Detecting Psychological Disorders with Stylometry

Juan Barrios Rudloff, Simon Gabay, Florian Cafiero, Martin Debbané

Detecting Psychological Disorders with Stylometry: the Case of ADHD in Adolescent Autobiographical Narratives

Juan Barrios[1,*], Simon Gabay[1], Florian Cafiero[2] and Martin Debbané[1]

[1]*Université de Genève*
[2]*École nationale des chartes - PSL, Centre Jean Mabillon*
[1]*Université de Genève*

Abstract

Attention-deficit/hyperactivity disorder (ADHD) is one of the most common psychological neurodevelopmental disorder among children and adolescents, with a prevalence of 5.6% in teenagers aged 12 to 18 years [1]. Its diagnosis is reliable and valid when evaluated with standard criteria for psychiatric disorders [2], but it is time consuming and requires a high level of expertise to arrive at a correct differential diagnosis. The development of low-cost, fast and efficient tools supporting the ADHD diagnosis process would therefore be important for practitioners, because it should help identify and prevent risks in different populations.

In this paper, we study the possibility of detecting ADHD with Natural Language Processing (NLP), based on the analysis of a specific type of adolescent's autobiographical narratives called Self-Defining Memories (SDMs). (1) We train a Support Vector Machine (SVM) to predict ADHD diagnosis, (2) we attempt to explain its results by exploring lexical information (3) and unfolding the results of the SVM to identify and analyse the linguistic markers associated with each groups.

With an accuracy of 92%, the SVM manages to classify texts from both group (ADHD vs Control), revealing a signal specific to autobiographical texts narratives written by people with ADHD. The quality of the detection is confirmed by the interpretative yield of the main markers identified. However, several methodological improvements remain necessary to improve the accuracy and the automation of ADHD diagnosis with stylometric methods.

Keywords

Stylometry, NLP, Psychology, Psychological disorder, ADHD, Self-Defining Memories

1. Introduction

The assumptions that the "words we use in natural language [...] reveal a tremendous amount of information about our social interactions and personality" and that Natural Language Processing

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France
*Corresponding author.
juan.barrios@unige.ch (J. Barrios); simon.gabay@unige.ch (S. Gabay); florian.cafiero@chartes.psl.eu (F. Cafiero); martin.debbane@unige.ch (M. Debbané)
https://www.unige.ch/fapse/psychoclinique/unites/upcd/membres/juan-barrios-rudloff (J. Barrios); https://www.unige.ch/lettres/humanites-numeriques/equipe/collaborateurs/dr-simon-gabay (S. Gabay); https://sites.google.com/view/florian-cafiero (F. Cafiero); https://www.unige.ch/fapse/psychoclinique/unites/upcd/membres/debbane (M. Debbané)
0000-0003-0709-8306 (J. Barrios); 0000-0001-9094-4475 (S. Gabay); 0000-0002-1951-6942 (F. Cafiero); 0000-0002-4677-8753 (M. Debbané)

(NLP) can help us access to this information is now widely shared by linguists, philologists or psychologists [3]. With regard for psychological disorders, such a "revelation" is not enough: because ADHD poses significant risks [4], we need to efficiently assist practitioners to detect such disorders as automatically and as reliably as possible. This paper proposes therefore a case study to evaluate the use of a NLP technique, stylometry, for the detection and possibly the understanding of Attention-deficit/hyperactivity disorder (ADHD).

Research crossing NLP and psychology is mainly conducted on English native speakers, but some papers exist on romance languages (such as Spanish [5]) or Asian languages (such as Korean [6, 7]). It is precisely on Korean that the rare studies on ADHD concluding with a specific style for people prone to this disorder have been carried out. If a specific style exists, stylometry should make it possible to identify it, and therefore presents itself as a promising method for diagnosing ADHD, and thus preventing the risks associated with it.

1.1. Attention-deficit/hyperactivity disorder (ADHD)

ADHD is a neurodevelopmental disorder that affects 5.6% in teenagers aged 12 to 18 years [1]. Its causes are multifactorial: several genetic and environmental risk factors act together to increase both susceptibility to the disorder and the extent of psychiatric comorbidities [2]. Regarding its diagnosis, it is reliable and valid when evaluated with standard criteria for psychiatric disorders, but the expression of symptoms varies as a function of patient developmental stage in both the social and the academic contexts [2]. All this makes ADHD diagnoses a time consuming task (up to one working day) requiring a high level of expertise.

Furthermore, during adolescence, young people with ADHD are especially prone to experience difficulties in interpersonal relationships, characterised by conflicts within the family or at school. In the latter case, conflicts with classmates may in turn lead to peer rejection, social isolation or school failure [8, 9, 10]. Besides these relational problems, adolescents with ADHD are also exposed to higher risks of neurocognitive dysfunction, substance abuse, low self-esteem and social disability [2]. In summary, this leads to a situation in which a complex diagnosis with extremely serious health and social implications for patients must be made, involving time and resources that are not always available and in a context of high prevalence.

