REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2022 PUCV, Chile ⇝ 55(110) 1045-1056

Storytelling, a(nother) Discourse Elicitation Instrument: Protocol and Materials

Narración de historias, otro instrumento de elicitación del discurso: protocolo y materiales

María Francisca Alonso-Sánchez
UNIVERSIDAD DE VALPARAÍSO
CHILE
mariafrancisca.alonso@uv.cl

Pedro Alfaro-Faccio
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
pedro.alfaro@pucv.cl

Héctor Allende-Cid
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
hector.allende@pucv.cl

Juan Zamora
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
juan.zamora@pucv.cl

Recibido: 02-IV-2022 / **Aceptado:** 07-XI-2022
DOI: 10.4067/S0718-09342022000301045

Abstract

The study of oral narrative discourse allows for the description of several features of the speakers (e.g., belonging to social or age groups, pathologies or special educational needs, stages of acquisition or learning, among many other attributes). This type of semi-structured tasks also allows speakers to generate ecological discourses with comparable lexical and semantic structures, as well as being easily replicable independently of the examiner. In this framework, we present a retelling instrument composed of three stories made from 15 static images each based on a central event. In order to demonstrate its usefulness, we presented an automatic analysis of discourses generated by 50 university students and 13 people diagnosed with schizophrenia. The results showed homogeneity of the texts, based on the comparison of the types of words prompted. We observed that one of the stories generates more abstract content, which makes it especially useful for the study of certain populations.

Keywords: Discourse, oral narratives, retelling, natural language processing, language assessment.

Resumen

El estudio del discurso narrativo oral permite describir diversas características de los hablantes (p. ej., pertenencia a grupos sociales o etarios, patologías o necesidades educativas especiales, etapas de la adquisición o el aprendizaje, entre muchas otras cualidades). Este tipo de tareas semiestructuradas permite, también, que los hablantes generen discursos ecológicos y con estructuras léxicas y semánticas comparables, además de ser fácilmente replicables de forma independiente del evaluador. En este marco, presentamos una herramienta de recontado compuesta por tres historias construidas a partir de 15 imágenes estáticas cada una basada en un evento central. A fin de demostrar su utilidad, presentamos un análisis automático de discursos generados por 50 estudiantes universitarios y 13 personas diagnosticadas con esquizofrenia. Los resultados indicaron la homogeneidad de los textos, a partir de la comparación de los tipos de palabras que los constituyen. Asimismo, se determinó que una de las historias genera contenido más abstracto, lo que la hace especialmente útil para el estudio de ciertas poblaciones.

Palabras Clave: Discurso, narración oral, recontado, procesamiento del lenguaje natural, evaluación del lenguaje.

INTRODUCTION

In recent years, language and speech have been in the spotlight of neurological and psychiatric disorders because of their potential use as a biomarker (de Boer, Brederoo, Voppel & Sommer, 2020; Palaniyappan, 2021; Barron, Baker, Budde, Bzdok, Eickhoff et al., 2021). Examples of these are pathologies that involve cognitive syndromes, such as Neurocognitive Disorders (Hoffman, Sajjadi, Patterson & Nestor, 2017) and Psychosis (Corcoran & Cecchi, 2020), or motor features by the acoustics patterns as in Parkinson's disease (Amato, Borzì, Olmo & Orozco-Arroyave, 2021). However, the development of language assessment methods is highly specific to certain pathologies diagnosed based on language patterns. For example, Aphasias have historically been assessed by particular tasks targeting linguistic macro-process (e.g., Repetition) (Shewan & Kertesz, 1980) while Developmental Language Disorders assessment (Leonard, 1982) focuses on language levels (e.g., Phonology). These methods are not transferable to other pathology assessments.

Advances in computational linguistics allow us to explore a variety of language features (e.g., word frequency, syntactic organization, or semantic dimensionality) during naturalistic/ecological tasks. For instance, text representation recognizes patterns that are not detectable for human evaluation through several methods (term-frequency/inverse document frequency, N-grams, Bag-of-words or count-based models). Computational linguistic methods also allow us to build automatic classifiers that make predictions by assigning classes (for example, with/without pathology) to oral or written productions. In particular, the generation of the classifiers (learning machines) is based on mathematical-computational models trained with a set of pre-classified text examples. As more sophisticated and robust methods are available, we increase the potential opportunities for the disorder's phenomenological description and early detection, but caution is required; biases and reductionism may swamp us (Rezaii, Wolff & Price, 2022). Despite the advantages and disadvantages of these methods, a critical point emerges regarding the collection of data to be analyzed (i.e., garbage in = garbage out). Data collection must be unbiased, blind, and even

(Pannucci & Wilkins, 2010), but it also should balance ecological tasks, be hypothesis-driven, and have large-scale applicable protocols to solve cross-linguistic questions, all of which is a challenge.

