phase 2: Conversation

  1. Watching television
     Do you watch TV?
     Which are your favorite TV shows?
     Can you describe a movie you have seen recently?

  2. Daily activities
     How do you usually spend your day?
     What did you do this morning?
     What are your favorite activities?

  3. Recipe
     Can you cook?
     Which is your signature dish?
     Can you describe its recipe to me?

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

- Phase 3: Narrative speech
  Do you remember the fairy tale of Little Red Riding Hood? Can you please tell me about it?

4 Reuse potential

The FRA corpus is an essential tool for the study of pathologies that arise in senescence, such as MCI, AD, or Parkinson's Disease. Studying healthy elderly people's language will make it possible to prodromally detect the altered linguistic manifestations generated by a neurological pathology at the initial stage of the disease, thus providing the chance of defining linguistic predictors of cognitive impairment. By the end of the project in October 2025, the pseudonymised FRA transcription files will distributed as open access. Textgrid and audio files will only be available upon request.

5 Ethical considerations

Data were pseudonymized by removing personally identifiable information from files. Depseudonymization of data is forbidden. Scholars who request access to the FRA corpus will not have access to the participants' sensitive data. In addition, scholars will have to consent not to release the transcriptions to third parties and to use them for scientific purposes only.

The DISAGE project was submitted to a review board for ethical approval as part of the PRIN Project 2022 *Senectus Ipsa Morbus* (SIM: Spontaneous Speech in Healthy Aging).[6] Participants signed for written informed consent to the participation of the study and data processing, which guarantees the confidentiality of the data. The data obtained will remain pseudonymous and may only be used for scientific purposes in accordance with DL 196/2003 on the processing of personal data.

6 Acknowledgments

We are grateful to Miriana Migliaccio and Alessia Salemme for their support in patient recruitment and neuropsychological testing. We also thank a second expert transcriber, Sundra Sorrentino, for her revisions.

7 Example and case studies

The FRA data may be employed for several purposes, such as phonetic, lexical, morphosyntactic, and pragmatic analyses. As an example, a preliminary analysis was conducted on dysfluencies ahead of the CLARe 6 conference, to assess whether the parameter of the duration of silent pauses could be used to distinguish between AD, MCI, and HC individuals. Our TextGrids, manually annotated, present separate Tiers

6. The Ethics Committee for Research with Human Subjects in Non-Biomedical Fields (Comitato Etico per la Ricerca con Soggetti Umani in campo non biomedico, CERSUB) of the University of Naples Federico II approved the project by means of protocol PG/2024/0011629 of January 29, 2024.

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

for Giver's and Follower's silent pauses; this allowed us to automatically extract the duration of all Follower's silent pauses by running the duration logger script in PRAAT. Subsequently, the average pause durations of AD, MCI, and HC patients were compared. It came out that silent pauses' duration varies according to the task type and that longer than 1 second silent pauses may help distinguish the three experimental groups, in agreement with previous studies on Italian (Dovetto et al. 2024).

Our data may also be employed to observe dysfluencies' typology and collocation; indeed, dysfluencies such as silent and filled pauses, prolongations, and repetitions are all tagged, and therefore easy to extract, in the transcriptions. Here we summarize some examples of dysfluencies' typology and collocation. From a morphosyntactic perspective, dysfluencies, marked by double slash in the examples, can be external as in (1), or internal as in (2); if external, they could separate independent clauses as in (3) or main and dependent clauses as in (4):

(1) È amp+ / andata col<ll> paniere nel bosco <pause dur="0,412s"/> e<ee> nel bosco ha trovato<oo> <inspiration> Cappuccett+

è *and-at-a* *co-l* *paniere* *ne-l* *bosco*
AUX go-PTCP.F INS-DET basket LOC-DET woods

'she went with the basket into the woods'

e *ne-l* *bosco* *ha* *trov-at-o* *il* *lupo*
and LOC-DET woods AUX find-PTCP.F DET wolf

'and into the woods, she found the wolf'

(DGmaB01N)