1.2. Psychology and Natural Language Processing (NLP)

For more than ten years now, the use of NLP has established itself as a valid method for different types of clinical applications in neuroscience and psychiatry [11], which explains an upward trend for this approach and a growing need to develop new detection methods [12]. Clinical applications include psychological profiling [13, 6], or the detection of depression [14, 15] and psychological distress [16] with different methods: word embeddings, sentiment analysis or stylometry. The latest research makes it possible not only to identify mental health risks (like patients who exhibit suicidal risk behaviour [17]), but also to predict depression [14] or schizophrenia in first-episode of psychosis [18, 19, 20] on a linguistic basis.

2. Data and features

2.1. Corpus

The corpus is made up of a series of 198 Self-Defining Memories (SDMs, cf. section 2.3) collected in two samples, one consisting of adolescents with a diagnose of ADHD and the other with control participants (cf. section 2.2), who all had to write a total of three SDMs each. All texts are written in French.

SDMs were handwritten by the participants and then transcribed by the person in charge of the experiment. If spelling errors can provide valuable linguistic information, we noticed that some errors were introduced or removed during computer entry of the text. We have therefore decided to correct all of them: a first time with a professional spell checker, a second time with the *Hunspell* package [21].

2.2. Participants

ADHD group Adolescents with ADHD were recruited as part of a research project conducted at the Unit of Developmental Clinical Psychology at the University of Geneva. The project was advertised in local parents' associations for children with ADHD and through collaborations established with local child psychiatrists. Diagnostic criteria were investigated by detailed anamnestic interviews and confirmed by a semi-structured parents interview using the ADHD Child Evaluation cf. [22]. All diagnostic assessments were conducted by experienced clinical psychologists specialised in ADHD.

Control group The control sample was recruited from the general population in Geneva by means of advertisements and personal contact. Participants were volunteers, native French or fluent French speakers and received a compensation for their participation.

Inclusion/Exclusion criteria Inclusion criteria were age (12-17), fluency in French and, for the ADHD group, meeting current diagnostic criteria for ADHD [23]. Exclusion criteria were history of psychotic disorders, borderline personality disorder, autism spectrum disorder or neurological disorders.

2.3. Self-defining memories

SDMs are written texts that represent a specific type of autobiographical memories associated with the self-concept and a sense of coherence and continuity in one's ongoing individual history [24]. They were collected with the SDM Task [25, 26], during which participants are asked to evoke personal memories of events with specific attributes. These events

1. occurred at least one year ago;
2. are important and generally vividly represented;
3. are meaningful and useful to help the participant (or a significant other) to understand who s.he is;
4. are related to an important and enduring theme;

5. are either positive or negative but must generate strong feelings;
6. were recalled many times on a voluntary basis or spontaneously.

While listening to this description, participants had a sheet of paper in front of them that summed up these principal points. Participants were then told to imagine a situation where they meet someone they like very much and with whom they agree during a walk to talk about who they really are, their "Real Me", sharing several personal past events that powerfully convey how they have become the person they currently are. Participants were then given three sheets of paper on which they had to write down a SDM on each of them.

3. Prediction task

As already said, previous studies have shown that persons with ADHD have a different linguistic style from non-ADHD groups [27, 7, 6]. It should therefore be possible to classify texts from the ADHD and control groups with stylometric methods.

3.1. Hypothesis

Function words (FW) are words used both in stylometry [28] and in psychology [3]. These words have little lexical meaning and a grammatical role in the sentence (e.g. articles, prepositions, conjunctions, auxiliary verbs…). As such, FW are opposed to content words, which have a semantic content (e.g. nouns, adjectives, verbs…), but there is no clear delimitation between the two groups. Indeed, some words can be classified in one or the other category such as pronouns. Pronouns have indeed attracted much attention from researchers, who have emphasised their particular status first in literature [29, 30] and then in psychology [31].

According to Chung and Pennebaker [3], examining the use of FW in natural language samples has provided a non-reactive way to explore personality processes. For example, it has been found that use of specific FW is related with affective states [32, 3], depression [33, 34], reactions to individual life stressors [35], reactions to socially-shared stressors [36, 37, 38, 39], deception [40, 41], status [42], sex [43] and age [44]. We think that the use of FW differs between our two working groups (ADHD vs control) and that it is therefore possible to automatically classify one and the other group on the basis of FW. For years now, stylometry has been used to classify all types of documents [45], and seems to be the most suitable method for this task.

In addition to FW, we also propose to use another traditional feature of stylometric research: characters 3-grams [46], which can capture lexical, and even grammatical preferences. Such a feature has already shown its relevance for French texts in previous studies on authorship attribution [47], and has also proven its capacity to capture more than the authorial signal [48].

3.2. Set up

Support Vector Machine (SVM) Often used in stylometry, unsupervised approaches [49, 50, 51, 47, 52] seemed less adapted than supervised techniques for this profiling task. A recent survey [53] has shown that classical machine learning methods still perform better for profiling in similar settings (short texts, boolean or few categories) than deep learning. We turned to

Support Vector Machines (SVMs) rather than random forest [54] or logistic regression [55, 56], as it allows for easy interpretation and have established themselves as a standard method in stylometry [57, 58].

