Defining Cohesion Features in the Study of Discourse Properties in Cognitive Impairment

Giorgia Albertin
University of Bologna
giorgia.albertin3@unibo.it

Dimitrios Kokkinakis
University of Gothenburg
dimitrios.kokkinakis@svenska.gu.se

Abstract

The analysis of discourse and pragmatics, which deteriorate alongside other linguistic levels in cognitive decline, can enhance our understanding of dementia-related language patterns and contribute to the improvement of automatic screening tools. This study focuses on discourse cohesion, specifically investigating three linguistic phenomena: reference, lexical repetition, and connectives. Six features related to these categories were defined and automatically extracted from an Italian corpus of semi-spontaneous speech, collected from patients with early dementia, MCI subjects, and healthy controls. Some of these features proved significant in distinguishing among the three groups. Additional quantitative analysis revealed notable differences in the use of these elements, suggesting a potential link between their degradation and cognitive decline.

1 Introduction

Dementia, or Major Neurocognitive Disorder (American Psychiatric Association et al., 2013), is characterized by language deterioration, which occurs within a wider frame of cognitive impairment, affecting memory, visuo-spatial skills, executive functions, and reasoning. In Alzheimer's disease, marked by episodic memory decline, linguistic deficits such as word-finding difficulties, reduced speech rate, and simplified syntax are well-documented (Catricalà et al., 2015; Orimaye et al., 2014). Discourse and pragmatics are also affected: speech is marked by an abundance of irrelevant details and difficulties in referencing key concepts, leading to reduced informativeness (Ahmed et al., 2013; Bschor et al., 2001). Moreover, errors related to the textual referential dimension, such as referent omission, has been observed and pronouns are used ambiguously (Drummond et al., 2015; Carlomagno et al., 2005). Disturbances in linguistic competence emerge from the onset, often preceded by an intermediate phase known as Mild Cognitive Impairment (MCI), in which cognitive decline is already present but the subject's independence in daily activities is preserved (Petersen, 2016). Therefore, language promises to be a viable approach for detecting subtle changes in cognitive status, even in pre-clinical stages, that could enhance screening and timely intervention (Vigo et al., 2022).

Supported by remarkable advancements in Natural Language Processing (NLP) and Machine Learning (ML), speech analysis has gained increasing importance in providing low-cost and portable tools for the prodromal detection of cognitive impairment (Petti et al., 2020). Many studies that pursued automatic language processing has focused on implementing features from the acoustic, lexical, and morpho-syntactic levels (Lindsay et al., 2021; Calzà et al., 2021), which are actually more straightforward to formalize. To step forward, incorporating discourse phenomena in the computational analysis would not only enrich the features used for classification but also enhance our understanding of how cognitive decline affects verbal competence. When speaking of discourse analysis, coherence immediately comes to mind: this property governs hearer's interpretation, ensuring continuity between utterances, which are organized as contextualized units to give rise to an intelligible text(Van Dijk, 1985). However, defining quantifiable indices related to coherence is not a simple task. For this reason, it was decided to begin by studying an aspect related to coherence, namely cohesion, the property of the superficial form of the text to display its internal unity through a network of *cohesive devices*, which are words or morphemes, that contribute to maintain relationships within the text (Ferrari, 2014).

In this work a method to design and formalize a set of cohesion features is proposed, with the aim of observing whether they significantly contribute to discriminate early dementia patients, MCI sub-

	Cohesive devices	Examples
Reference	personal pronouns	io (*I*), tu (*you*), essa/lei (*she*), egli/lui (*he*)
	possessive pronouns/adjectives	mio (*mine*), tuo (*your*), suo (*her/his*)
	demonstrative pronouns	questo (*this*), quello (*that*)
	indefinite pronouns	tutti (*all, everyoe*), alcuni (*some*)
	deictics	qua (*here*), là (*there*), sopra (*above*)
Connectives	single word	e (*and*), quindi (*therefore*), tuttavia (*however*)
	complex expressions	a causa di (*because of, due to*), da allora (*since then*)
	correlatives	da un lato... dall'altro (*on the one hand ... on the other hand*)