(2) {<repetition> mia<aa> <pause dur="0,597s"/> mia </repetition>} zia

mia *mia* *zia*
POSS POSS aunt

'my aunt'

(DGmaL01N)

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

(3) Cappuccetto Rosso era una bambina <pause dur="0,705s"/> <inspiration> andava a scuola

CappuccettoRosso era *una bambina*
RedRidingHood be-3SG.COP.IPFV DET child

'RedRidingHood was a child'

andava *a* *scuola*
go-3SG.IPFV ALL school

'she attended school'

(DGmaL01N)

(4) nel bosco non ti fermare <inspiration> <eeh> <pause dur="0,240 s"/> {<dialect> 'nsomm' a <pp>parlà cu nnisciun' </dialect>}

nel *bosco* *non* *ti* *fermare*
LOC-DET wood NEG 2SG stop-INF

'into the woods don't stop'

a *parlare* *con* *nessuno*
to talk-INF COM anyone

'and talk to anybody'

(DGmciF01N)

Some phenomena may help lexical retrieval. That is the case of hesitation phenomena in (5), which precedes a content word.

(5) e poi alla fine il lupo mangiò<oo> <eeh> anche<ee> Cappuccetto

e *poi* *il* *lupo* *mangiò* *anche* *Cappuccetto*
and ADV DET wolf eat-3SG.PST ADV LittleHood

'then the wolf also ate LittleRedRidingHood'

(DGmciQ01N)

And last, dysfluencies may help the speaker in repairing errors, as in (6).

(6) E incontra il le+ / <eh> il lupo

e *incontra* *il* *le* *il* *lupo*
and meet-3SG.PRS DET lio DET wolf

'and she meets the wolf'

(DGmaB01N)

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1522

7.1 Little Red Riding Hood

The following transscript shows an example of the transcript file format used. This is the narration of the story *Little Red Riding Hood* narrated by DGmaB01N.

```
G#89:    <inspiration> Va bene <pause dur="0,226s"/> <lip smacking>
         <inspiration> le faccio un' altra domanda <pause dur="0,398s"/>
         lei conosce la storia di Cappuccetto Rosso ?
         <note> The Follower produces the discursive marker "mh" during
         the <pause dur="0,398s"/> pause </note>
         <pause dur="1,309s"/>
F#90:    <lip smacking> Cappuccetto Rosso che va nel bosco ?
G#91:    Sì
F#92:    E *<cc>contro il le+ / <eeh> <repetition> il lupo <pause
         dur="0,254s"/> il lupo </repetition> cattivo <pause
         dur="0,475s"/> <inspiration> <dialect> e <pp>po' nun<nn> è
         </dialect> che mi<ii> impegno troppo <dialect> chessa </dialect>
         Cappuccetto mi <dialect> aggio </dialect> dimenticato un po' con
         la testa
         <note> "<cc>contro" stands for "incontra"; the Giver produces
         the discursive marker "sì" during the <pause dur="0,254s"/>
         pause </note>
         <pause dur="0,710s"/>
G#93:    Ho capito <inspiration> quindi <pause dur="1,293s"/> <lip
         smacking> <pause dur="0,697s"/> possiamo provare a raccontarla
         da capo
F#94:    <lip smacking> <eh> #<G#95> <eh>#
G#95:    #<F#94> Allora# <pause dur="0,417s"/> chi era Cappuccetto
         Rosso ?
F#96:    È una bimba ch+ / che vivev+ con la nonna
         <pause dur="1,031s"/>
G#97:    E che cosa le è successo ?
F#98:    È amp+ / andata col<ll> paniere nel bosco <pause dur="0,412s"/>
         e<ee> nel bosco ha trovato<oo> <inspiration> Cappuccett+ <eeh>
         / <ll>le+ / <dialect> comm' s' chiamm' </dialect> / il lupo
         <pause dur="0,514s"/> <lip smacking> il lupo cattivo <dialect>
         e <repetition> <mm>mo mo </repetition> nun me ricord' 'a
         </dialect> <inspiration> <pause dur="0,288s"/> / *memorizzi
         <dialect> chissu fatt' 'stu lup' cattiv' ch' ha fatt' po'
         </dialect> <pause dur="0,428s"/> ecco <eh> <laugh>
         <note> the Giver produces the discursive marker "ho capito"
         during the <pause dur="0,428s"/> pause </note>
         <pause dur="0,488s"/>
G#99:    Il lupo cattivo la segue <pause dur="0,474s"/> nel #<F#100>
         bosco#
F#100:   #<G#101> <eh> <unclear># Mo per esempio non mi riesco
         <repetition> a<aa> a </repetition> ricordare l' attre<ee>+
         <inspiration> / *attravessamento che ha fatto
         <pause dur="0,332s"/>
G#101:   Ho capito <pause dur="0,412s"/> quindi non si ricorda come va a
         finire la storia ?
F#102:   <pause dur="0,490s"/> No
```