SuperStyl In this study, all analyses were implemented with the Python *SuperStyl* package [59]. This package has been used to build stylistic profiles with very good results [60]. *SuperStyl* use internally the SVM and pipeline facilities from scikit-learn [61].

Data All participants' SDMs are collected in a single file, with the exception of those written by two people from the control group and two from the ADHD group for a final blind test. Several word sample sizes are tested (1'000, 1'250, 1'500 and 2'000).

Parameters We tried to use two types of features: FW and character 3-grams for the reasons previously exposed. Because the length of SDMs varies a lot and we have more SDMs in the control group than in the ADHD groups (cf. section 4.3), we have tested different sampling methods (no sampling, downsampling, upsampling, Tomek links) and the use of class weights. All tests are conducted with a linear kernel and a 10-fold validation, on data normalised by using z-scores for variables and applied Euclidean vector-length normalisation (L2 normalisation).

3.3. Results

Scores Best results are achieved with 1'500 words samples, 10-fold validation, class weights and Tomek Links (for FW) / downsampling (for 3-grams). Accuracy is slightly lower with FW (0.85) than with 3-grams (0.92), which appears here as a promising indicator for research in psycholinguistics. However, the results remain surprisingly good in both cases. The recall for the ADHD group deserves special attention: for obvious reasons, a maximum number of people with ADHD must be identified and a minimum must be misclassified. With 0.75, results are not satisfactory yet, but promising.

	Precision	Recall	f1-score	support
control	0.89	0.89	0.89	9
TDAH	0.75	0.75	0.75	4
Accuracy			0.85	13

	Precision	Recall	f1-score	support
control	0.90	1.00	0.95	9
TDAH	1.00	0.75	0.86	4
Accuracy			0.92	13

Table 1
Results of the experiment for FW (left) and character 3-grams (right).

Test The SDMs of four participants (2 ADHD and 2 control) were not used for training and kept for a final test. The model perfectly classify ADHD and non-ADHD SDMs.

4. Lexical exploration

In a medical context, predicting is not enough: it is necessary to explain the results. The study of our corpus could make it possible to give initial explanations as to the quality of the prediction,

but also to confirm several hypotheses.

4.1. Hypotheses

Two standard measures have been used to explore the lexical information in the corpus:

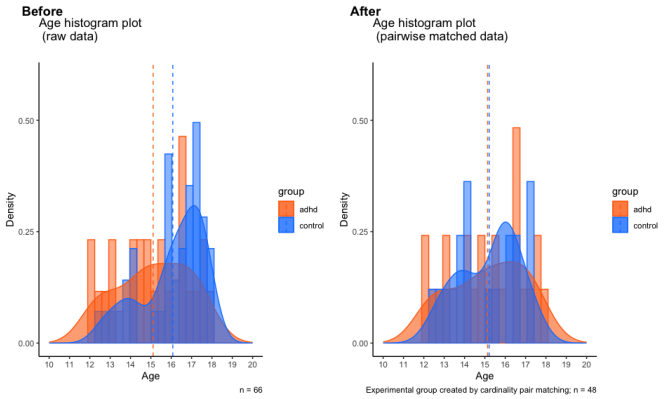
1. **Lexical Density** measures the structure and complexity of a text. It provides a measure of the proportion of lexical items (types) in relation to the total number of tokens in the text [62], and thus helps evaluating the amount of information in a given document [63]. Diversity evolves during human lifespan and it is influenced by different factors like education or the communication style of the family. Other studies in the field of psychology have found a relationship between neurodevelopmental disorders and lexical density, where children with autism spectrum disorders tend to have a lower density than the average population [27]. To the best to our knowledge, no study has yet explored lexical density in the linguistic production of adolescent with ADHD. Considering the fact that ADHD is a neurodeveloppmental disorder, measures of lexical density could differentiate adolescents with ADHD from those with a typical development.
2. **Lexical Diversity** measures how many types are used in each adolescent's narrative. SDMs that are lexically diverse use a wide range of vocabulary, of synonyms and a precise language to describe their memories. Previous studies have shown that measures of lexical diversity do not differentiate adolescents with ADHD from those with a typical development [64], we should therefore expect a similar result.

4.2. Set up

Pairwise matching is used to create samples balancing the ADHD and the control groups with respect to the means of participants' sex[1] and age in order to make them comparable.

Type	Raw	Pairwised
Boys	32	24
Girls	34	24
<=15 years old	21	20
>15 years old	45	28
ADHD	25	24
Control	41	24
Total	66	48

Table 2
Effect of pairwise matching on the distribution of participants (× 3 SDMs).



[1]We do not use gender, but biological sex as a category, as many other studies do. Our objective here is not to study the difference between these two groups (male vs. female) but to obtain two samples as comparable as possible (ADHD vs. control). The question of gender could, however, be of interest in future studies, especially because ADHD is possibly because ADHD is generally more likely to be diagnosed and treated for boys.

