IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Topic coherence analysis for the classification of Alzheimer's disease

Anna Pompili[1,2], *Alberto Abad*[1,2], *David Martins de Matos*[1,2], *Isabel Pavão Martins*[3,4]

[1] L^{2}F – Spoken Language Systems Laboratory – INESC-ID Lisboa
[2] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3] Laboratório de Estudos de Linguagem, Faculty of Medicine, University of Lisbon, Portugal
[4] Instituto de Medicina Molecular, Lisbon, Portugal

anna@l2f.inesc-id.pt

Abstract

Language impairment in Alzheimer's disease is characterized by a decline in the semantic and pragmatic levels of language processing that manifests since the early stages of the disease. While semantic deficits have been widely investigated using linguistic features, pragmatic deficits are still mostly unexplored. In this work, we present an approach to automatically classify Alzheimer's disease using a set of pragmatic features extracted from a discourse production task. Following the clinical practice, we consider an image representing a closed domain as a discourse's elicitation form. Then, we model the elicited speech as a graph that encodes a hierarchy of topics. To do so, the proposed method relies on the integration of various NLP techniques: syntactic parsing for sentence segmentation into clauses, coreference resolution for capturing dependencies among clauses, and word embeddings for identifying semantic relations among topics. According to the experimental results, pragmatic features are able to provide promising results distinguishing individuals with Alzheimer's disease, comparable to solutions based on other types of linguistic features.

Index Terms: natural language processing, topic coherence, Alzheimer's disease

1. Introduction

In 2006 the prevalence of Alzheimer's disease (AD) was 26.6 million people worldwide. Due to the increase of average lifespan, it is expected that this data will quadruple by 2050, affecting 1 in 85 persons worldwide [1]. No treatments stop or reverse the progression of the disease, though some may temporarily improve the symptoms. AD is currently diagnosed through an analysis of the patient history and through neuropsychological tests assessing cognitive decline in different domains (memory, reasoning, language, and visuospatial abilities). In fact, although the prominent symptom of the disease is memory impairment, language problems are also prevalent and existing literature confirms they are an important factor [2, 3, 4, 5, 6]. Impairments in language abilities are usually the result of a decline either in the semantic or pragmatic levels of language processing. Semantic processing is related with the content of language and involve words and their meaning. Deficits in this domain are typically characterized by difficulties in word finding, word comprehension, semantic paraphasia, and by the use of a reduced vocabulary. Pragmatic processing, on the other hand, is concerned with the inappropriate use of language in social situations. Deficits in this domain may include difficulties in understanding questions, in following conversations, and in identifying the key points of a story, getting lost in the details. It is likely that the semantic and pragmatic levels are interdependent, and that semantic deficits in word finding may contribute to pragmatic deficits that lead, for example, to the problem of maintaining the topic of a conversation [7].

In the recent years, there has been a growing interest from the research community in the computational analysis of language impairment in AD. Overall, existing studies targeted the automatic assessment of lexical, syntactical, and semantic deficits through an extensive amount of linguistic features [8, 9, 10]. More recently, semantic changes have been also investigated through vector space model representations [11, 12]. On the other hand, up to our knowledge, there are no works facing language impairments at an higher level of processing, considering macro-linguistic aspects of discourse production such as cohesion and coherence. While cohesion expresses the semantic relationship between elements, coherence is related to the conceptual organization of speech, and may be analyzed through the study of local, global, and topic coherence.

In this work, we investigate the possibility of automatically discriminate AD exploring a novel approach, based on the analysis of topic coherence. To this end, we model discourse transcripts into graphs encoding a hierarchy of topics on which we compute a relatively small set of pragmatic features. In the following, in Section 2, topic coherence analysis is briefly introduced, followed by an overview of the current state of the art for AD classification. Then, in Section 3 and 4, we present the dataset used in this study and a description of our methodology. Finally, the features related with topic coherence and classification results are reported in Section 5 and 6, respectively. Conclusions are summarized in Section 7.

2. Related work

2.1. Introduction to topic coherence analysis

The notion of topic and subtopic was introduced in 1991 in the work of Mentis & Prutting [13], whose target was an analysis of topic introduction and maintenance. A topic was defined as a clause identifying the question of immediate concern, while a subtopic being an elaboration or expansion of one aspect of the main topic.