Table 1: Examples of cohesive devices of reference and connectives.

jects and healthy peers in a corpus of Italian elderly speakers. We focused on three of the major classes of cohesive devices, according to Halliday and Hasan (2014), i.e. reference, lexical iteration and connectives. The study is exploratory in nature, as it consist in an attempt to encompass discourse properties in automatic language analysis of cognitive impaired population for Italian. Therefore, although the significance of at least some of the designed features is expected, further analyses will be needed in the future to observe whether their interaction with other linguistic levels improve groups classification through ML algorithms.

2 Designing Features of Discourse Cohesion

Reference. Reference is involved when an expression occurs in the discourse that requires referring to another element for its interpretation (Halliday and Hasan, 2014). This mechanism operates either through the repetition of the antecedent, or through its substitution with other forms (Ferrari, 2014), such as the use of anaphora. The features related to this group focused on the second modality. Starting from a review of the relevant literature, an exhaustive list of referential elements in Italian was selected (see Prandi and De Santis (2006); Andorno (2003); Ferrari and Zampese (2000)). In the group were included personal, demonstrative, indefinite, and possessive pronouns, possessive adjectives and deictics. Table 1 are provided some examples of the particles considered.

The occurrences of these groups were counted and divided by the total number of words (COE_REF). We also computed *pronoun density* (COE_PRON_DENS), which is defined as the ratio between pronouns and nouns (Louwerse et al., 2004).

Lexical iteration. According to Halliday and Hasan (2014), the iteration of a lexical item is a specific use of the referential mechanism, which acquires cohesive force on its own because it is typically used when the referent is farther in the text. This set of features focuses on the repetition of three main open-class categories, namely nouns, (main) verbs, and adjectives. The use of words from these classes affects vocabulary´s richness, reflecting the speaker's tendency toward lexical variation. Lexical iteration features include the repetitions of lemmas divided by the total number of words (COE_RIP_LEM) and the average number of repetitions for repeated lemmas (COE_MEDRIP_LEM).

Connectives. As defined by Ferrari (Ferrari, 2010), connectives are morphologically invariable forms that explicit logical relations within parts of the text. Elements from different grammatical classes can be used as connectives and classified based on their function in the linguistics context, which usually reflects their meaning (e.g., temporal, causal, additive). To create a comprehensive list of connectives, we draw from the Lexicon of Italian Connectives - *LICO*[1] (Feltracco et al., 2018, 2016). LICO contains 173 entries, including single words, complex expressions, and correlatives, along with lexical or orthographic variants, POS category, the semantic relations conveyed according to the PDTB 3.0 schema (Webber et al., 2016), examples of usage, and correspondences of the forms with connectives in other languages. Some examples of those elements are reported in Table 1. A feature was devoted to computing the occurrences of connectives by the total number of words (COE_TC).

Additionally, a comprehensive feature that mea-

[1] http://connective-lex.info/

	Controls	MCI subjects	Early dementia
Inclusion criteria	No neurological/sensory deficits or intellectual disabilities MMSE $\geq$ 24; MoCA $\geq$ 18	No problem in daily living activities MMSE $\geq$ 18	Need of support in daily living activities MMSE $\geq$ 18
age	61.60 ± 6.93	64.34 ± 7.33	66.38 ± 6.70
education	13.00 ± 3.92	11.28 ± 4.35	9.38 ± 4.01

Table 2: Inclusion criteria, i.e. MMSE and MoCA scores and clinician's impressions, age and years of education (mean and st.dev) of OPLON corpus participants, as reported in Calzà et al. (2021).

Features	CON-MCI	CON-DEM
COE_REF	**0.017**	**0.000**
COE_PRON_DENS	**0.0211**	**0.000**
COE_RIP_LEM	0.989	**0.050**
COE_AVGRIP_LEM	**0.031**	**0.022**
COE_TC	0.183	0.338
COE_TOT	0.733	0.070

Table 3: Results of Kolmogorov-Smirnov test for the binary classification of controls (CON) vs MCI subjects (MCI) and earlu dementia patients (DEM). The cohesion features are reported along with their p-value, significant ones are marked in bold.