We matched the two samples using the cardinality matching method to find the largest matched set (in this case by age and sex) with the additional constraint that the ratio between the number of adolescents in both groups had to be equal to 1. This method allowed us to avoid differences between groups by sex or age with minimal loss of ADHD cases selecting the best-fitting control cases in function of the ADHD group.

Final samples In both groups the final sample meeting and pairwise inclusion criteria consisted of 24 adolescents (12 females and 12 males) in both groups (cf. tab. 2). The age mean weight in ADHD group was 15.14 (σ = 1.83) and 15.21 (σ = 1.44) for the control group. A Student two-samples t-test showed that the difference was statistically not significant (t(43.65) = -0.16, p = 0.87). Finally, as a result of this pairwise matching, the total number of SDMs per group has been reduced to 72 SDMs (24 participants $\times$ 3 SDMs) per group.

4.3. Results

Lexical diversity (TTR = $\frac{V}{N}$) is higher for the control group (209, σ = 63,9), indicating its members use significantly more different words to describe their memories than the ADHD group (153, σ = 54,5, p < 0.001), contradicting the results of Redmond [64]. This fact could be explained by the significant difference in size between the SDMs of the two groups (cf. fig. 1), those of the control group being generally longer ($\bar{x}$ >100 token) than those of the ADHD group ($\bar{x}\approx$60): as the length of the text increases, there is a greater statistical likelihood of finding a new word.

On the contrary lexical density (L_d = $\frac{N_{lex}}{N} * 100$) is not significantly different between the control group (0.027, σ = 0.008) and the ADHD group (0.029, σ = 0.01), which invalidates our hypothesis based on Yoder [27].

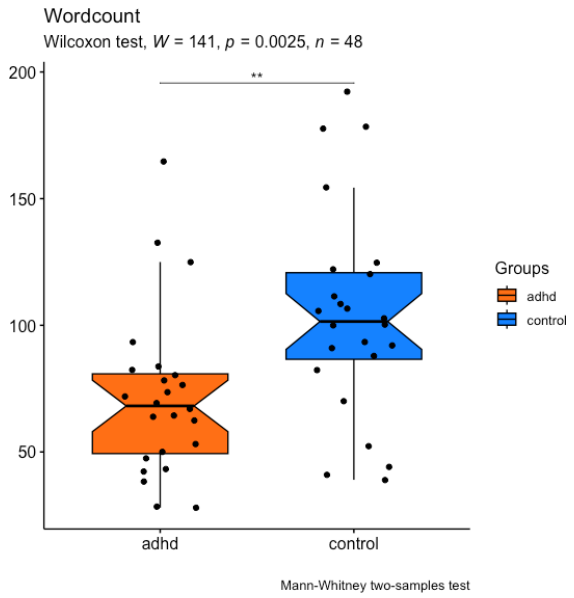


Figure 1: Number of tokens per SDM in ADHD group vs control group

4.4. Functors as markers

Analyzing which features allowed our SVM to distinguish between our two groups can help us understand further the differences between ADHD-diagnosed respondents and individuals from the control group. The size of our corpus does not allow us to comment on content words, not frequent enough to be considered as reliably interpretable. But we can gain significant insights from function words (cf. fig. 2).

For the ADHD group the markers include the neutral pronoun (*on*) combined with the third person auxiliaries and the abundance of words with syntactic function (*donc*, *et*, *avec*).

Marker	Context of meaning
"donc"	(...) nous avons **donc** été confiné ensemble (...)"
"et"	(...) **on** a pris de l'extasie **et** de la cocaïne **et on** s'est baladé toute la nuit (...)
"sur"	(...) **On** s'était couché surla plage **et** d'un coup, un nuage de libellules est arrivé (...)
"avec"	(...) C'était en 2020, en France, à la campagne **avec** toute ma famille (...)

Table 3
Examples of markers in their context of meaning (ADHD group).

On the other hand, the control group is very marked by first person pronouns (*je*, *me*/*m*) and the plural (*des*, *plus*).

Marker	Context of meaning
"je"	(...) **j**'avais peur de faire des toboggans quand un ami **m**'a forcé à le descendre."
"des"	(...) C'était compliqué de se concentrer à cause **des** distractions de la maison (...)
"plus"	(...) **j**'ai rencontré ma copine et ça **m**'a rendu **plus** heureuse et sûre de moi (...)
"fois"	(...) C'était la première fois que **j**'assistais à un enterrement. (...)

Table 4
Examples of markers in their context of meaning (Control group).

5. Discussion

Our findings indicate that adolescent's with ADHD diagnosis show significant differences in their autobiographical narratives style of language from their control group (non-ADHD cohort), which is enough to have them detected by the machine. Additionally, different markers for each group were found which are meaningful from a psychological perspective. Indeed, the use of the neutral pronoun in the ADHD group is very different from the use of the first singular pronoun in the case of the Control group. In terms of agency, the latter identifies the author as the subject performing and feeling things while the narrated event occurred. In the case of the ADHD group, the agent is subsumed in an undefined neutral mass (*on*) that performs things in a distinctly less personal way. In their model of the relationships between self, memory, and visual perspective, Sutin and Robins [65] argued that in the first person perspective, individuals see the event through their own eyes, while, in the third person perspective, individuals see themselves and the event from the perspective of an external observer. According to their model, a reduced use of the first person may serves a distancing function helping to reduce emotional reliving and to distance the current self from the self in the memory. In this sense, adolescents with ADHD show at least difficulties to connect with their emotional experiences, either by way of a defence mechanisms or by functional difficulties related to executive functions. This result is even more relevant to consider if one takes into account that the task asks for memories lived by the person that have marked him/her and that, therefore, easily elicit first-person experiences.