Several years later, Bradie *et al.* [14] analyzed topic coherence and topic maintenance in individuals with right hemisphere brain damage. This work extends the one of Mentis & Prutting [13] with the inclusion of the notion of sub-subtopic and sub-sub-subtopic. Topic and sub-divisional structures were further categorized as new, related, or reintroduced.

In a following study Mackenzie *et al.* [15] used discourse samples elicited through a picture description task to determine the influences of age, education, and gender on the concepts and topic coherence of 225 healthy adults. Results confirmed education level as a highly important variable affecting the per-

10.21437/IberSPEECH.2018-59

formance of healthy adults.

More recently, Miranda [16] investigated the influence of education in the macro-linguistic dimension of discourse evaluation, considering concepts analysis, local, global and topic coherence, and cohesion. The study was performed on a population of 87 healthy, elderly Portuguese participants. Results corroborated the ones obtained by Mackenzie *et al.* [15], confirming the effect of literacy in this type of analysis.

2.2. Computational work for AD classification

From a computational point of view, language impairment in AD has been extensively assessed through the analysis of linguistic and acoustic features. Among the various works, we mention the study of Fraser *et al.* [8], where the authors considered more than 350 features to capture lexical, syntactic, grammatical, and semantic phenomena. Using a selection of 35 features the authors achieved state of the art classification accuracy of over 81% in distinguishing individuals with AD.

Yancheva *et al.* [11] presented a generalizable method to automatically generate and evaluate the information content conveyed from the description of the Cookie Theft picture. The authors created two cluster models, one for each group, from which they extracted different semantic features. Classification accuracy results achieved an F-score of 0.74. By combining semantic features to the set of lexicosyntactic features used in the work of Fraser *et al.* [8] the F-score improves to 0.80.

To the extent of our knowledge, the first work approaching coherence and cohesion computationally is the study of dos Santos *et al.* [17], although the target of the authors is the detection of Mild Cognitive Impairment (MCI). To this purpose, they model discourse transcripts with a complex network enriched with word embeddings. Classification is performed using topological metrics of the network and linguistic features, among which referential cohesion. Accuracy varies among 52%, 65%, and 74%, depending on the dataset used.

3. The Cookie Theft corpus

Data used in this work are obtained from the DementiaBank database[1], which is part of the larger TalkBank project [18, 19]. The collection was gathered in the context of a yearly basis, longitudinal study; demographics data, together with the education level, are provided. Participants included elderly controls, people with MCI and different type of Dementia. Among other assessments, participants were required to provide the description of the Cookie Theft picture, shown in Figure 1. Each speech sample was recorded and then manually transcribed at the word level following the TalkBank CHAT (Codes for the Human Analysis of Transcripts) protocol [20]. Data are in English language.

For the purposes of this study, only participants with a diagnosis of AD were selected, resulting in 234 speech samples from 147 patients. Control participants were also included, resulting in 241 speech samples from 98 speakers.

4. Modeling discourse transcripts as a hierarchy of topics

The topics used during discourse production may be subject to an internal, structural organization in order to achieve an information hierarchy. This organizational structure allows a gradual organization of information that is essential for an effective communication [22]. In fact, being important for both the speaker and the listener, this type of organization highlights the key concepts and indicates the degrees of importance and relevance within the discourse. Mackenzie *et al.* [15] in their work provided an example of a topic hierarchy based on the Cookie Theft picture description task, which was later extended in the study of Miranda [16]. To better understand the problem at hand, an excerpt of this hierarchy is also reported in Figure 2.