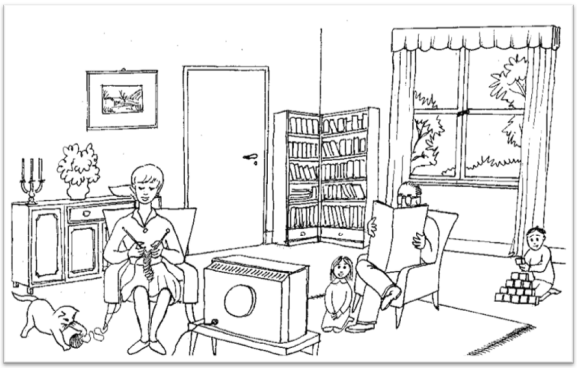

Figure 1: Image used for picture description task in *Esame del Linguaggio II* (Ciurli et al., 1996).

sures the overall impact of the classes of cohesion considered was computed by summing referential-substutive elements, lexical iteration items and connectives, divided by the total number of words (COE_TOT).

3 Corpus description

We extracted the cohesion features from the corpus collected for the OPLON (OPportunities for active and healthy LONgevity) project [2]. OPLON focused on the automatic extraction of linguistic features from Italian semi-spontaneous speech samples of cognitively impaired individuals and healthy controls, which were used to train machine learning classifiers to distinguish between the different groups (Beltrami et al., 2018). The dataset included 96 participants, comprising 48 healthy controls and 48 cognitively impaired individuals, of whom 32 were diagnosed with MCI and 16 with early-stage dementia. Inclusion criteria were based on the Mini-Mental State Examination (MMSE) (Magni et al., 1996) and the Montreal Cognitive Assessment (MoCA) (Conti et al., 2015) scores, along with an anamnesis conducted by a clinician. Further details on the demographic composition of the corpus are reported in Table 2 (Calzà et al., 2021).

[2] http://smartdata.cs.unibo.it/oplon_project

Speech was elicited using the picture description task from Esame del Linguaggio II [Language Examination II] (Ciurli et al. (1996); see Figure 1) and two semi-structured interview questions: "Could you please describe your typical working day?" and "Could you please describe the last dream you remember?". Audio recordings were transcribed and annotated, both manually and semi-automatically, with the aim of observing whether there were any significant differences in the subsequent analysis between the two approaches, which were not detected. A multidimensional parameter analysis was conducted to extract acoustic, rhythmic, lexical, readability, morpho-syntactic and syntactic features using the computational pipeline developed by Gagliardi and Tamburini (2022).

The cohesion features presented in this study were extracted from the `.conll` file generated through data annotation, using a python script.

4 Significance of cohesion features for discrimination

Individual statistical significance was assessed using the Kolmogorov-Smirnov non-parametric test to determine whether the cohesion features could contribute to the binary classification of the three groups. Table 3 reports significant p-

