Letras de Hoje
Estudos e debates em linguística, literatura e língua portuguesa
Programa de Pós-Graduação em Letras da PUCRS
e-ISSN: 1984-7726 | ISSN-L: 0101-3335

Let. Hoje, v. 53, n. 1, p. 48-58, jan.-mar. 2018
http://dx.doi.org/10.15448/1984-7726.2018.1.30955

Detecting mild cognitive impairment in narratives in Brazilian Portuguese: first steps towards a fully automated system

Detecção de comprometimento cognitivo leve em narrativas em Português Brasileiro: primeiros passos para um sistema automatizado

Marcos Vinícius Treviso
Leandro Borges dos Santos
Christopher Shulby
Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences, University of São Paulo
Lilian Cristine Hübner
Department of Linguistics, Pontifical Catholic University of Rio Grande do Sul
National Council for Scientific and Technological Development – CNPq
Letícia Lessa Mansur
Department of Physiotherapy, Speech Pathology and Occupational Therapy. School of Medicine, University of São Paulo
Sandra Maria Aluísio
Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences, University of São Paulo

Abstract: In recent years, Mild Cognitive Impairment (MCI) has received a great deal of attention, as it may represent a pre-clinical state of Alzheimer´s disease (AD). In the distinction between healthy elderly (CTL) and MCI patients, automated discourse analysis tools have been applied to narrative transcripts in English and in Brazilian Portuguese. However, the absence of sentence boundary segmentation in transcripts prevents the direct application of methods that rely on these marks for the correct use of tools, such as taggers and parsers. To our knowledge, there are only a few studies evaluating automatic sentence segmentation in transcripts of neuropsychological tests. The purpose of this study is to investigate the impact of the automatic sentence segmentation method DeepBond on nine syntactic complexity metrics extracted of transcripts of CTL and MCI patients.

Keywords: Clinical diagnosis; Mild cognitive impairment; Automatic sentence segmentation; Syntactic complexity metrics; Automated discourse analysis tools

Resumo: Nos últimos anos, o Comprometimento Cognitivo Leve (CCL) tem recebido bastante atenção, uma vez que pode representar um estado pré-clínico da Doença de Alzheimer (DA). Na distinção entre idosos saudáveis (CTL) e pacientes com CCL, ferramentas de análise automática do discurso têm sido aplicadas a transcrições de narrativas em inglês e em português brasileiro. No entanto, a ausência da segmentação dos limites da sentença em transcrições impede a aplicação direta de métodos que empregam essas pontuações para o uso correto de ferramentas, como taggers e parsers. Segundo nosso conhecimento, há poucos estudos avaliando a segmentação automática de sentenças em transcrições de testes neuropsicológicos. O propósito deste estudo é investigar o impacto do método DeepBond para segmentação automática de sentenças em nove métricas de complexidade sintática extraídas de transcrições de CTL e de pacientes com CCL.

Palavras-chave: Diagnóstico clínico; Comprometimento cognitivo leve; Segmentação automática de sentença; Métricas de complexidade sintática; Ferramentas de análise do discurso

1 Introduction

The ageing of the population is a well-known social trend in developed countries that has become increasingly pronounced in developing countries. In Brazil, for example, the population pyramid is changing in shape, according to the IBGE (Brazilian Institute of Geography and Statistics) census from 2000 and 2010. Increased life expectancy with a high quality of life is priceless for every citizen; however, it raises serious financial and social issues, particularly in health, because aging may be accompanied by neurodegenerative diseases, requiring new resources and medical facilities. A recent study by Engedal and Laks (2016) reveals figures about the prevalence of dementia disorders worldwide and puts the total at 44 million individuals and possibly rising to 140 million by 2050. With regards to Brazil, they also draw some estimates about people with dementia as being 1.6 million, 1.2 million of which are not diagnosed at all, based on Herrera et al. (2002) and Scazufca et al. (2008) as well as on a recent study by Nakamura et al. (2015). One particular problem in Brazil is the low average level of education, since lower education is a known risk factor for dementia and imposes problems for its diagnosis and treatment (cf. CÉSAR et al., 2016).

Recently, the study of the syndrome known as Mild Cognitive Impairment (MCI), which is defined as a cognitive decline greater than expected in individuals at the same age and level of education, has become more and more important (cf. LEHR et al., 2012; TÓTH et al., 2015; VINCZE et al., 2016; and SANTOS et al., 2017). The interference caused by MCI in day-to-day activities is minimal, since it can only be perceived in complex situations and cannot be considered a type of dementia. However, the most frequent type, amnestic MCI, has the highest conversion rate to Alzheimer's disease (AD) (15% per year, versus 1-2% of the total population) (CLEMENTE; RIBEIRO-FILHO, 2008).

There are several instruments to identify pre-clinical and manifested dementias, such as the use of biomarkers, Magnetic Resonance Imaging and molecular neuroimaging (MCKHANN et al., 2011; MAPSTONE et al., 2014), but none of these are inexpensive solutions for public hospitals. Language is one of the most efficient information sources for assessing cognitive functions. Changes in language are frequently observed in patients with dementia and normally being the first to be observed by themselves and their family members. Therefore, the automatic analysis of discourse production is seen as a promising solution for diagnosing MCI, because its early detection ensures a greater chance of success in addressing potentially reversible factors or maintenance of functionality (MUANGPAISAN et al., 2012).