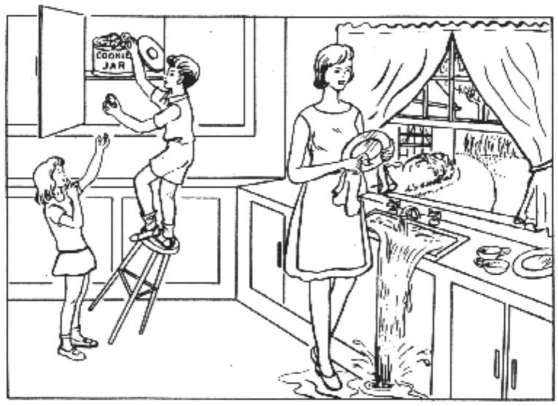

Figure 1: *The Cookie Theft picture, from the Boston Diagnostic Aphasia Examination [21].*

The number of topics that can be described observing the Cookie Theft picture, is somehow limited to the concepts that are explicitly represented in the scene, and to those ones that can be inferred from the previous (e.g., climate). Taking this into account, the problem of building a topic hierarchy from a transcript can be modeled with a semi-supervised approach in which a predefined set of topics clusters is used to guide the assignment of a new topic to a level in the hierarchy.

Both for the creation of the topics clusters, and for the analysis of a new discourse sample, a multistage approach is used to prepare, enhance, and transform the original transcriptions in a representation suitable for the subsequent analysis. Initially the transcriptions are preprocessed, then syntactical information is used to separate sentences into clauses and to identify coreferential expressions. Finally, we compute the vector representation of each clause by averaging the embeddings extracted for each word. To build the graph representing the topic hierarchy we develop an algorithm based on the cosine similarity that first evaluates the membership of a clause to the topic clusters, and then assigns the clause to a node in the hierarchy. We account for new and repeated topics. Each stage of this process is better described in the following sections.

4.1. Preprocessing

The Cookie Theft corpus provides the textual transcriptions of the participant's speech samples together with the morphological analysis and an extensive set of manual annotations (i.e., disfluencies, pauses, repetitions, and other more complex events). Among these, *retracing* and *reformulation* are used to indicate abandoned sentences where the speaker starts to say something, but then stops. While in the former the speaker may maintain the same idea changing the syntax, the latter involves a complete restatement of the idea. In order to prepare the transcriptions for the next stage of the pipeline, all the annotations were removed, and in the case of a retrace or a reformulation, also

[1] https://dementia.talkbank.org

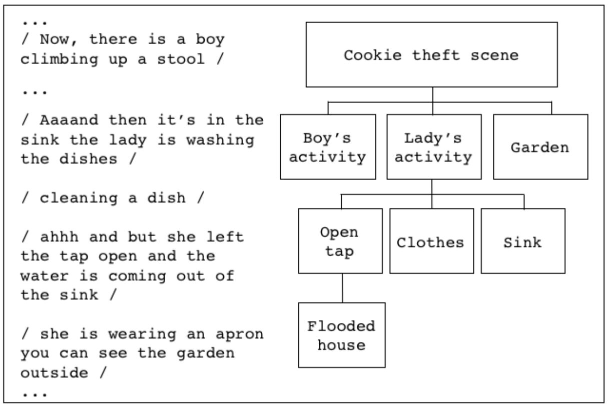

Figure 2: *An excerpt of a topic hierarchy for the Cookie Theft picture found in the work of Miranda [16] (text was translated from Portuguese).*

the text marked as being substituted was ignored. Additionally, disfluencies were disregarded and contractions were expanded to their canonical form. At this stage of processing, stopwords were not removed. Once the preprocessing phase is concluded, Part of Speech (POS) tags are automatically generated using the lexicalized probabilistic parser of the Stanford University [23].

4.2. Clause segmentation

The next step requires to face the problem of identifying dependent and independent clauses. In fact, while the Cookie Theft corpus already provides a segmentation of the input speech into sentences, this is not sufficient for the purposes of this work. Complex, compound or complex-compound sentences may contain references to multiple topics. The following excerpt shows an example of this problem, a complex sentence composed of a dependent and an independent clause: */and the mother is washing dishes while the water is running over in the sink on the floor /.*

A possible way to cope with the separation of different sentences types is by using syntactic parse trees. Thus, in a similar way to the work of Feng et al. [24], POS tags are used for the identification of dependent and independent clauses. For the former, the tag SBAR is used, while for the latter, the proposed solution checks the sequence of nodes along the tree to verify if the tag S or the tags [NP VP] appear in the sequence[2].