When studying discourse, the elicitation task is a key component of the outcome interpretation, especially if we are using coherence, similarity, or semantic network metrics. Among structured methods, narratives are either from a single picture, a series of pictures or storyboard, a personal experience/dream, or a story retelling. However, comparing semantic metrics between dream reports may not be very successful since the background content is not comparable. Instead of examining language, we may end up exploring the detail with which dreams are recalled and even how bizarre the dreams are. To use any of the semantic metrics, the content of the speech must be controlled (e.g., the same number of characters or events). Moreover, the level of abstraction of the content and the use of metaphors will be very relevant in comparing content. In the same line, the syntactic complexity from a single picture description may be terse and with limited variability while a storyboard enables us to elaborate more extensively. Therefore, the different types of stimuli can produce very different results independent of the participant's speech skills. Furthermore, the requirement of cognitive resources is widely different between retelling a story previously presented and the narration of a personal episodic event (e.g., working memory versus episodic memory). Also, there is a cultural background influence on the well-known story task (e.g., Cinderella is popular in Western but not in Eastern cultures or there is bias risk related to—current—women's role in society) (McCabe, 1997). For instance, the 'Cookie Theft' and the 'Dinner party' tasks are widely used for aphasic discourse elicitation. The first is a picture description (Goodglass & Kaplan, 1983) while the second is a pictorial script of eight black and white cartoons (Fletcher & Birt, 1983). Some reports show higher diversity of content and richer descriptions using storytelling than picture descriptions (Alyahya, Halai, Conroy & Lambon Ralph, 2020). However, the 'Dinner party' task reveals most of the story content during the instructions: *"Mr. Smith invited his boss, Mr. Plummer to dinner - but they forgot about the cat... Look at the pictures and tell the story. Here are some words to help you: to invite, a salmon, to lay the table, to be horrified, to rush out"*. This intervention from the interviewer generates an influence on the narration and even on the comprehension of the storyboard. The task must be designed for valid comparisons in terms of content, thematic progression, cognitive load, and cultural background but should enable the participant to develop the story independently. In this context, this study aims to generate a framework for a valid discourse task. To this end, we first present the materials: three stories in a series of pictures and the protocol. Then, we show the features extracted from a healthy sample, and, finally, we summarize the linguistic features of a sample of patients with schizophrenia previously reported (Allende-Cid, Zamora, Alfaro-Faccio & Alonso-Sánchez, 2019).

1. Methods

1.1. Corpus

To collect the corpus of oral discourse, a semi-structured narration elicitation task was developed. This task consisted of three stories, each one made up of 15 static images without text—created by the Chilean illustrator Pedro Prado for this study. The content of these three stories is related to each other, based on a central event around which they take place: an earthquake that occurred in a port city. The first story is about a man who returns home concerned about his family's well-being; the second story is about a woman who captains a ship and must guide her team; the third is about an old lightkeeper who must help a ship during the night. These three stories were designed with the same structure, length and graphic style. The material was presented in a letter sized colour image book (Appendix 1). The authors have the copyright of this material.

This type of task was chosen for two reasons. Firstly, previous research has shown that the narrative discourse —as opposed to others such as descriptive, explanatory, argumentative, and instructive discourse—is more easily mastered, both ontogenetically, that is, through language acquisition in childhood and during aging, as through language and cognitive pathologies (Schneider, 1996; Gazella & Stockman, 2003; Isbell, Sobol, Lindauer & Lowrance, 2004). Secondly, semi-structured tasks allow the speakers to maintain the topic and, consequently, the type of lexical and syntactic structures through the generated texts. This provides lexical and syntactic homogeneity to the corpus and allows subsequent comparisons among texts (Eisenbeiss, 2010).

1.2. Participants

These three narrative tasks were performed by a group of 50 university students with more than 14 years of formal education, without psychiatric, neurological, sensory, or language pathologies. Then, 13 participants diagnosed with schizophrenia completed the same task. All patients included were diagnosed by the clinical team of the Servicio de Salud (Chile) independently of this study and were on stable medication according to their treating physician. In the patient group, ages ranged from 19 to 74 years (x = 37.8 σ =19.6) and a range of schooling between 8 and 12 years of formal education. Patients with psychiatric and neurological comorbidities were excluded. The variability of the sample was planned to represent the wide age spectrum of patients. All participants provided written informed consent before assessment and ethics approval was granted by the Human Research Ethics Board at Universidad Santo Tomás, Viña del Mar, Chile.