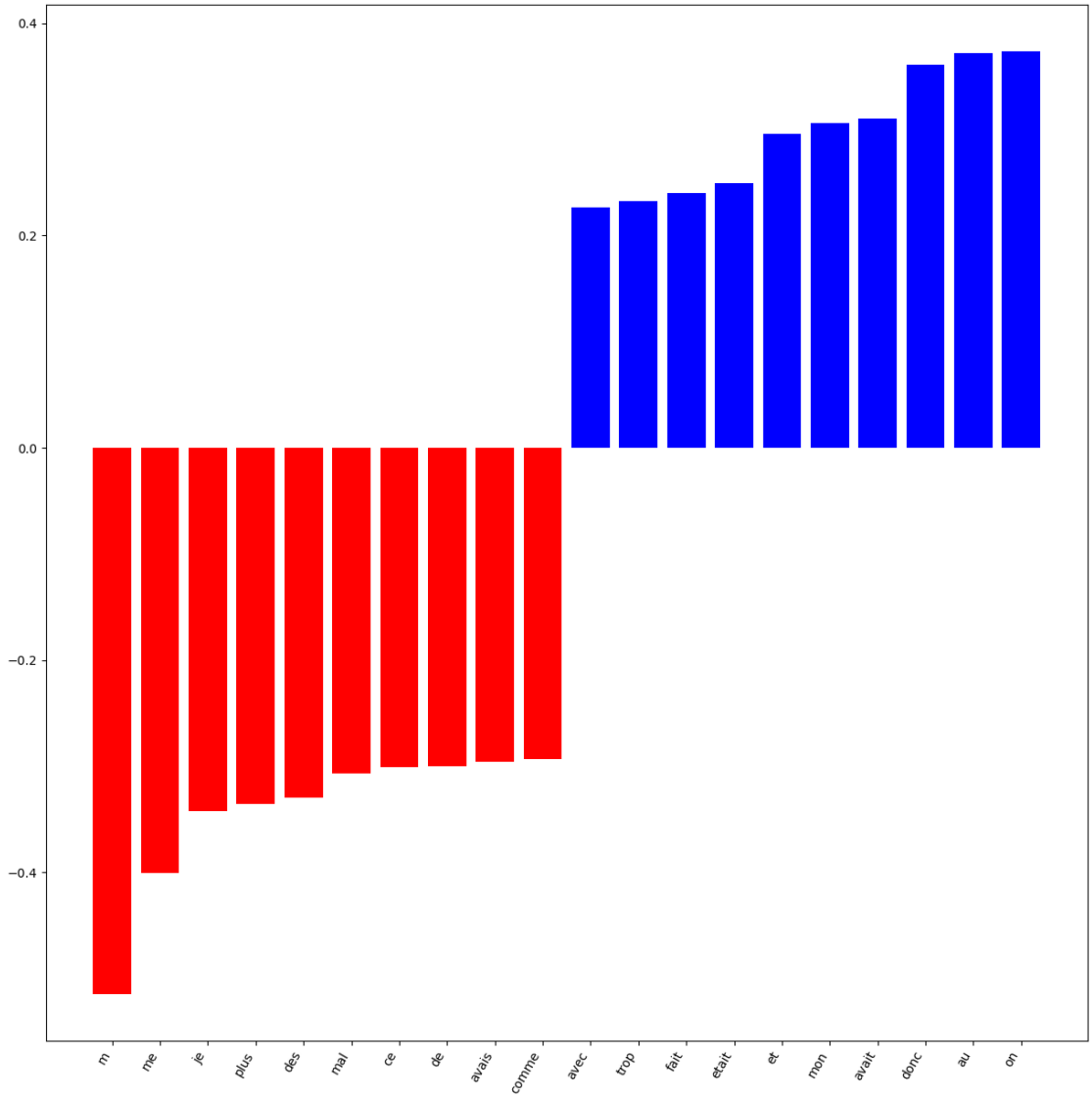


Figure 2: Most important FW for stylometric classification. In red the control group, in blue the ADHD group.

With respect to lexical density, our results are not consistent with previous finding and does not confirm differences between groups based on the presence or absence of ADHD. However, this could be due to the small size of the ADHD group narratives.

6. Further work

Future studies are needed to increase the accuracy and fine grain detection of ADHD. First the potential differences by sex or gender in language markers should be addressed. Second,

refining the precision of the detection in order to identify the different modalities of ADHD (predominantly inattentive, hyperactive or combined) is another relevant challenge to consider for future studies. Third, self-reported affect reported by participants in the Task of SDMs should be considered in order to evaluate to what extend this metadata could improve the detection of the signal of ADHD in the adolescent's autobiographical narratives. Finally, considering the potential challenge for young adolescents to write a memoir, another interesting aspect would be to modify the task to make it a verbal task and, eventually, also modify the instruction to make it simpler and eventually more stimulating for the target population.