4.3. Coreference analysis

The analysis of coreference proves to be particularly useful in higher level NLP applications that involve language understanding, such as an extended discourse [25]. Strictly related with the notions of anaphora and cataphora, coreference resolution goes beyond the relation of dependence implicated by these concepts. It allows to identify when two or more expressions refer to the same entity in a text. In this work, the analysis of coreference has been performed with the Stanford coreference resolution system [26], using the results of the segmentation performed in the previous step. During the process of building the hierarchy, the coreference information is used to guide the assignment of a subtopic to the corresponding level in the hierarchy. To this purpose, we constraint the results provided by the coreference system to those mentions whose referent is the subject of the sentence. We are not interested in considering other coreferential expressions because a subtopic, being a specialization of a topic, is typically referred to the subject of the sentence.

[2]http://www.surdeanu.info/mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html

4.4. Sentence embeddings

In the last step of the pipeline, discourse transcripts are transformed in a representation suitable to compare and measure differences between sentences. In particular, the transformed transcripts should be robust to syntactic and lexical differences and should provide the capability to capture semantic regularities among sentences. To this purpose, we rely on a pre-trained model of word vector representations containing 2 million word vectors, in 300 dimensions, trained with fastText on Common Crawl [27]. In the process of converting a sentence into its vector space representation, we first perform a selection of four lexical items (nouns, pronouns, verb, adjectives), then, for each word we extract the corresponding word vector and finally we compute the average over the whole sentence.

4.5. Topic hierarchy analysis

To create a topic hierarchy from a transcript, we follow a methodology that is partly inspired by current clinical practice. Thus, in modelling the problem we do not want to impose a predefined order or structure in the way topics and subtopics may be presented, as this, of course, will depend on how the discourse is organized. However, we can take advantage of the closed domain nature of the task to define a reduced number of clusters of broad topics that will help to guide the construction of the hierarchy and the identification of off-topic clauses.

4.5.1. Topic clusters definition

As mentioned, the proposed solution relies on the supervised creation of a predefined number of clusters of broad topics. Each cluster contains a representative set of sentences that are related with the topic of the cluster. 10 clusters were defined: *main scene, mother, boy, girl, children, garden, climate, not-related, incomplete,* and *no-content*. The cluster *not-related* was used to model those sentences in which the participant is not performing the task (e.g., questions directed to the interviewer). The clusters *incomplete* and *no-content* are instead used to explicitly model sentences that may be characteristics of a language impairment. The former contains fragments of text that do not represent a complete sentence (e.g., /overflowing sink/), the latter identifies those expressions that do not add semantic information about the image (e.g., /fortunately there is nothing happening out there/, /what is going on/). To build the clusters, 30% of the data from both the AD and the control group is used. Each sentence has been manually annotated with the corresponding cluster label and clusters are simply modelled by the complete set of sentences belonging to them.

4.5.2. Topic hierarchy building algorithm

The algorithm to build the topic hierarchy relies on the cosine similarity between sentence embeddings. The first step consists in verifying to which topic cluster belongs the current sentence. This is achieved by computing the cosine similarity between the current sentence embeddings and each sentence embeddings in each topic clusters. The highest result determines the cluster for the new sentence. In the following step, we need to assign the current sentence embeddings to a level in the current hierarchy. This implies to establish if we are dealing with a new or a repeated topic and its level of specialization (i.e., subtopic, subsubtopic, etc.). This is achieved by first identifying, in the current hierarchy, the sub-graph whose nodes belong to the same cluster of the current sentence (e.g., the sub-graph corresponding to the *mother* cluster). Then, we compute the cosine similarity between the current sentence and each nodes of this subgraph. The new sentence is considered a *son* of the closest node if the similarity is lower than a threshold. Otherwise, it is considered a repeated topic. If there is no sub-graph, the sentence embedding is added as a new topic. If the new topic results to be a coreferential expression, this kind of information supersede the cosine metric strategy, and the new topic is added directly as a *son* of its referent.

Although the algorithm developed resembles the analysis performed in the standard clinical practice, the aim of this work is not the comparison of the automatic method with the manual one. Instead, our focus is understanding if pragmatic features related with topic coherence analysis may be relevant to discriminate AD. The type of features computed, as well as the results of classification experiments are described in the following sections.