1.3. Procedures

The assessment was done individually by a trained research assistant authorized by the Ethics Review Board. The participants were requested to see the images and try to understand the story told therein. Then, they were asked if they had understood the story and, if necessary, they were able to see the book as many times as they wished and for as long as necessary. After stating that they had understood the story, they were asked to retell it orally with the images on display. This procedure sought to ensure their understanding of the story and to avoid the influence of working memory.

The oral narratives were audio recorded—prior authorization using an informed consent of the participants—and later transcribed orthographically. This procedure resulted in 150 oral narratives produced by the student group and 39 produced by the group of participants with schizophrenia.

1.4. Text processing

Each oral narration was transcribed, digitalized, and processed by extracting and counting Part-Of-Speech (Pos) tags. The total number of labels corresponded to nine types of words: adjectives, verbs, adpositions, adverbs, conjunctions, pronouns, nouns, determiners, and interjections. All these tags denote the kind of linguistic information that the automatic classifiers will use. The abstract versus concrete content of the text was done with a logistic regression from sklearn (Python), trained with a set of 50 nouns related to the topic of the stories. After the Pos tag, nouns were selected and classified with the binary regression. We assessed the logistic regression model performance with the confusion matrix.

1.5. Data analysis

We analyzed the distribution of word types between stories, comparing an alternative model (difference between groups) to a null model (no difference between groups) with Bayesian ANOVA, considering the participant as a random factor of the model. The priors of the model were set by default with an r-scale for a fixed effect of 0.5 and a random effect of 1. Post hoc analyses were made with a posterior odd corrected for multiple testing by fixing to 0.5 the prior probability that the null hypothesis holds across all comparisons. For categorical variables (abstract versus concrete), we computed the amount on each story and contrasted it with a Bayesian independent multinomial comparison. We performed a Bayesian t-test for each Pos tag variable in the group comparison. In this model comparison, the alternative hypothesis specifies that the student group is greater than the patient group. For the interpretation of the Bayes Factor on each model we consider, as Jeffreys proposed, BF_{ij}= 1-3: Not worth more than a bare mention, 3-10 Substantial evidence for H_j, 10-

30 Strong evidence for H_j, 30-100 Very strong evidence for H_j, and >100 Decisive evidence for H_j (Lee, 2004; Jarosz & Wiley, 2014).

2. Results

2.1. Pos tag comparison between stories

In story comparison, there was no difference in the use of adjectives, verbs, adpositions, adverbs, conjunctions, pronouns, determiners, or interjections (anecdotal degree of evidence to support the alternative model hypothesis). These results are shown in Table 1.

Table 1. Comparison between stories A, B, and C. The data are presented with means and standard deviations. Alternative model Bayes Factor (BF$_{10}$) with Post Hoc (PH) analysis is reported.

	A		B		C		Model comparison			
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**BF$_{10}$**	**PH a-b**	**PH a-c**	**PH b-c**
Adjectives	8.4	5.9	10.25	5.81	8.56	5.69	2.01	0.55	0.23	0.48
Verbs	60.9	42.2	55.9	32	57.3	27.2	0.17	0.27	0.25	0.23
Adposition	33.7	20.6	32.2	20.9	32	17.3	0.11	0.24	0.24	0.23
Adverbs	16	18.4	15.1	12	15.3	9.52	0.09	0.24	0.23	0.23
Conjunctions	23.1	20.8	21	13.6	22.8	12.8	0.13	0.26	0.23	0.27
Pronouns	24.9	18.7	26	15	24.6	11.9	0.1	0.24	0.23	0.25
Nouns	48.8	31.8	47.3	27.3	48.7	22	0.09	0.23	0.23	0.23
Determiners	43.2	24.4	42.5	23.7	44.5	20.4	0.1	0.23	0.23	0.24
Interjections	5.48	6.02	4	2.81	3.63	2.52	0.7	0.47	0.66	0.3

2.2. Abstraction/concreteness comparison between stories

The independent multinomial comparison of abstraction/concreteness between stories showed strong evidence supporting the alternative model over the null model (BF$_{10}$= >10.000). As shown in Figure 1, the comparison among the stories showed a wide difference in the use of abstract and concrete words, the BF$_{10}$ was higher than 10.000 for all the combinations (A-B, A-C, B-C).