Funding

The University of Naples Federico II and Banca Intesa San Paolo co-funded the research.

Ethics statement

The DISAGE project has been submitted to a review board for ethical approval as part of the PRIN Project 2022 *Senectus Ipsa Morbus* (SIM: Spontaneous Speech in Healthy Aging). The Ethics Committee for Research with Human Subjects in Non-Biomedical Fields (Comitato Etico per la Ricerca con Soggetti Umani in campo non biomedico – CERSUB) of the University of Naples Federico II approved the project by means of protocol PG/2024/0011629 of January 29, 2024.

Conflict of interest

The authors have no conflicts of interest to declare.

References

Albano Leoni, Federico, Alberto A. Sobrero, and Andrea Paoloni. 2007. "Corpora e lessici di italiano parlato e scritta CLIPS." *Bolletino di Italianistica,* 121–148. https://doi.org/10.7367/71826.

Alzheimer's Association. 2024. *Alzheimer's Disease facts and figures.* Chicago: Alzheimer's Association. https://www.alz.org/alzheimers-dementia/facts-figures.

Beltrami, Daniela, Laura Calzà, Gloria Gagliardi, Enrico Ghidoni, Norino Marcello, Rema Rossini Favretti, and Fabio Tamburini. 2016. "Automatic identification of mild cognitive impairment through the analysis of Italian spontaneous speech productions." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16),* edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, et al., 2086–2093. Portoroz: European Language Resources Association.

Boersma, Paul, and David Weenink. 2024. *Praat: Doing phonetics by computer [Computer Program].* V. 6.4.17. http://www.praat.org/.

Brückl, Markus, and Walter F. Sendlemeier. 2003. "Aging female voices: An acoustic and perceptive analysis." In *Proceedings of VOQUAL '03.* Geneva.

Capasso, Rita, and Gabriele Miceli. 2001. *Esame neuropsicologico per l'afasia.* Milan: Springer.

De Mauro, Tullio. 2008. *Lezioni di linguistica teorica.* Rome: Laterza.

DiCanio, Christian. 2011. *Duration script for PRAAT.* V. 2.0. Lyon & Buffalo: Laboratoire Dynamique du Langage & University at Buffalo. https://www.acsu.buffalo.edu/~cdicanio/scripts/Get_duration_2.0.praat.

Dovetto, Francesca M., and Monica Gemelli. 2013. *Il parlar matto: Schizofrenia tra fenomenologia e linguistica: Il corpus CIPPS.* 2nd ed. Rome: Aracne.

Dovetto, Francesca M., Alessia Guida, Anna Chiara Pagliaro, and Raffaele Guarasci. 2024. "Silences and disfluencies in a corpus of patients with Alzheimer's Disease (CIPP-ma)." In *Proceedings of DiSS 2021: The 10th Workshop of Disfluency in Spontaneous Speech,* edited by Ralph L. Rose and Robert Eklund, 125–126. Saint-Denis: Université Paris VIII Vincennes—Saint-Denis. https://doi.org/10.18463/DISS-2021-001.