Neuropsychological tests that require some degree of memorization are usually included in verbal memory tests. This is the case of the logical memory test, in which an individual reproduces a story after listening to it. The higher the number of recalled elements from the narrative, the higher the memory score (WECHSLER, 1997; BAYLES; TOMOEDA, 1993; MORRIS et al., 2006). The evaluation of language from another standpoint presents, in discourse production (mainly in narratives), an attractive alternative because it allows for the analysis of linguistic microstructures (ANDREETTA et al., 2012), including phonetic-phonological, morphosyntactic and semantic-lexical components, as well as semantic-pragmatic macrostructures. Since it is a natural form of communication, it favors the observation of the patient's functionality in everyday life. Moreover, it provides data for observing the language-cognitive skills interface, such as executive functions (planning, organizing, updating and monitoring data). However, the main difficulties are: (i) the time required, since it is a manual task; and (ii) the subjectivity of the clinician in checking the presence of the main ideas of the narrative retold by the patient.

In terms of distinction between healthy aging adults, refereed in our study as CTL, and MCI patients, several studies have shown that discourse production is a sensitive task to differentiate individuals with MCI from controls using the Wechsler Logical Memory (WLM) test (PRUD'HOMMEAUX et al., 2011, PRUD'HOMMEAUX; ROARK, 2015). The original narrative used in this test is short, allowing the use of the output of Automatic Speech Recognition (ASR) methods of patients' speeches even without capitalization and sentence segmentation, as shown Lehr et al. (2012) for data in English. They based their method on automatic alignments of the original and patient transcripts in order to calculate the number of recalled elements from the narrative. Moreover, automated discourse analysis tools based on Natural Language Processing (NLP) resources and tools aiming at the diagnosis of language-impairing dementias via machine learning methods are already available for the English language (FRASER et al., 2015a; YANCHEVA et al., 2015). Yet a comprehensive NLP environment publicly available, designed for Brazilian Portuguese (BP), called Coh-Metrix-Dementia (ALUÍSIO et al., 2016a), was only recently developed.

Coh-Metrix-Dementia is based on a previous tool for discourse analysis, named Coh-Metrix-Port (SCARTON; ALUÍSIO, 2010), which was already used in a clinical discourse analysis study to classify written descriptions of healthy adults (TOLEDO et al., 2014). Based on previous studies using metrics and machine learning classifiers for the English language in clinical settings (e.g. CHAND et al., 2012; ROARK et al., 2011), 25 new metrics were

added to the existing 48 metrics for measuring syntactic complexity, semantic content of language via idea density (CUNHA et al., 2015), and text cohesion through latent semantics.

Although Coh-Metrix-Dementia is publicly available, there are major issues for its wide use in clinical settings: (i) the current need for manual narrative transcription and (ii) the absence of capitalization and boundary segmentation of the transcript, preventing the direct application of NLP methods that rely on these marks for the correct use of tools, such as taggers and parsers. In this paper we will focus on nine syntactic metrics of Coh-Metrix-Dementia for which the performance of segmentation has high impact.

The task of predicting sentence boundaries has been treated by many researchers. Liu et al. (2006) investigated the imbalanced data problem, since there are more non boundary words than ones with boundaries; their study was carried out using two speech corpora: conversational telephone and broadcast news, both for the English language. More recent papers have focused on Conditional Random Field (CRF) models. Wang et al. (2012) and Hasan et al. (2014) use CRF based strategies to identify word boundaries in speech corpora datasets, more specifically on English broadcast news data and English conversational speech (lecture recordings), respectively.

Although there are several methods of sentence segmentation for BP datasets (SILLA; KAESTNER, 2004; BATISTA, 2013; LÓPEZ; PARDO, 2015), none of which are adopted in transcriptions used in clinical settings for elderly people with dementias and related syndromes. The study most similar to our scenario is Fraser et al.'s (2015b), which proposes a segmentation method for aphasic speech based on lexical, Part of Speech (PoS) and prosodic features using tools and a generic acoustic model trained on resources for English. Their approach is based on a CRF model, which classifies a word by taking its context into account. With this model better results were obtained for broadcast news data, where speech is prepared, but the results on patient data were generally similar to the controls' data, allowing the use of several syntactic complexity metrics.

In this paper we present the first steps taken towards the wide use of Coh-Metrix-Dementia, working together with the DeepBond method for sentence segmentation. DeepBond (TREVISO et al., 2017a) uses recurrent convolutional neural networks with prosodic, PoS features, and also word embeddings and it was evaluated intrinsically on impaired, spontaneous speech and on normal, prepared speech. Moreover, when comparing the administration of the CRF method presented in Fraser et al. (2015b) and DeepBond method on our data, DeepBond method presents better results. In Section 2 we present Coh-Metrix-Dementia tool to automatically analyse text productions using several metrics. Section 3 presents DeepBond and details of the task of sentence segmentation, formally called Sentence Boundary Detection. Section 4 presents the extrinsic evaluation of DeepBond, using syntactic complexity metrics of Coh-Metrix-Dementia in order to measure the impact of using DeepBond to automatically segment narratives of neuropsychological tests.

2 Coh-Metrix-Dementia

Coh-Metrix-Dementia has been used to extract 73 features of oral narrative productions based on a sequence of pictures from the Cinderella story. Narratives of CTL, AD, and MCI patients were used in experiments with machine learning classification and regression methods (ALUÍSIO et al., 2016b). In their study, it was possible to separate CTL, AD, and MCI with an F1 score of 0.817, and separate CTL and MCI with an F1 score of 0.900. As for machine learning regression, the best results for Mean Absolute Error (MAE) were 0.238 and 0.120 for scenarios with three (CTL, AD and MCI) and two classes (CTL and MCI), respectively. The most discriminative features for the classifier and regressor were: dependence distance, Yngve and Frasier syntactic complexity metrics, the informativeness metric idea density (CUNHA et al., 2015) and disfluencies metrics, such as average duration of pauses, average number of short pauses, average number of vowel prolongations.