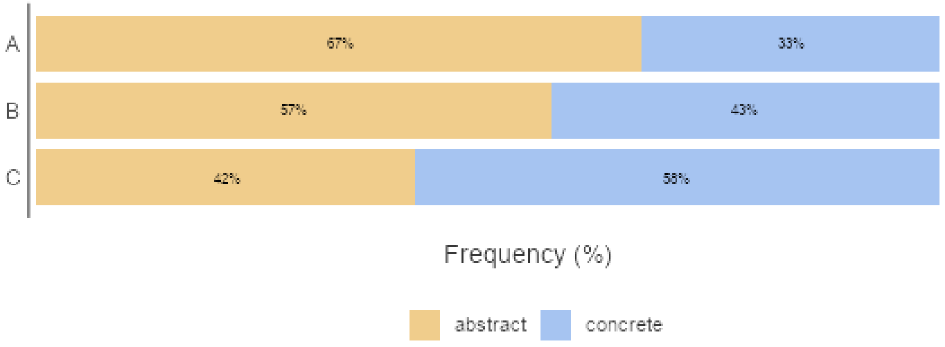

Figure 1. Use of abstract vs concrete words comparison between stories.

2.3. Comparison between groups

In the count of word types, the control and patient groups did not differ in the number of adjectives, conjunction, determiners, nouns, pronouns, adverbs, or interjections (anecdotal evidence for the alternative hypothesis model) as is shown in Table 2. Although differences appeared between the two groups in most of the features, those that exhibited higher contrasts were adpositions, verbs, and interjections, but only the adposition showed more than anecdotal evidence of a difference between groups (Figure 2).

Table 2. Comparison between groups. The data is presented with means, standard deviations, and 95% credible interval. Alternative model Bayes Factor (BF_{10}) with Effect size (δ-95% Credible interval) analysis is reported.

	Control		95% Credible Interval		Patients		95% Credible Interval		BF_{10}	δ-95% CI
	Mean	SD	Lower	Upper	Mean	SD	Lower	Upper		
Adjective	9.07	5.82	8.01	10.1	8.3	8.22	5.52	11	0.348	0.010, 0.487
Conjunction	22.3	16	19.4	25.2	21.6	18.1	15.6	27.6	0.236	0.007, 0.423
Determiner	43.4	22.7	39.3	47.5	41.8	39.4	29.1	54.6	0.25	0.007, 0.429
Noun	48.2	27.1	43.3	53.2	51.4	44.3	37	65.7	0.137	0.004, 0.336
Pronoun	25.1	15.3	22.4	27.9	27	17.3	21.2	32.8	0.132	0.004, 0.332
Adverb	15.4	13.72	12.9	17.9	18.6	14.9	13.6	23.5	0.097	0.003, 0.281
Adposition	32.6	19.56	29.1	36.1	24	21.6	16.8	31.1	4.281	0.075, 0.755
Verb	58	34.2	51.8	64.2	59.8	43.5	45.7	73.9	0.244	0.005, 0.362
Interjection	4.37	4.13	3.45	5.28	2.4	2.06	1.25	3.54	1.997	0.048, 0.967

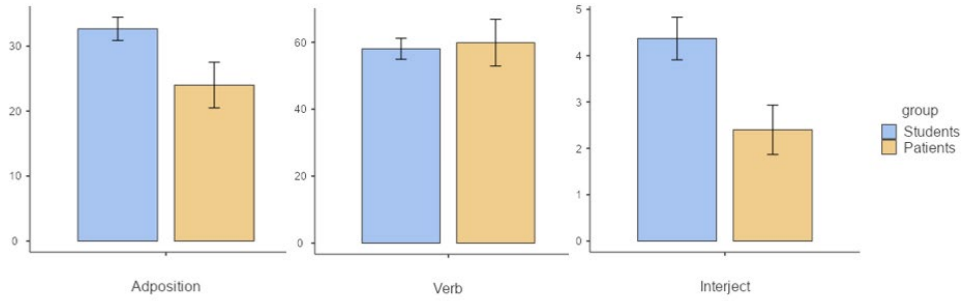

Figure 2. Barplots comparison between groups within adpositions, verbs and interjections.