Acknowledgements

JB was funded by the Chilean National Agency for Research and Development (ANID Chile) through the PhD Abroad Scholarship, 2018 award. The PI (Prof. Martin Debbané) was funded by the Swiss National Science foundation (Grant number 100014_179033), as well as the Marina Picasso Prize from AEMD Foundation 2018.

References

[1] N. Salari, H. Ghasemi, N. Abdoli, A. Rahmani, M. H. Shiri, A. H. Hashemian, H. Akbari, M. Mohammadi, The global prevalence of ADHD in children and adolescents: a systematic review and meta-analysis, Italian Journal of Pediatrics 49 (2023-04-20) 48. doi:10.1186/s13052-023-01456-1.

[2] S. V. Faraone, P. Asherson, T. Banaschewski, J. Biederman, J. K. Buitelaar, J. A. Ramos-Quiroga, L. A. Rohde, E. J. S. Sonuga-Barke, R. Tannock, B. Franke, Attention-deficit/hyperactivity disorder, Nature Reviews. Disease Primers 1 (2015) 15020. doi:10.1038/nrdp.2015.20.

[3] C. Chung, J. W. Pennebaker, The psychological functions of function words, Social communication (2007).

[4] J. Jogia, A. H. Sharif, F. A. Nawaz, A. R. Khan, R. H. Alawami, M. A. Aljanahi, M. A. Sultan, Comorbidities associated with attention-deficit/hyperactivity disorder in children and adolescents at a tertiary care setting, Global Pediatric Health 9 (2022-02-20) 2333794X221076607. doi:10.1177/2333794X221076607.

[5] A. Leis, F. Ronzano, M. M. A., L. I. Furlong, F. Sanz, Detecting signs of depression in tweets in spanish: Behavioral and linguistic analysis, Journal of Medical Internet Research 21 (2019) e14199. doi:10.2196/14199.

[6] K. Kim, S. Lee, C. Lee, College students with ADHD traits and their language styles, Journal of Attention Disorders 19 (2015) 687–693. doi:10.1177/1087054713484512.

[7] K. Kim, C. H. Lee, Distinctive linguistic styles in children with ADHD, Psychological Reports 105 (2009) 365–371. doi:10.2466/PR0.105.2.365-371.

[8] C. L. Bagwell, B. S. Molina, W. E. Pelham, B. Hoza, Attention-deficit hyperactivity disorder and problems in peer relations: Predictions from childhood to adolescence, Journal of the American Academy of Child & Adolescent Psychiatry 40 (2001) 1285–1292. doi:10.1097/00004583-200111000-00008.

[9] R. A. Barkley, K. E. Fletcher, Adolescents with ADHD: Patterns of behavioral adjustment, academic functioning, and treatment utilization, Journal of the American Academy of Child & Adolescent Psychiatry (1991). URL: 10.1016/s0890-8567(10)80010-3.

[10] J. A. Simmons, K. M. Antshel, Bullying and depression in youth with ADHD: A systematic review, Child & Youth Care Forum 50 (2021) 379–414. doi:10.1007/s10566-020-09586-x.

[11] C. Crema, G. Attardi, D. Sartiano, A. Redolfi, Natural language processing in clinical neuroscience and psychiatry: A review, Frontiers in Psychiatry 13 (2022). doi:10.3389/fpsyt.2022.946387.

[12] T. Zhang, A. M. Schoene, S. Ji, S. Ananiadou, Natural language processing applied to mental illness detection: a narrative review, npj Digital Medicine 5 (2022-04-08) 1–13. doi:10.1038/s41746-022-00589-7.

[13] J. Noecker Jr, M. Ryan, P. Juola, Psychological profiling through textual analysis, Literary and Linguistic Computing 28 (2013) 382–387. doi:10.1093/llc/fqs070.

[14] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, H. A. Schwartz, Facebook language predicts depression in medical records, Proceedings of the National Academy of Sciences of the United States of America 115 (2018) 11203–11208. doi:10.1073/pnas.1802331115.

[15] A. Pérez, J. Parapar, A. Barreiro, Automatic depression score estimation with word embedding models, Artificial Intelligence in Medicine 132 (2022) 102380. doi:10.1016/j.artmed.2022.102380.

[16] M. Manabe, K. Liew, S. Yada, S. Wakamiya, E. Aramaki, Estimation of psychological distress in japanese youth through narrative writing: Text-based stylometric and sentiment analyses, JMIR Formative Research 5 (2021) e29500. doi:10.2196/29500.

[17] P. Tchounwou, Environmental research and public health, International Journal of Environmental Research and Public Health 1 (2004) 1–2. doi:10.3390/ijerph2004010001.