The architecture of the Coh-Metrix-Dementia environment is depicted in **Figure 1**. It receives, as input, two versions of the narratives to be analyzed: (i) the original transcription, with several kinds of annotations and (ii) a clean transcription of a patient's speech sample separated in sentences and capitalized.

In the original transcript, segments with hesitations or repetitions of more than one word or segments of a single word are annotated. The labels used for this kind of

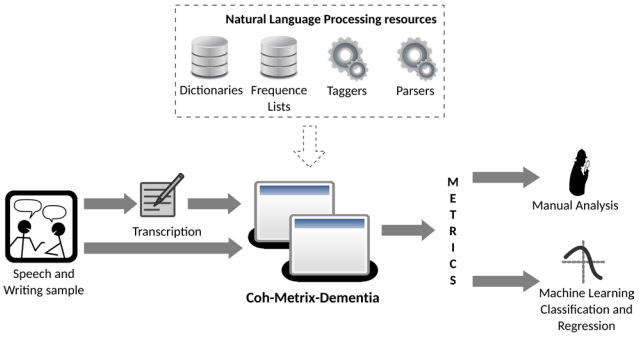

Figure 1. Architecture of Coh-Metrix-Dementia.

disfluency are <disf> and </disf>. Repetitions of unique words are captured automatically by the tool and do not require manual annotation. Empty emissions, which are comments that are not related to the topic of narration or confirmations, such as "*né*" (alright), are also annotated. Empty emissions are delimited by <empty> and </empty>. Prolongations of vowels (indicated by :::), short pauses (indicated by ...) and long pauses (indicated in seconds ((*pausa* XX *segundos*))) are also annotated. The six metrics related to these annotations are calculated as averages over the length of the narrative.

In the study by Aluísio et al. (2016b), narrative samples were recorded and transcribed by a trained researcher and sentence boundaries were marked later by a single researcher according to semantic and syntactic cues and the annotation of short and long pauses included in the original transcription; there was no distinction among the several types of disfluencies besides vowel prolongations.

A refined categorization of the types of disfluencies is welcome in order to be used as features to better distinguish the groups of interest, such as MCI and CTL and also to automate their removal to guarantee a successful parsing. In this paper, we have annotated several types of disfluencies over the same dataset in a double blind annotation experiment to compare the inter-annotator agreement between annotators following a manual based on Saffran et al. (1989). The removal of disfluencies is used as a first step for the sentence segmentation phase. Therefore, here we follow a manual to annotate sentence boundaries, different from the annotation used in Aluísio et al. (2016b). More details of the proposed manual annotation of narratives are presented in Section 4.3.

ela morava com a madrasta as irmã <empty>né</empty> e ela era diferenciada das três era maltratada ela tinha que fazer limpeza na casa toda no castelo alias e as irmãs não faziam nada...((pausa 3 segundos)) até que um dia chegou um convite do rei ele ia fazer um baile… ((pausa 3 segundos)) e a madrasta então <empty>é::: </empty> colocou que todas as filhas elas iam menos a cinderela... ((pausa 6 segundos)) bom como ela não tinha o vestido sapato as coisas tudo então ela mesmo teve que fazer a roupa dela começou a fazer

1. Ela morava com a madrasta, as irmã e ela era diferenciada das três, era maltratada.
2. Ela tinha que fazer limpeza na casa toda, no castelo aliás, e as irmãs não faziam nada.
3. Até que um dia chegou um convite do rei.
4. Ele ia fazer um baile e a madrasta então colocou que todas as filhas elas iam menos a cinderela.
5. Bom, como ela não tinha o vestido, sapato, as coisas tudo, então ela mesmo teve que fazer a roupa dela.
6. Começou a fazer.

Figure 2. Excerpt from a narrative of a patient with MCI, showing the original transcript and the clean transcript, capitalized, and segmented in sentences.

Figure 2 shows the two versions of a narrative that are expected in the Coh-Metrix-Dementia environment. The clean transcript is enumerated here to be compared with a second annotation proposed in this study.

After analyzing the versions of a narrative, Coh-Metrix-Dementia outputs a set of 73 textual metrics, divided in 14 categories: Ambiguity, Anaphoras, Basic counts, Connectives, Constituents, Coreferences, Disfluencies, Frequencies, Hypernyms, Logic operators, Latent Semantic Analysis, Semantic density, Syntactical complexity, and Pronouns, Types & Tokens. More detail can be found in the help section of the environment[1].

3 DeepBond: automatic sentence segmentation of narratives

The Sentence Boundary Detection (SBD) is the name of the task of segmenting narratives in neuropsychological tests which use audio transcriptions. SBD attempts to break a text into sequential units that correspond to sentences, and can be applied to either written text or audio transcriptions which do not necessarily end in final punctuation marks but are complete thoughts nonetheless. To perform SBD in speech texts is more complicated due to the lack of information such as punctuation and capitalization. Moreover text output is susceptible to recognition errors, in case of Automatic Speech Recognition (ASR) systems are used for automatic transcriptions (GOTOH; RENALS, 2000).

The work of Treviso et al. (2017a)2 proposed an automatic SBD method for impaired speech in Brazilian Portuguese, to allow a neuropsychological evaluation based on discourse analysis. The method uses RCNNs (Recurrent Convolutional Neural Networks) which independently treat prosodic and textual information, reaching state-of-the-art results for impaired speech. Also, this study showed that it is possible to achieve good results when comparing them with prepared speech, even when practically the same quantity of text is used.

In a follow-up study, Treviso et al. (2017b) showed that by using only a good word embedding model to represent textual information it is possible to achieve similar results with the state-of-the-art for impaired speech. Their study was set to verify which embedding induction method works best for the sentence boundary detection task, specifically whether it be those which were proposed to capture semantic, syntactic, or morphological similarities.