5. Topic coherence features

Through the multistage approach and the final hierarchy of topics we identified sixteen measurements: (1-4) the number of topics, subtopics, sub-subtopics and sub-sub-subtopics introduced, (5-6) the proportion of dependent and independent clauses to the total number of sentences, (7) the total number of coreferential mentions, (8) the total number of topics, subtopics, sub-subtopics and sub-sub-subtopics repeated, (9-11) the number of sentences that were classified as not-related, incomplete, or no-content in the first step of the main algorithm, (12) the coefficient of variation (the ratio of the standard deviation to the mean) of the cosine similarity between two temporally consecutive topics, (13) the length of the longest path from the root node to all leaves, (14) the average number of outgoing edges of all nodes, (15) the total number of sentences, (16) the ratio of dependent to independent clauses.

6. Results and discussion

Classification experiments were performed with a Random Forest classifier, using the 70% of the remaining data of the Cookie Theft corpus, once that 30% of the data was retained to model the topic hierarchy. A stratified *k*-fold cross validation per subject strategy was implemented, with *k* being equal to 10.

Initial results, using the set of features described previously, provided an average accuracy of 74% in distinguishing AD patients from healthy controls. Then, in order to understand the importance of each feature, we implemented a forward feature selection method. This is an iterative approach in which the model is trained with a varying number of features. Starting with no features, at each iteration we test the accuracy of the model by adding, one at a time, each of the features that were not selected in a previous iteration. The accuracy is evaluated with a stratified 10-fold cross validation. The feature that yields the best accuracy is retained for further processing. The results of this method are shown in Figure 3. With this approach, we identified the first six features as the most relevant in discriminating the disease.

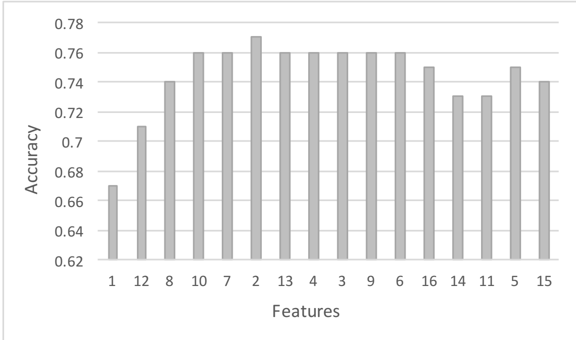

Figure 3: *Variation of the classification accuracy while increasing the number of features.*

Our results, achieved with the 70% of the data, provided an average accuracy of 77%, and an average F-score of 77% in classifying AD. Interestingly, the number of topics was the first feature selected, providing, alone, an average accuracy of 67%. Comparing these results with current state of the art, we acknowledge that Fraser *et al.* [8] achieved a higher accuracy (81%) using a set of lexicosyntactic features. On the other hand, we also recognized that these results are slightly better than the ones achieved by Yancheva *et al.* [11] (F-score 74%) using only a set of 12 semantic features. However, when the authors combine lexicosyntactic and semantic information, the F-score improves to 80%. These considerations are interesting for multiple reasons, in fact, on one side they confirm the relevance of pragmatic features related with topic coherence in the task of classifying AD. On the other hand, they also highlight that lexicosyntactic features are extremely important in characterizing the disease and should be used in a complementary way with other features.

7. Conclusions

In this work, we approached the problem of exploting topic coherence analysis to automatically classify AD. To this purpose, we proposed an algorithm inspired by the type of assessment conducted by clinicians to construct the topic hierarchy of a picture description task, from which we extract a reduced set of pragmatic features for automatic classification. Initial experimental results show comparable AD classification performance to current state of the art approaches using different types of consolidated linguistic features. As future work, we plan to integrate the proposed pragmatic features with lexicosyntactic features and to explore the extension of this kind of analysis to other types of discourse production tasks, including open-domain tasks.

8. Acknowledgments

The authors want to express their gratitude to Dr. Filipa Miranda for her precious advice and help provided during the development of this work. This work was supported by Portuguese national funds through – Fundação para a Ciência e a Tecnologia (FCT), under Grants SFRH/BD/97187/2013 and Project with reference UID/CEC/50021/2013.

9. References

[1] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of alzheimers disease," *Alzheimer's & dementia*, vol. 3, no. 3, pp. 186–191, 2007.