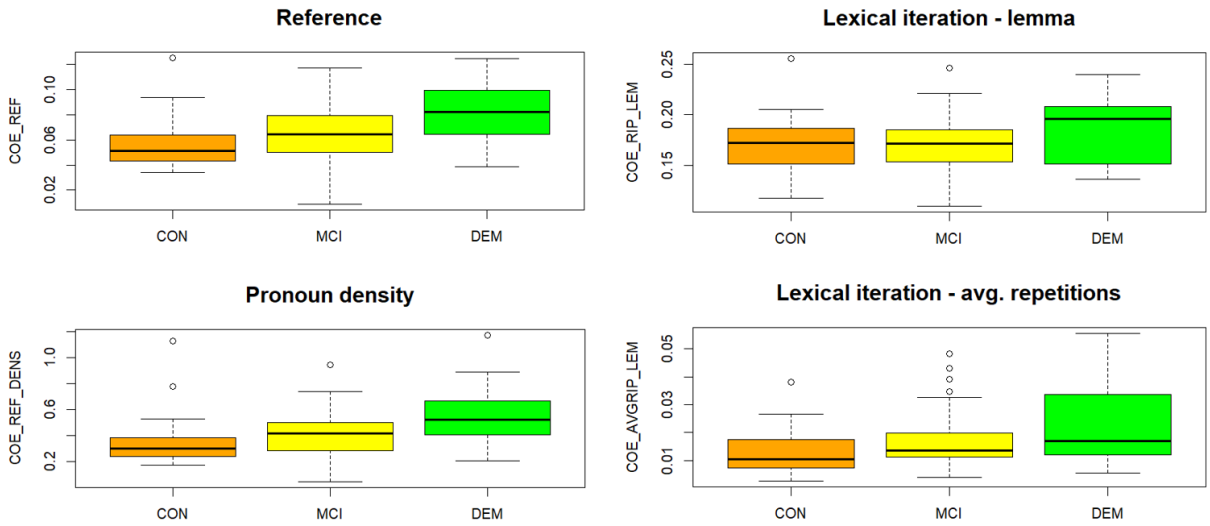

Figure 2: Distribution plots of significantly discriminative features: COE_REF indicates the n. of referential elements by the total n. of words and COE_PRON_DENS is the measure of pronoun density, COE_RIP_LEM and COE_AVGRIP are related to the repetitions of lemmas of nouns, adjectives and verbs,

values in discriminating the control group (CON) from the pathological cohort, i.e. MCI subjects (MCI) and early dementia patients (DEM). The results show that three of the designed features significantly differentiate CON from MCI. These features include the ones related to reference (COE_REF and COE_PRON_DENS) and the one concerning the average repetition of lemmas (COE_AVGRIP_LEM). All three features also significantly contribute to the discrimination between the CON and DEM groups, with the addition of COE_RIP_LEM, another feature of lexical iteration. The distribution of these features across the three groups can be visualized in Figure 2.

We observe that, compared to the control group, both early dementia and MCI subjects produced a greater number of referential expressions, in the total of uttered words, and exhibited higher pronoun density, with the DEM group showing higher indices than the other two. These results seem to suggest a change in linguistic competence with regard to reference processing, which may become increasingly evident as the disease worsens given the different level of significance in distinguishing CON vs MCI and CON vs DEM. While a qualitative analysis would be necessary to verify their non-canonical use, we can suggest that this preference might represent a compensatory strategy, favoring the use of less salient elements over full forms, which require their lexical access.

The distribution of lexical iteration features follows a similar trend to that observed for reference elements. Specifically, the DEM group shows a higher number of lemma repetitions, both in terms of the total number of words uttered and average repetitions, compared to the MCI group, which in turn exhibits higher values than the CON group. The trend in lexical iteration, which increases as cognitive status deteriorates, suggests that also the repetition of the same words may serve as a compensatory mechanism for difficulties in retrieving lexical forms due to word-finding problems, likely resulting in semantically impoverished speech.

5 Conclusion

In this study, we introduced a methodology for identifying linguistic features of cohesion to monitor changes in discourse properties in the speech of cognitive impaired subjects compared to healthy peers. Focusing on three cohesion categories - reference, lexical iteration, and connectives - we defined a set of features that were automatically extracted from an Italian corpus of semi-spontaneous speech, gathered from early dementia patients, MCI subjects and controls. The application of the Kolmogorov-Smirnov test revealed that features related to reference and lexical iteration, significantly contributed to the binary classification between CON-MCI and CON-DEM. Additionally, the quantitative distribution of these features shows interesting differences in the use of cohesive elements along the groups which seem to highlight a general decline in discourse properties with cognitive impairment.

References

Samrah Ahmed, Anne-Marie F Haigh, Celeste A de Jager, and Peter Garrard. 2013. Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. *Brain*, 136(12):3727–3737.