Here, we used the version of DeepBond presented in Treviso et al. (2017a). A boundary is defined as a period,

[1] <http://nilc.icmc.usp.br/coh-metrix-dementia/>.
[2] DeepBond can be downloaded at <http://github.com/mtreviso/deepbond>

exclamation mark, question mark, colon or semicolon, i.e., our problem is one of binary classification. DeepBond consists of a linear combination of two models. The first model is responsible for treating only lexical information, while the second treats only prosodic information. In order to obtain the most probable class, a linear combination was created between the two models, where one receives the pondered complement of the other.

4 Extrinsic evaluation using syntactic complexity metrics

4.1 *Datasets*

We used three datasets in the study of this paper. For all of them we have removed information about the capitalization and left all disfluencies intact in order to simulate a high-quality ASR system. In **Table 1**, statistics relevant to each dataset used are presented. The demographic information about the datasets are presented in **Table 2**.

4.1.1 *Spontaneous speech: MCI and healthy controls narratives*

The first dataset of discourse tests is a set of spontaneous speech narratives, based on a book of sequenced pictures from the well-known Cinderella story. In the test, an individual receives the book, then verbally tells the story to the examiner. The narrative is manually transcribed by a trained annotator who scores the transcription by counting the number of recalled propositions. This dataset consists of 60 narrative texts of BP speakers, 20 controls, 20 AD patients, and 20 MCI patients, diagnosed at Medical School of the University of São Paulo (FMUSP) and also used in Aluísio et al. (2016b). Counting all patient groups, this dataset has an average of 30.72 sentences per narrative, and each sentence averages of 12.92 words.

The second dataset of neuropsychological tests is available from the *Bateria de Avaliação da Linguagem no Envelhecimento* (BALE) ("Battery of Language Assessment in Aging", in English), under a process of validation (JERONIMO, 2015; HÜBNER et al. [in preparation]). Including tasks assessing naming, episodic verbal memory, semantic judgement, semantic categorization at the word level, metaphor comprehension and completion at the sentence level, as well as narrative production based on a sequence of story scenes, free narrative production on a given topic (news and funny event) and narrative retelling from a story orally presented, the battery aims at tackling some of the language impairments normally associated with MCI and AD. Moreover, the tasks were developed so that their administration is adjusted to include illiterate and lower educational level participants' linguistic data, population samples which are very common in the Brazilian public health system.

Here, we used the transcription of 10 narratives taken from the narrative production test based on the presentation of a set of seven pictures telling a story of a boy who hides a dog that he found on the street (The dog story (LE BOEUF, 1976)). The participants are asked to carefully observe the pictures, displayed in the correct sequence, and as soon as they feel confident to start telling the story, their production is recorded. The test administrator tries not to interfere, as in the previous task. Assessment includes the amount and quality of recorded propositions from the text. Complementary assessment included comprehension questions approaching the micro and macrostructural aspects of the narrative, as well as semantic and syntactic quantitative and qualitative aspects. Because this dataset is also composed of patient narratives, we can evaluate how well our model behaves on data from the same domain, where the story and vocabulary of the narratives are different from the ones in which the model has been tested. The average number of sentences and the average size of the sentences in this dataset are 16.60 and 6.58, respectively. When compared with the first dataset, this one is composed of less sentences and the sentences have fewer words on average.

4.1.2 *Prepared speech: Brazilian Constitution*

The third dataset was made available by FalaBrasil, a project at the Federal University of Pará's Signal Processing Laboratory (BATISTA, 2013). This dataset is composed of articles of Brazil's 1988 constitution, in which the speech is read. Each file has an average of 30 seconds of transcribed speech. To use these files in our scenario a preprocessing step was necessary, which removed lexical tips that indicate the beginning of articles, sections and paragraphs. This removal was carried out on both the transcripts and audio. In addition, we separated the new dataset organized by articles, yielding 357 texts in total. Then we marked the end of each article, paragraph, and inserted punctuation at the end. Titles and chapters have been ignored in this process. We randomly selected 60 texts from this dataset, only following the condition that the number of sentences of each text sentence should be higher than 12. We refer to the largest dataset as Constitution L, and the dataset with the 60 texts as Constitution S. The average number of sentences in each text of Constitution L is 7.56, and the average size of these sentences is 23.45 words while Constitution S has 23.48 sentences on average, and these sentences have an average of 21.66 words.

Table 1. Narrative statistics for each dataset.

Dataset:	Cinderella			Constitution		BALE	
	CTL	MCI	AD	L	S	CTL	MCI
Texts	20	20	20	357	60	6	4
Sentences	621	552	670	2698	1409	98	68
Words	8096	7904	7807	63275	30521	611	482
Total Duration	1h 9m	1h 12m	1h 50m	7h 39m	3h 43m	4m	4m

Table 2. Demographic information of the participant groups.

Dataset:	Cinderella			BALE	
	CTL	MCI	AD	CTL	MCI
Avg. Age (SD)	74.8 (11.3)	73.3 (5.9)	78.2 (5.1)	73.3 (2.5)	67.2 (10.5)
Avg. Years of Education (SD)	11.4 (2.6)	10.8 (4.5)	8.6 (5.5)	6.8 (1.8)	4.0 (1.6)
Sex	16 F	14F	10F	2F	5F

4.2 *Segmentation of the speech corpora*

Word and sound boundaries must be identified in order to label useful audio excerpts. Fluent listeners hear speech as a sequence of discrete sounds even when there are no pauses in the waveform. This segmentation is not as trivial for a machine which receives a single signal. In our algorithm we segment the audio excerpts from the corpus in phone and word boundaries by forced alignment (YUAN; LIBERMAN, 2008). Forced alignment uses a trained acoustic model to predict phoneme sequences, then uses the orthographic transcriptions to force the recognized phonemes into their likely transcriptions based on the words present and attempts to join the transcriptions with the correct timestamps present in the audio signal. The forced alignment method requires two indispensable components: (i) a robust acoustic model; and (ii) a well-designed pronunciation model since the more phoneme sequences are added, the more training examples are required, and the more closely related the phoneme sequences occurring in similar phonological contexts are, the more difficulties will be encountered by the model. In this work a dictionary was built using automatic transcriptions from Petrus (MARQUIAFÁVEL, 2015) for all words in the scripts used for training and testing. Since this was a pilot experiment, adaptations were not made for multiple pronunciations, which will be included in the future.