[18] A. Figueroa-Barra, D. Del Aguila, M. Cerda, P. A. Gaspar, L. D. Terissi, M. Durán, C. Valderrama, Automatic language analysis identifies and predicts schizophrenia in first-episode of psychosis, Schizophrenia 8 (2022) 53. doi:10.1038/s41537-022-00259-3.

[19] S. X. Tang, R. Kriz, S. Cho, S. J. Park, J. Harowitz, R. E. Gur, M. T. Bhati, D. H. Wolf, J. Sedoc, M. Y. Liberman, Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders, npj Schizophrenia 7 (2021) 25. doi:10.1038/s41537-021-00154-3.

[20] N. Rezaii, E. Walker, P. Wolff, A machine learning approach to predicting psychosis using semantic density and latent content analysis, npj Schizophrenia 5 (2019) 9. doi:10.1038/s41537-019-0077-9.

[21] J. Ooms, hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker, 2023. URL: https://docs.ropensci.org/hunspell, R package version 3.0.3.

[22] S. Young, ADHD Child Evaluation (ACE), A diagnostic interview of ADHD in children, Psychology Services Limited, London, 2015. URL: https://www.psychology-services.uk.com/adhd.

[23] A. P. Association, Diagnostic And Statistical Manual Of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR), American Psychiatric Association, Washington D.C., 2013. doi:10.1176/appi.books.9780890425787.

[24] P. S. Blagov, J. A. Singer, K. M. Oost, J. A. Goodman, Self-defining memories-narrative features in relation to adaptive and maladaptive personality traits (replication and extension of blagov & singer, 2004), Journal of Personality 90 (2022) 457–475. doi:10.1111/jopy.12677.

[25] A. Thorne, K. C. McLean, Manual for Coding Events in Self-Defining Memories, University of California, Santa Cruz, Santa Cruz, 2001. URL: http://www.self-definingmemories.com/Thorne___McLean_SDM_Scoring_Manual.pdf, unpublished manuscript.

[26] J. A. Singer, P. S. Blagov, Classification System & Scoring Manual for Self-Defining Memories, Connecticut College, New London, CT, 2022. URL: http://www.self-definingmemories.com/Classification_System___Scoring_Manual_for_SDMs.pdf.

[27] P. J. Yoder, Predicting lexical density growth rate in young children with autism spectrum disorders, American Journal of Speech-Language Pathology 15 (2006) 378–388. doi:10.1044/1058-0360(2006/035).

[28] M. Kestemont, Function Words in Authorship Attribution: From Black Magic to Theory?, in: Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), Gothenburg, Sweden, 2014, pp. 59–66. doi:10.3115/v1/W14-0908.

[29] C. Muller, Sur quelques scènes de Molière, essai d'un indice du style familier, in: Langue française et linguistiques quantitatives, Slatkine, 1979, pp. 107–124.

[30] M. Kastberg Sjöblom, L'indice pronominal est-il encore d'actualité ?, Lexicometrica 5 (2004). URL: http://lexicometrica.univ-paris3.fr/article/numero5.htm.

[31] J. W. Pennebaker, The Secret Life of Pronouns: What Our Words Say About Us, Bloomsbury Press, New York, 2013.

[32] Weintraub, Verbal Behavior in Everyday Life, Springer, New York, 1989.

[33] S. Rude, E.-M. Gortner, J. W. Pennebaker, Language use of depressed and depression-vulnerable college students, Cognition and Emotion 18 (2004) 1121–1133. doi:10.1080/02699930441000030.

[34] M. R. Mehl, The lay assessment of subclinical depression in daily life, Psychological Assessment 18 (2006) 340–345. doi:10.1037/1040-3590.18.3.340.

[35] J. W. Pennebaker, T. C. Lay, Language use and personality during crises: Analyses of mayor rudolph giuliani's press conferences, Journal of Research in Personality 36 (2002) 271–282. doi:10.1006/jrpe.2002.2349.

[36] R. S. Campbell, J. W. Pennebaker, The secret life of pronouns: Flexibility in writing style and physical health, Psychological Science 14 (2003) 60–65. doi:10.1111/1467-9280.01419.

[37] L. D. Stone, J. W. Pennebaker, Trauma in real time: Talking and avoiding online conversations about the death of princess diana, Basic and Applied Social Psychology 24 (2002) 173–183. doi:10.1207/S15324834BASP2403_1.

[38] E.-M. Gortner, J. W. Pennebaker, The archival anatomy of a disaster: Media coverage and community-wide health effects of the texas A&M bonfire tragedy, Journal of Social and Clinical Psychology 22 (2003) 580–603. doi:10.1521/jscp.22.5.580.22923.

[39] M. A. Cohn, M. R. Mehl, J. W. Pennebaker, Linguistic markers of psychological change surrounding september 11, 2001, Psychological Science 15 (2004) 687–693. doi:10.1111/j.0956-7976.2004.00741.x.