DSMTF American Psychiatric Association, DS American Psychiatric Association, et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Cecilia Andorno. 2003. *Linguistica testuale. Un'introduzione*. Carocci.

Daniela Beltrami, Gloria Gagliardi, Rema Rossini Favretti, Enrico Ghidoni, Fabio Tamburini, and Laura Calzà. 2018. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Frontiers in aging neuroscience*, 10:369.

Tom Bschor, Klaus-Peter Kühl, and Friedel M Reischies. 2001. Spontaneous speech of patients with dementia of the alzheimer type and mild cognitive impairment. *International psychogeriatrics*, 13(3):289–298.

Laura Calzà, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini. 2021. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language*, 65:101113.

Sergio Carlomagno, Anna Santoro, Antonella Menditti, Maria Pandolfi, and Andrea Marini. 2005. Referential communication in alzheimer's type dementia. *Cortex*, 41(4):520–534.

Eleonora Catricalà, Pasquale A Della Rosa, Valentina Plebani, Daniela Perani, Peter Garrard, and Stefano F Cappa. 2015. Semantic feature degradation and naming performance. evidence from neurodegenerative disorders. *Brain and language*, 147:58–65.

Paolo Ciurli, Paola Marangolo, and Anna Basso. 1996. *Esame del Linguaggio II. Manuale e materiale d'esame*. Giunti, Firenze.

Silvia Conti, Stefano Bonazzi, Marcella Laiacona, Marco Masina, and Mirco Vanelli Coralli. 2015. Montreal cognitive assessment (moca)-italian version: regression based norms and equivalent scores. *Neurological Sciences*, 36:209–214.

Cláudia Drummond, Gabriel Coutinho, Rochele Paz Fonseca, Naima Assunção, Alina Teldeschi, Ricardo de Oliveira-Souza, Jorge Moll, Fernanda Tovar-Moll, and Paulo Mattos. 2015. Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. *Frontiers in aging neuroscience*, 7:96.

Anna Feltracco, Elisabetta Ježek, and Bernardo Magnini. 2018. Enriching a lexicon of discourse connectives with corpus-based data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Anna Feltracco, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. 2016. Lico: A lexicon of italian connectives. *CLiC it*, page 141.

Angela Ferrari. 2010. Connettivi. *Enciclopedia dell'italiano*.

Angela Ferrari. 2014. Linguistica del testo. *Principi, fenomeni, strutture, Roma, Carocci*.

Angela Ferrari and Luciano Zampese. 2000. *Dalla frase al testo: una grammatica per l'italiano*. Zanichelli.

Gloria Gagliardi and Fabio Tamburini. 2022. The automatic extraction of linguistic biomarkers as a viable solution for the early diagnosis of mental disorders. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5234–5242, Marseille, France. European Language Resources Association.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in english*. Routledge.

Hali Lindsay, Johannes Tröger, and Alexandra König. 2021. Language impairment in alzheimer's disease—robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. *Frontiers in aging neuroscience*, 13:642033.

Max M Louwerse, Philip M McCarthy, Danielle S McNamara, and Arthur C Graesser. 2004. Variation in language and cohesion across written and spoken registers. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Eugenio Magni, Giuliano Binetti, Angelo Bianchetti, Renzo Rozzini, and Marco Trabucchi. 1996. Mini-mental state examination: a normative study in italian elderly population. *European Journal of Neurology*, 3.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality*, pages 78–87.

Ronald C Petersen. 2016. Mild cognitive impairment. *CONTINUUM: lifelong Learning in Neurology*, 22(2):404–418.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797.

Michele Prandi and Cristiana De Santis. 2006. Le regole e le scelte. *Introduzione alla grammatica italiana, UTET, Torino*.

Teun A Van Dijk. 1985. Semantic discourse analysis. *Handbook of discourse analysis*, 2:103–136.

Ines Vigo, Luis Coelho, and Sara Reis. 2022. Speech- and language-based classification of alzheimer's disease: a systematic review. *Bioengineering*, 9(1):27.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined vps. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31.