4.3 *Segmentation and annotation of the narratives*

The segmentation of the narratives which were the basis for the assessment of Coh-Metrix-Dementia reported in Aluísio et al. (2016b) was performed by a single person, without the support of a spontaneous speech annotation method on clinical data nor an annotation manual. Therefore, this work re-annotated the disfluencies and segmentation in sentences based on the work of Saffran et al. (1989) using a 3 step process for annotation of propositions of an ungrammatical narrative: (1) Removal (by annotation) of text excerpts, here termed "non-word" narratives; (2) Segmentation of sentences and judgment if (+) or (-), that is, (+) is for sentences prosodically, syntactically and semantically well formed in the argumental structure of the sentence; and (3) Annotation propositions in the well-formed sentences.

The annotation was performed by peers, using the brat annotation tool[3], and the agreement between peers was measured by the kappa statistic. The kappa statistics for steps (1) and (2), that are of interest for this work, were calculated for the task of selection and categorization of the selected excerpt. In step 1 the categorization involves the following types of 10 non-words: (1) Neologisms, i.e. word creation; (2) Patient comments not related to the topic of narration or confirmations. For example, <empty>né ("alright?")</empty>; (3) False starts: then, well then, so; (4) Coordinating conjunctions (and, but, or) which join two complete sentences; (5) Direct discourse markers, like: "He said, X"; "The prince said, Y"; (6) Repetions: [Cinderela wanted] Cinderela wanted; (7) Interruptions: [The dad hates Romeo] the dad hates Romeo's dad; (8) Corrections: the dad hates [Romeo's mom] Romeo's dad; (9) Elaborations: Dad hates [dad] Romeo's dad; and (10) Deictics (mainly spatial), e.g. the

[3] <http://brat.nlplab.org/>.

adverbs with locative value: aqui ("here"), ali ("there"), cá ("over here"), lá ("over there").

And in the step 2 if the sentence is well-formed (+) or not (-).

The kappa value for non-words selection was 0.81 and for categorization was 0.91; for sentence boundary identification it was 0.84, all of them very high, but for sentence categorization it was 0.14 as the judges diverged strongly on the concept of which arguments are needed for a specific verb. This involves the knowledge of semantic role labeling theory, which is not an easy concept and deserves a manual by itself for annotation. The categorization is not important for this work, as we have to segment all the sentences regardless.

After the evaluation of our annotation manual and inter annotator agreement, the segmentation in sentences was carried out before the non-words were removed. This was done in order to simulate an ASR system.

Of particular interest to this article are the narrative non-words (many of them are disfluencies) and sentence segmentation, using semantic, syntactic and prosodic cues. From the 10 non-words, 9 were removed from the original transcript for the extrinsic evaluation of this study. The conjunctions were maintained as they are well resolved by parsers, as long as they are separated by commas throughout the narrative.

Figure 3 shows the same excerpt from **Figure 2**, using the new annotation for segmentation, with annotated disfluencies (marked with *) to be removed.

1. ela morava com a madrasta as irmã né* . ("she lived with the stepmother the sister alright*")
2. e ela era diferenciada das três era maltratada . ("and she was differentiated from the three was mistreated")
3. ela tinha que fazer limpeza na* casa* toda* no castelo alias* . ("she had to do the cleaning in the* entire* house* actually* in the castle")
4. e as irmãs não faziam nada . ("and the sisters didn't do anything")
5. até que um dia chegou um convite do rei . ("until one day the king's invitation arrived")
6. ele ia fazer um baile . ("he would invite everyone to a ball ")
7. e a madrasta então é* colocou que todas as filhas elas* iam menos a cinderela . ("and then the stepmother is* said that all the daughters they* would go except for cinderella")
8. bom* como ela não tinha o* vestido* sapato* as coisas tudo então ela mesmo teve que fazer a roupa dela ("Well* since she didn't have a* dress* shoes* all the things she had to make her own clothes")
9. começou a fazer . ("she started to make them")

Figure 3. Segmentation annotation with emphasis on the removal of disfluencies, informed by the annotation manual.

Comparing the resulting annotations of **Fig. 2** and **Figure 3**, one can see that 6 sentences in **Figure 2** were transformed into 9 sentences. As for the disfluencies removed, **Figure 3** shows, for example, two reformulations in sentence 3 and in sentence 8 which were removed (marked with an "*"); sentence 8 contains a false start and sentence 1 a comment. This annotation process generates short sentences, allowing for higher success for analysis by the parsers.

4.4 *Metrics evaluated*

Some of the metrics used in automatic evaluation studies of speech in the clinical field are influenced by the method in which the transcription is segmented because it depends on the robustness of the parser being used and the characteristics of the annotation used in the datasets on which the parsers were trained. In our study of BP narratives, the syntactic metrics of Coh-Metrix-Dementia used the dependency parser MALT-parser (NIVRE et al., 2006), trained with the dataset for the task CoNLL-X 2006 Multi-lingual Dependency Parsing, and the constituency parser LX-parser (SILVA et al., 2010). The former was the parser with better performance and the choice of the latter was because it is the only freely available constituency parser for Portuguese.