[40] J. W. Pennebaker, L. A. King, Linguistic styles: Language use as an individual difference, Journal of Personality and Social Psychology 77 (1999) 1296–1312. doi:10.1037/

0022-3514.77.6.1296.
[41] M. L. Newman, Pennebaker, D. S. James W., Berry, J. M. Richards, Lying words: Predicting deception from linguistic styles, Personality and Social Psychology Bulletin 29 (2003) 665–675. doi:10.1177/0146167203029005010.
[42] J. W. Pennebaker, D. M., Pronoun use and dominance, Department of Psychology, University of Texas at Austin, Austin, TX, 2006.
[43] M. L. Newman, D. S. Pennebaker, James W.and Berry, J. M. Richards, Lying words: Predicting deception from linguistic styles, Personality and Social Psychology Bulletin 29 (2003) 665–675. doi:10.1177/0146167203029005010.
[44] J. W. Pennebaker, L. D. Stone, Words of wisdom: Language use over the life span, Journal of Personality and Social Psychology 85 (2007) 291–301. doi:10.1037/0022-3514.85.2.291.
[45] E. Segev (Ed.), Affaires de style : du cas Molière à l'affaire Grégory, la stylométrie mène l'enquête, Le Robert, Paris, 2022.
[46] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, Journal of the American Society for Information Science and Technology 60 (2009) 9–26. doi:10.1002/asi.20961.
[47] F. Cafiero, J.-B. Camps, Why Molière most likely did write his plays, Science Advances 5 (2019) eaax5489. doi:10.1126/sciadv.aax5489.
[48] S. Gabay, Beyond idiolectometry? on racine's stylometric signature, in: Conference on Computational Humanities Research 2021, 2021, pp. 359–376. URL: https://hal.science/hal-03402994.
[49] J.-B. Camps, F. Cafiero, Setting bounds in a homogeneous corpus: a methodological study applied to medieval literature, Revue des Nouvelles Technologies de l'Information (2013) 55–84. URL: https://shs.hal.science/halshs-00765651.
[50] H. Gómez-Adorno, C. Martín-del Campo-Rodríguez, G. Sidorov, Y. Alemán, D. Vilariño, D. Pinto, Hierarchical clustering analysis: the best-performing approach at pan 2017 author clustering task, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9, Springer, 2018, pp. 216–223. URL: 10.1007/978-3-319-98932-7_20.
[51] R. McCarthy, J. O'Sullivan, Who wrote wuthering heights?, Digital Scholarship in the Humanities 36 (2021) 383–391. doi:10.1093/llc/fqaa031.
[52] C. Schmidt-Petri, M. Schefczyk, L. Osburg, Who authored on liberty? stylometric evidence on harriet taylor mill's contribution, Utilitas 34 (2022) 120–138.
[53] Y. HaCohen-Kerner, Survey on profiling age and gender of text authors, Expert Systems with Applications 199 (2022) 117140.
[54] C. Ikae, Unine at pan-clef 2022: Profiling irony and stereotype spreaders on twitter, in: CLEF, 2022, pp. 1613–0073. doi:https://pan.webis.de/clef22/pan22-web/author-profiling.html.
[55] P. Modaresi, M. Liebeck, S. Conrad, Exploring the effects of cross-genre machine learning for author profiling in pan 2016, Components of an Automatic Single Document Summarization System in the News Domain (2017) 100–107. URL: https://ceur-ws.org/Vol-1609/16090970.pdf.

[56] J. K. Ward, F. Cafiero, R. Fretigny, J. Colgrove, V. Seror, France's citizen consultation on vaccination and the challenges of participatory democracy in health, Social Science & Medicine 220 (2019) 73–80. doi:10.1016/j.socscimed.2018.10.032.

[57] M. Eder, Rolling stylometry, Digital Scholarship in the Humanities 31 (2015) 457–469. doi:10.1093/llc/fqv010.

[58] F. Cafiero, J.-B. Camps, 'Psyché' as a Rosetta stone? assessing collaborative authorship in the french 17th century theatre, in: Proceedings of the Conference on Computational Humanities Research 2021, CEUR Workshop Proceedings, 2021, pp. 377–391. URL: https://ceur-ws.org/Vol-2989/long_paper51.pdf.

[59] J.-B. Camps, SUPERvised STYLometry (SuperStyl), 2021. URL: https://github.com/SupervisedStylometry/SuperStyl.

[60] F. Cafiero, J.-B. Camps, Who could be behind QAnon? authorship attribution with supervised machine-learning, Digital Scholarship in the Humanities (2023). doi:10.48550/arXiv.2303.02078.

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: https://jmlr.org/papers/v12/pedregosa11a.html.

[62] V. Johansson, Lexical diversity and lexical density in speech and writing, Working papers / Lund University, Department of Linguistics and Phonetics 53 (2008). URL: https://journals.lub.lu.se/LWPL/article/view/2273.

[63] M. A. K. Halliday, Spoken and written language, Oxford University Press [Oxford], 1990.

[64] S. M. Redmond, Conversational profiles of children with ADHD, SLI and typical development, Clinical Linguistics & Phonetics 18 (2004) 107–125. doi:10.1080/02699200310001611612.

[65] A. R. Sutin, R. W. Robins, When the "I" looks at the "Me": Autobiographical memory, visual perspective, and the self, Consciousness and Cognition 17 (2008) 1386–1397. doi:10.1016/j.concog.2008.09.001.