3. Discussion

This study aimed to examine the suitability of three storyboards as a discourse elicitation method. To this end, we compared the Pos tag of 50 students and the amount of abstract and concrete words used in each narration. We report two major results. First, the three storyboards are comparable in eliciting the type of words. Second, the use of abstract or concrete words is different between stories.

The three stories generate similar linguistic structures and therefore the combination of the corpora extends the richness of the discourse sample. Thus, these corpora may be useful to compare the discourse of different populations with or without pathologies. Conversely, the stories elicited a different amount of concrete/abstract content. For instance, the story about the man who returns home concerned for the state of his family is the task that elicited a greater number of abstract words. The story about the woman who captains a ship and must guide her team was the most balanced elicitation of abstract and concrete words. Finally, the story about the old lightkeeper who must help a ship during the night generated the highest number of concrete words. Overall, the three stories are suitable for discourse analysis, but the difference in the level of abstraction may be worth noticing depending on the study population.

When telling a story, we use abstract representations to reference complex mental states (e.g., happiness), situations (e.g., conflicts), and relationships (e.g., seniority) while we use concrete words to talk about material objects (e.g., apple). Several reports have shown that concrete words are easier to recognize and elicit faster responses in lexical decision tasks but only when the context is not available (Barbe, Ottenb, Koustac & Vigliocco, 2013; Mkrtychian, Blagovechtchenski, Kurmakaeva, Gnedykh, Kostremina & Shtyrov, 2019). Storytelling based on an image sequence mostly relies on context so even though the three stories are showing contexts with the same structure, these stories have a unique potential to explore the inference of intentional and emotional states. However, caution is needed due to the overinterpretation that computational linguistic methods may generate about cognition. Although these tools allow us to observe features that other tools do not, we must be aware of their limitations. These methods are trained on another dataset that may have its own biases. Furthermore, the computational representation of texts is not necessarily a reflection of cognitive processing.

Notwithstanding the scope of this study was limited to the elicitation method, we also examined the speech elicitation application to a clinical sample (Allende-Cid et al., 2019). The most interesting finding was that the clinical sample did not differ in most PoS analyses. The adpositions were the only PoS that showed evidence of a difference between the groups. These findings must be interpreted with caution because we did not evaluate the effect of demographic and clinical features on speech performance since this was not the purpose of this research.

CONCLUSION

In this investigation, we introduced three stories for speech elicitation. These stories are suitable for comparison according to the type of words that they prompt. Story retelling is an outstanding data collection approach since it allows the creation of large corpora. This semi-structured technique enables the generation of samples based on an ecological task where spontaneous speech emerges. Moreover, a large corpus is more suitable for the use of computational techniques, which were previously almost exclusive for analyzing written texts. Finally, this approach has a high degree of reliability (inter-rater), supporting large-scale studies that allow for comparison across sites, cultures, and linguistic varieties. The authors of this study provide the materials freely available for non-commercial use.

REFERENCES

Allende-Cid, H., Zamora, J., Alfaro-Faccio, P. & Alonso-Sánchez, MF. (2019). A Machine Learning Approach for the Automatic Classification of Schizophrenic Discourse. *IEEE*, 7, 45544-45553. DOI: 10.1109/ACCESS.2019.2908620

Alyahya, R. S. W., Halai, A. D., Conroy, P. & Lambon Ralph, M. A. (2020). A Unified Model of Post-Stroke Language Deficits Including Discourse Production and Their Neural Correlates. *Brain, 143*(5), 1541-1554. DPI: https://doi.org/10.1093/brain/awaa074

Amato, F., Borzì, L., Olmo, G. & Orozco-Arroyave, J. R. (2021). An Algorithm for Parkinson's Disease Speech Classification Based on Isolated Words Analysis. *Health Information Science and Systems*, *9*(1), 1-15. DOI: https://doi.org/10.1007/s13755-021-00162-8

Barber, H. A., Ottenb, L., Koustac, S. T. & Vigliocco, G. (2013). Concreteness in Word Processing: ERP and Behavioral Effects in a Lexical Decision Task. *Brain and Language,* *125*(1), 47-53. DOI: https://doi.org/10.1016/j.bandl.2013.01.005

Barron, D. S., Baker, J. T., Budde, K. S., Bzdok, D., Eickhoff, S. B., Friston, K. J., Fox, P. T., Geha, P., Heisig, S., Holmes, A., Onnela, J. P., Powers, A., Silbersweig, D. & Krystal, J. H. (2021). Decision Models and Technology Can Help Psychiatry Develop Biomarkers. *Frontiers in Psychiatry*, *12*(September), 1-14. DOI: https://doi.org/10.3389/fpsyt.2021.706655