Duboisdindien, Guillaume, Cyril Grandin, Dominique Boutet, and Anne Lacheret-Dujour. 2020. "A multimodal corpus to check on pragmatic competence for mild cognitive impaired aging people." *Corpus* 19. https://doi.org/10.4000/corpus.4295.

Fleming, Valarie B. 2014. "Early detection of cognitive-linguistic change associated with Mild Cognitive Impairment." *Communication Disorders Quarterly* 35 (3): 146–157. https://doi.org/10.1177/1525740113520322.

Folia, Vasiliki, Ioannis Liampas, Eva Ntanasi, Mary Yannakoulia, Paraskevi Sakka, Georgios Hadjigeorgiou, Nikolaos Scarmeas, Efthimios Dardiotis, and Mary H. Kosmidis. 2022. "Longitudinal trajectories and normative language standards in older adults with normal cognitive status." *Neuropsychology* 36 (7): 626–639. https://doi.org/10.1037/neu0000843.

Folstein, M. F., S. E. Folstein, and P. R. McHugh. 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.

Fraser, Kathleen, Jed A. Meltzer, and Frank Rudzicz. 2016. "Linguistic features identify Alzheimer's Disease in narrative speech." *Journal of Alzheimer's Disorders* 49 (2): 407–422. https://doi.org/10.3233/JAD-150520.

Hilviu, Dize, Ilaria Gabbatore, Alberto Parola, and Francesca M. Bosco. 2022. "A cross sectional study to assess pragmatic strengths and weaknesses in healthy ageing." *BMC Geriatrics* 22:699. https://doi.org/10.1186/s12877-022-03304-z.

Jicha, Gregory A., Joseph E. Parisi, and Dennis W. Dickson. 2006. "Neuropathologic outcome of Mild Cognitive Impairment following progression to clinical dementia." *Archives of Neurology* 63 (5): 674–681. https://doi.org/10.1001/archneur.63.5.674.

Lemos, Raquel, Ana Afonso, Cristina Martins, James H. Waters, Filipe Sobral Blanco, Mário R. Simões, and Isabel Santana. 2016. "Selective reminding and free and cued selective reminding in Mild Cognitive Impairment and Alzheimer Disease." *Applied Neuropsychology: Adult* 23 (2): 85–89. https://doi.org/10.1080/23279095.2015.1012761.

López-de-Ipiña, K., U. Martinez-de-Lizarduy, and P. M. Calvo. 2020. "On the analysis of speech and disfluencies for automatic detection of Mild Cognitive Impairment." *Neural Computing & Applications* 32:15761–15769. https://doi.org/10.1007/s00521-018-3494-1.

Petersen, R. C. 2004. "Mild Cognitive Impairment as a diagnostic entity." *Journal of Internal Medicine* 256 (3): 183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x.

Pistono, Aurélie, Mélanie Jucla, Emmanuel J. Barbeau, Laure Saint-Aubert, Béatrice Lemesle, Benjamin Clevet, Barbara Köpke, Michèle Puel, and Jérémie Pariente. 2016. "Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's Disease." *Journal of Alzheimer's Disease* 50 (3): 687–698. https://doi.org/10.3233/JAD-150408.

Pistono, Aurélie, Jérémie Pariente, Catherine Bézy, Josette Pastor, Thi Mai Tran, Antoine Renard, Marion Fossard, Jean-Luc Nespoulous, and Jucla. 2017. "Inter-individual variability in discourse informativeness in elderly populations." *Clinical Linguistics and Phonetics* 31 (5): 391–408. https://doi.org/10.1080/02699206.2016.1277390.

Raven, John C. 1938. *Progressive matrices: A non-verbal test of a person's present capacity for intellectual activity.* London: H. K. Lewis.

Savy, Renata. 2006. "Specifiche per la trascrizione ortografica annotata dei testi raccolti." In *Italiano parlato: Analisi di un dialogo,* edited by F. Albano Leoni and R. Giordano, 1–37. Naples: Liguori.

Wright, Heather Harris. 2016. *Cognition, language and aging.* Amsterdam: Benjamins.