[2] K. E. Forbes, A. Venneri, and M. F. Shanks, "Distinct patterns of spontaneous speech deterioration: an early predictor of alzheimer's disease." *Brain Cogn*, vol. 48, no. 2-3, pp. 356–361, Mar-Apr 2002.

[3] J. Reilly, J. Troche, and M. Grossman, "Language processing in dementia," *The handbook of Alzheimer's disease and other dementias*, pp. 336–368, 2011.

[4] V. Taler and N. A. Phillips, "Language performance in alzheimer's disease and mild cognitive impairment: a comparative review." *J Clin Exp Neuropsychol*, vol. 30, no. 5, pp. 501–556, Jul 2008.

[5] G. Oppenheim, "The earliest signs of alzheimer's disease." *J Geriatr Psychiatry Neurol*, vol. 7, no. 2, pp. 116–120, Apr-Jun 1994.

[6] D. Kempler, "Language changes in dementia of the alzheimer type," *Dementia and communication*, pp. 98–114, 1995.

[7] S. H. Ferris and M. R. Farlow, "Language impairment in alzheimers disease and benefits of acetylcholinesterase inhibitors," in *Clinical interventions in aging*, 2013.

[8] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech." *J Alzheimers Dis*, vol. 49, no. 2, pp. 407–422, 2016.

[9] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden, "Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances," in *Proceedings of the 1st Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*. sn, 2014, pp. 78–87.

[10] M. Yancheva, K. C. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for alzheimer's disease and related dementias," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 134–139.

[11] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting alzheimer's disease." in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2337–2346.

[12] K. C. Fraser and G. Hirst, "Detecting semantic changes in alzheimer's disease with vector space models," in *Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016), Monday 23rd of May 2016*, no. 128. Linköping University Electronic Press, Linköpings universitet, 2016, p. 1 to 8.

[13] M. Mentis and C. A. Prutting, "Analysis of topic as illustrated in a head-injured and a normal adult," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 3, pp. 583–595, 1991.

[14] M. Brady, C. Mackenzie, and L. Armstrong, "Topic use following right hemisphere brain damage during three semi-structured conversational discourse samples," *Aphasiology*, vol. 17, no. 9, pp. 881–904, 2003.

[15] C. Mackenzie, M. Brady, J. Norrie, and N. Poedjianto, "Picture description in neurologically normal adults: Concepts and topic coherence," *Aphasiology*, vol. 21, no. 3-4, pp. 340–354, 2007.

[16] F. Miranda, "Influência da escolaridade na dimensão macrolinguística do discurso," Master's thesis, Universidade Católica Portuguesa, Instituto de ciências da saúde, 2015.

[17] L. Santos, E. A. Corrêa Júnior, O. Oliveira Jr, D. Amancio, L. Mansur, and S. Aluísio, "Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017, pp. 1284–1296. [Online]. Available: http://www.aclweb.org/anthology/P17-1118

[18] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[19] B. MacWhinney, S. Bird, C. Cieri, and C. Martell, "Talkbank: Building an open unified multimodal database of communicative interaction," *4th International Conference on Language Resources and Evaluation*, pp. 525–528, 2004.

[20] M. B, "The childes project: Tools for analyzing talk, 3rd edition." Lawrence Erlbaum Associates, Mahwah, New Jersey, 2000.

[21] H. Goodglass, E. Kaplan, and B. Barresi, *The Boston Diagnostic Aphasia Examination*, Baltimore: Lippincott, Williams & Wilkins, 2001.

[22] H. Ulatowska and S. Chapman, *Discourse Analysis and Applications. Studies In Adult Clinical Populations*. Hillsdale: Lawrence Elbaum Associates, 1994, ch. Discourse macrostructure in aphasia, pp. pp. 29–46.

[23] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 423–430.

[24] S. Feng, R. Banerjee, and Y. Choi, "Characterizing stylistic elements in syntactic structure," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 1522–1533.

[25] S. Boytcheva, P. Dobrev, and G. Angelova, "Cgextract: Towards extraction of conceptual graphs from controlled english." Contributions to ICCS-2001, 9th International Conference of Conceptual Structures,, 2001.

[26] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[27] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.