Corcoran, C. M. & Cecchi, G. A. (2020). Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(8), 770-779. DOI: https://doi.org/10.1016/j.bpsc.2020.06.004

de Boer, J. N., Brederoo, S. G., Voppel, A. E. & Sommer, I. E. C. (2020). Anomalies in Language as a Biomarker for Schizophrenia. *Current Opinion in Psychiatry*, *33*(3), 212-218. DOI: https://doi.org/10.1097/YCO.0000000000000595

Eisenbeiss, S. (2010). Production Methods in Language Acquisition Research. En E. Blom & S. Unsworth (Eds.), *Experimental Methods in Language Acquisition Research* (pp. 11-34). Ámsterdam: John Benjamins

Fletcher, M. & Birt, D. (1983). *Storylines: Picture Sequences for Language Practice*. London: Longman.

Gazella, J. & Stockman, I. J. (2003). Children's Story Retelling Under Different Modality and Task Conditions: Implications for Standardizing Language Sampling Procedures. *American Journal of Speech-Language Pathology*, *12*(1), 61-72. DOI: https://doi.org/10.1044/1058-0360(2003/053)

Goodglass, H. & Kaplan, E. (1983). *Boston Diagnostic Aphasic Examination* (2nd ed.). Pennsylvania: Lea & Febiger.

Hoffman, P., Sajjadi, S. A., Patterson, K. & Nestor, P. J. (2017). Data-Driven Classification of Patients with Primary Progressive Aphasia. *Brain and Language*, 174, 86-93. DOI: https://doi.org/10.1016/j.bandl.2017.08.001

Isbell, R., Sobol, J., Lindauer, L. & Lowrance, A. (2004). The Effects of Storytelling and Story Reading on the Oral Language Complexity and Story Comprehension of Young Children. *Early Childhood Education Journal*, *32*(3), 157-163. DOI: https://doi.org/https://doi.org/10.1023/B:ECEJ.0000048967.94189.a3

Jarosz, A. & Wiley, J. (2014). What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *Journal of Problem Solving*, 7, 2-9. DOI: https://doi.org/10.7771/1932-6246.1167

Lee, H. K. H. (2004). Priors for Neural Networks. En D. Banks, F. R. McMorris, P. Arabie & W. Gaul (Eds.), *Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation* (pp. 141-150). Berlin: Springer. DOI: https://doi.org/10.1007/978-3-642-17103-1_14

Leonard, L. B. (1982). Phonological Deficits in Children with Developmental Language Impairment. *Brain and Language*, *16*(1), 73-86. DOI: https://doi.org/10.1016/0093-934X(82)90073-6

McCabe, A. (1997). Cultural Background and Storytelling. *The Elementary School Journal*, *97*(5), 453-473.

Mkrtychian, N., Blagovechtchenski, E., Kurmakaeva, D., Gnedykh, D., Kostremina, S. & Shtyrov, Y. (2019). Mental Representations to Functional Brain Mapping. *Frontiers in Human Neuroscience,* 13, 267. DOI: https://doi.org/10.3389/fnhum.2019.00267

Palaniyappan, L. (2021). More than a Biomarker: Could Language Be a Biosocial Marker of Psychosis? *Npj Schizophrenia*, *7*(1), 13-15. DOI: https://doi.org/10.1038/s41537-021-00172-1

Pannucci, C. J. & Wilkins, E. G. (2010). Identifying and Avoiding Bias in Research. *Plastic and Reconstructive Surgery*, *126*(2), 619-625. DOI: https://doi.org/10.1097/PRS.0b013e3181de24bc

Rezaii, N., Wolff, P. & Price, B. (2022). Natural Language Processing in Psychiatry: The Promises and Perils of a Transformative Approach. *The British Journal of Psychiatry, 220*(5), 251-253. DOI:10.1192/bjp.2021.188

Schneider, P. (1996). Effects of Pictures Versus Orally Presented Stories on Story Retellings by Children with Language Impairment. *American Journal of Speech-Language Pathology*, *5*(1), 86-95. DOI: https://doi.org/10.1044/1058-0360.0501.86

Shewan, C. M. & Kertesz, A. (1980). Reliability and Validity Characteristics of the Western Aphasia Battery (WAB). *Journal of Speech and Hearing Disorders*, *45*(3), 308-324. DOI: https://doi.org/10.1044/jshd.4503.308