In this study, we intended to assess whether or not automatic segmentation had an impact on the syntactic metrics, therefore disfluencies were removed from the narratives with manual and automatic annotation. In order not to have to evaluate two variables, we selected 9 syntactic metrics from Coh-Metrix-Dementia to see if there was any significant difference between the manual and automatic segmentation.

The metrics mean Yngve's complexity (YNGVE, 1960), Frazier's complexity (FRAZIER, 1985), mean clauses per sentence, noun phrase incidence, modifiers per noun phrase and pronouns per noun phrase depend on the constituent structure and the dependency distance metric is calculated by the dependency parser. The first two depend on the success of analyzing the tree as a whole; the third depends on the correct identification of verb phrases; and the last three depend on the correct identification of noun phrases. Words per sentence and number of sentences are correlated, since the greater the number of sentences, the smaller their size.

4.5 *Results and discussion*

The results are given in **Table 3.** Such comparisons were analyzed using the Wilcoxon for paired data, with a significance level of 5% (p-value <0.05), the null hypothesis is that the metrics have equal averages for manual and automatic segmentation. Only the metric modifiers per noun phrase for the MCI group presented a significant statistical difference.

Table 3. Values shown are mean (standard deviation) and *p*-value. Bold values denote statistical significance at the $p < 0.05$ level.

Metrics	CTL			MCI		
	Manual	Auto	*p*	Manual	Auto	*p*
Yngve Complexity	2.08 (0.11)	2.05 (0.16)	0.76	2.08 (0.15)	2.01 (0.11)	0.12
Frazier Complexity	6.96 (0.28)	7.01 (0.20)	0.57	6.89 (0.24)	6.77 (0.30)	0.11
Dependency Distance	33.13 (6.82)	35.48 (14.62)	0.60	34.83 (7.74)	32.59 (7.52)	0.19
Number of Sentences	28.65 (12.52)	28.60 (13.18)	0.90	26.30 (12.49)	28.90 (14.93)	0.08
Words per sentence	11.02 (1.62)	11.29 (3.43)	0.86	11.58 (2.13)	10.66 (1.69)	0.09
Mean Clauses per Sentence	1.86 (0.27)	1.88 (0.42)	0.80	1.88 (0.33)	1.77 (0.26)	0.19
Pronouns per Noun Phrase	0.24 (0.09)	0.25 (0.08)	0.47	0.24 (0.09)	0.23 (0.09)	0.15
Noun Phrase Incidence	324.06 (29.49)	323.71 (24.62)	0.84	305.12 (78.38)	325.92 (36.53)	0.19
Modifiers per Noun Phrase	0.40 (0.22)	0.39 (0.07)	0.19	0.41 (0.08)	0.39 (0.07)	0.05

Since manual segmentation was used only for lexical information, and the beginning of the sentences are well defined by the discourse markers "*então*" ("then") and "*ai*" ("there") and finally by the confirmation marker "*né*" ("alright"), our method was able to learn this information, but occasionally the label "*então*" ("then") is used as a conjunction and our method could end up adding a period before this marker, this fact can be seen in the first example below.

Depending on how the sentences are segmented, the parser can not generate a noun phrase due to an error in the model or the sentence does not really have a noun phrase; this fact is shown in the second pair of examples in which the second sentence ("*Mas não coube nelas*.") does not have a noun phrase. This discordance between manual and automatic segmentation ends up generating a difference in metrics, and since the modifiers per noun phrase metric makes an analytical differentiation within a small context it may be more susceptible to these small variations. **Figure 4** shows three pairs of examples with manual and automatic segmentation showing the metric value modifiers per noun phrase in parentheses.

We also analyzed whether any of the 9 metrics were able to distinguish the groups. Although we know that in a machine learning approach the features in tandem help to correctly classify groups, for a manual analysis of the clinical metrics results this distinction is important.

Table 4 shows the p value results between CTL and MCI for manual and automatic segmentation. Group differences were measured using Mann-Whitney non-parametric statistical tests for unpaired data, with a significance level of 5% (p value <0.05), the null hypothesis is that the mean CTL is the same as the mean MCI.

Manual: Como ela era muito bonita e a madrasta tinha ciúmes dela em relação as filhas então dava tudo de bom pras filhas e deixava a cinderela de escanteio. (0.33)

Automatic: E ela então ficou a madrasta como ela era muito bonita. E a madrasta tinha ciúmes dela em relação as filhas . Então dava tudo de bom pras filhas e deixava a cinderela de escanteio. (0.37)

Manual: As irmãs experimentaram o sapatinho mas não coube nelas. (1.0)

Automatic: As irmãs experimentaram o sapatinho. Mas não coube nelas. (0.5)

Manual: No palácio tinha essas três megeras acho que era avó a mãe e a filha. (0.33)

Automatic: No palácio tinha essas três megeras acho que era avó. A mãe e a filha. (0.25)

Figure 4. Pairs of examples with manual and automatic segmentation.

Table 4. A comparison of CTL and MCI. Bold values denote statistical significance at the $p < 0.05$ level.

Metric	CTL vs MCI	
	Manual	Auto
Yngve Complexity	0.90	0.42
Frazier Complexity	0.24	**0.05**
Dependency Distance	0.49	0.82
Number of Sentences	0.59	0.82
Words per sentence	0.49	0.67
Mean Clauses per Sentence	0.86	0.42
Pronouns per Noun Phrase	0.64	0.51
Noun Phrase Incidence	0.66	0.88
Modifiers per Noun Phrase	0.95	0.82

Table 4 indicates that the Frazier Complexity has a statistically significant difference in the automatically segmented samples. One can also see that this metric in manual segmentation has a low p-value but not enough to reject the null hypothesis. Even though through most metrics it is not possible to affirm the existence of statistical difference between the groups we know that some syntactic metrics help in the automatic classification process as presented by Roark et al. (2011) and Aluísio et al. (2016b). An adequate segmentation and the removal of disfluencies help the parser and do not generate incorrect trees.

5 Conclusions and future work

We showed that our model, using a recurrent convolutional neural network, is benefited by word embeddings and can achieve promising results even with a small amount of data. We found that our method is better for cases where speech is planned, since the prosodic features lend more weight for classification. The results of our evaluation indicate that only the metric modifiers per noun phrase for the MCI group presented a significant statistical difference on automatically and manually segmented transcripts. These results suggest that DeepBond is robust to analyze impaired speech and can be used in automated discourse analysis tools to differentiate narratives produced by MCI individuals and healthy controls and similar studies. As for future work, we intend to analyze the impact of segmentation in other tests.

An ideal system for automatic detection of cognitive impairments would need to be fully automated. Our future work includes a pipeline featuring a full ASR system for Brazilian Portuguese which would be robust enough to handle impaired speech. The recognition output would then be piped into a disfluency detection stage, followed by the sentence segmentation presented in this work, so that it could be properly treated by the NLP tools in Coh-Metrix-Dementia and finally a classification stage to output whether or not the subject can be classified as having MCI. This approach might permit the screening of MCI through a computerized test using tablets. Automatic Speech recognition for this purpose presents a series of challenges. Firstly, one can imagine that the eventual administrator of this tool will not be in a noise free environment, so the acoustic model must be robust enough to function well in a noisy hospital or clinic. Secondly, cognitively impaired speech differs in many ways from "normal" speech. This becomes a double-bladed sword for our task as this piece of information is very important for MCI detection but it is very cumbersome for the ASR because most systems are not robust enough to anticipate these differences. We plan to work on both of these fronts by building methods as well as corpora suitable for the ultimate task.

References

ALUÍSIO, S. et al. Computational Tool for Automated Language Production Analysis Aimed at Dementia Diagnosis. In: SILVA, J. et al. (Org.). *International Conference on Computational Processing of the Portuguese Language, Demonstration Session*, 2016a.

ALUÍSIO, S.; CUNHA, A.; SCARTON, C. Evaluating Progression of Alzheimer's Disease by Regression and Classification Methods in a Narrative Language Test in Portuguese. In: SILVA, J. et al. (Org.). *International Conference on Computational Processing of the Portuguese Language*, p. 374-384, Springer International Publishing, 2016b.

ANDREETTA, S.; CANTAGALLO, A.; MARINI, A. Narrative discourse in anomic aphasia. *Neuropsychologia*, v. 50, n. 8, p. 1787-1793, 2012.

BATISTA, Pedro dos Santos. Avanços em Reconhecimento de Fala para Português Brasileiro e Aplicações: Ditado no Libre Office e Unidade de Resposta Audível com Asterisk. Dissertação (Mestrado) – Universidade Federal do Pará. Belém, 2013.

BAYLES, K. A.; TOMOEDA, C. K.; TROSSET, M. W. Alzheimer's disease: Effects on language. *Developmental Neuropsychology*, v. 9, n. 2, p. 131-160, 1993.

CÉSAR, K. G. et al. Prevalence of Cognitive Impairment Without Dementia and Dementia in Tremembé, Brazil. *Alzheimer Dis Assoc Disord*, v. 30, n. 3 p. 264-71, 2016.

CHAND, V. et al. Rubric for Extracting Idea Density from Oral Language Samples. *Current Protocols in Neuroscience*, v. 10, n. 5, 2012.

CLEMENTE, R. S. G.; RIBEIRO-FILHO, S. T. Comprometimento cognitivo leve: aspectos conceituais, abordagem clínica e diagnóstica. *Revista Hospital Universitário Pedro Ernesto*, v. 7, n. 1, p. 68-77, 2008.

CUNHA, A. L. V. et al. Automatic Proposition Extraction from Dependency Trees: Helping Early Prediction of Alzheimer's Disease from Narratives. Paper presented at the *28th International Symposium on Computer-Based Medical Systems*, Institute of Electrical and Electronics Engineers, 2015. p. 127-130.

ENGEDAL, K.; LAKS, J. Towards a Brazilian dementia plan? Lessons to be learned from Europe. *Dementia & Neuropsychologia*, v. 10, n. 2, p. 74-78, 2016.

HÜBNER, L. C. et al. BALE: Bateria de Avaliação da Linguagem no Envelhecimento. In: FONSECA, R. P.; ZIMMERMANN, N. et al. *Tarefas de avaliação neuropsicológica para adultos: memórias e linguagem*. São Paulo: Memnon. (em preparação)

FRASER, K. C.; MELTZER, J. A.; RUDZICZ, F. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, v. 49, n. 2, p. 407-422, 2015a.

FRASER, K.C. et al. Sentence segmentation of aphasic speech. *2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies* (NAACL-HLT-2015), Denver, Colorado, 2015b.

FRAZIER, L. Syntactic complexity. In: DOWTY D. R.; KARTTUNEN, L.; ZWICKY, A. M. (Org.). *Natural language parsing: Psychological, computational, and theoretical perspectives*. Cambridge: Cambridge University Press, 1985. p. 129-189.

JERONIMO, G. M. *Produção de narrativas orais no envelhecimento sadio, no comprometimento cognitivo leve e na doença de Alzheimer e sua relação com construtos cognitivos e escolaridade*. 2015. 201 f. Tese (Doutorado em Linguística) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2015.

HASAN, M.; DODDIPATLA, R.; HAIN, T. Multi-Pass Sentence-End Detection of Lecture Speech. Paper presented at the *15 Conference of the International Speech Communication Association*, Singapore, 2014.

HERRERA JR, E. et al. Epidemiologic survey of dementia in a community-dwelling Brazilian population. *Alzheimer Disease & Associated Disorders*, v. 16, n. 2, p. 103-108, 2002.

LE BOEUF, C. *Raconte: 55 historiettes en images*. L'École, 1976.

LEHR, M. et al. Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment. Paper presented at the *13 Annual Conference of the International Speech Communication Association*, Portland, OR, 2012.

LIU, Y. et al. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech & Language*, v. 20, n. 4, p. 468-494, 2006.

LÓPEZ, R.; PARDO, T. A. Experiments on Sentence Boundary Detection in User-Generated Web Content. In: GELBUKH, A (Org.). International Conference on Intelligent Text Processing and Computational Linguistics, Springer International Publishing, p. 227-237, 2015.

MAPSTONE, M. et al. Plasma phospholipids identify antecedent memory impairment in older adults. *Nature medicine*, v. 20, n. 4, p. 415-418, 2014.

MARQUIAFÁVEL, V. S. *Ambiente de suporte à transcrição fonética automática de lemas em verbetes de dicionários do português do Brasil*. Dissertação (Mestrado) – Universidade Estadual Paulista Júlio de Mesquita Filho, São José do Rio Preto, 2015. (não publicada)

MCKHANN, G. M. et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, v. 7, n. 3, p. 263-269, 2011.

MORRIS, J. C. et al. The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Disease & Associated Disorders*, v. 20, n. 4, p. 210-216, 2006.

MUANGPAISAN, W.; PETCHARAT, C.; SRINONPRASERT, V. Prevalence of potentially reversible conditions in dementia and mild cognitive impairment in a geriatric clinic. *Geriatrics & gerontology international*, v. 12, n. 1, p. 59-64, 2012.

NAKAMURA, A. E. et al. Dementia underdiagnosis in Brazil. *The Lancet*, v. 385, n. 9966, p. 418-419, 2015.

NIVRE, J.; HALL, J.; NILSSON, J. Maltparser: A data-driven parser-generator for dependency parsing. Paper presented at the *6th International Conference on Language Resources and Evaluation*, Genoa, 2006.

PRUD'HOMMEAUX, E. T.; MITCHELL, M.; ROARK, B. Using Patterns of Narrative Recall for Improved Detection of Mild Cognitive Impairment. In: *Annual RESNA Conference/3rd International Conference on Technology and Aging (ICTA)*. RESNA (Rehabilitation Engineering and Assistive Technology Society of North America), 2011.

PRUD'HOMMEAUX, E.; ROARK, B. Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, v. 41, n. 4, p. 549-578, 2015.

ROARK, B. et al. Spoken language derived measures for detecting mild cognitive impairment. *Institute of Electrical and Electronics Engineers Transactions on Audio, Speech, and Language Processing*, v. 19, n. 7, p. 2081-2090, 2011.

SANTOS, L. B. et al. Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. In: *Annual Meeting of the Association for Computational Linguist*, Vancouver, 2017.

*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistic*s (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2017. p. 1284-1296.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática*, v. 2, n. 1, p. 45-61, 2010.

SCAZUFCA, M. et al. High prevalence of dementia among older adults from poor socioeconomic backgrounds in Sao Paulo, Brazil. *International Psychogeriatrics*, v. 20, n. 02, p. 394-405, 2008.

SILVA, J. et al. Out-of-the-box robust parsing of portuguese. In: PARDO, T. A. S. et al. (Org.). *International Conference on Computational Processing of the Portuguese Language*. Springer Berlin Heidelberg, 2010. p. 75-85.

SAFFRAN, E. M.; BERNDT, R. S.; SCHWARTZ, M. F. The quantitative analysis of agrammatic production: Procedure and data. *Brain and language*, v. 37, n. 3, p. 440-479, 1989.

SILLA JR, C. N.; KAESTNER, C. A. An analysis of sentence boundary detection systems for English and Portuguese documents. In: GELBUKH, A (Org.). *International Conference on Intelligent Text Processing and Computational Linguistics*, Seoul, Korea. Springer Berlin Heidelberg, 2004. p. 135-141.

TÓTH, L. et al. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: *Proceedings of the 16th Annual Conference of the International Speech Communication Association*. International Speech and Communication Association, 2015. p. 2694-2698.

TOLEDO, C. M. et al. Automatic classification of written descriptions by healthy adults: an overview of the application of natural language processing and machine learning techniques to clinical discourse analysis. *Dementia & Neuropsychologia*, v. 8, n. 3, p. 227-235, 2014.

VINCZE, V. et al. Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In: *Proceedings of the 54th Annual Meeting of the Association Computer Linguistics*. Association for Computational Linguistics, 2016.

WANG, X.; NG, H. T.; SIM, K. C. Dynamic Conditional Random Fields for Joint Sentence Boundary and Punctuation Prediction. In: *Interspeech*, 2012. p. 1384-1387.

WECHSLER, D. *Wechsler Adult Intelligence Scale-III*, New York: Psychological Corporation, 1997.

YANCHEVA, M.; FRASER, K.; RUDZICZ, F. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. Paper presented at the *6 Workshop on Speech and Language Processing for Assistive Technologies* (SLPAT2015), Dresden, 2015.

YNGVE, V. A Model and an Hypothesis for Language Structure. *Proceedings of the American Philosophical Society*, v. 104, n. 5, p. 444-466, 1960.

YUAN, J.; LIBERMAN, M. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, v. 123, n. 5, p. 3878, 2008.

Recebido: 28/11/17
Aprovado: 10/01/18

Contato:
Marcos Vinícius Treviso <marcosvtreviso@gmail.com>