

Female Autism in Natural Language – A Corpus Paper

Nadine Probol^{1,2}, Margot Mieskes¹

¹University of Applied Sciences Darmstadt

²Otto-Friedrich University of Bamberg

nadine.probol@h-da.de, margot.mieskes@h-da.de

Abstract

Autism Spectrum Disorder (ASD) is a condition, which gets diagnosed more frequently in recent years. This increase in people on the autism spectrum raises the need for more and also more efficient indicators for the developmental disorder. The aim of this work is to show how it is possible to create a reproducible data set of speech and transcribed data of women and men on the spectrum. The data is collected from YouTube, Instagram, and TikTok with the explicit consent of the people donating their recordings. This includes problems due to ASD, tricks and tips to improve life and help to understand neurotypical (NT) reactions to their behaviour. This dataset should not be used in any way to replace a professional diagnosis, but rather to find linguistic indicators of ASD which might be of use to support a diagnosis.

1 Introduction

Closing gaps in research on severely under-represented groups often requires more creative approaches, as data tends to be sparse and hard to collect. One such gap is the study of females on the autism spectrum.

Autism Spectrum Disorder (ASD) is a condition, which gets diagnosed more frequently in recent years. Although the exact numbers vary according to the sources, they show the same trend. The Centers for Disease Control (CDC) report that 1 in 36 children is on the autism spectrum and the numbers are increasing.¹

This leads to two considerations: First, this increase in people on the spectrum raises the need for more but also more efficient indicators for the developmental disorder. Additionally, there is a huge difference in the diagnosis of women versus men. Although the CDC observes the likelihood of an autism diagnosis to be four times more common in boys and men than in girls and women,² this discrepancy has decreased in recent years. Breddemann et al. (2023) find a "strong gender bias", as women are diagnosed about seven to eleven years later than men. There is even a paragraph in

the ICD-11 on how girls and women are often not recognised, which might be the reason for this discrepancy between male and female patients.³ A reason for this discrepancy might be due to females on the spectrum displaying "fewer restricted, repetitive interests and behaviours" than their male counterparts.⁴

Second, this difference in the diagnosis of female and male patients requires more attention, in order to also give female patients adequate help early on.

In order to explore the reasons for this discrepancy further, we need datasets containing data for both male and female individuals on the spectrum. This is especially important, as most previous studies in this context do not provide enough female data (Probol and Mieskes 2024). The goal of this work is to create a reproducible data set of speech and transcribed data of women and men on the spectrum.

In this work, we show a way of collecting data from a potentially vulnerable and under-researched group (autistic individuals).

2 Related Work

Most research focuses on children on the spectrum (i.e. Parish-Morris et al. (2016); Ashwini, Narayan, and Shukla (2023)). When looking at the audio aspects of autistic speech, the age has already been identified as a large influence on the results of NLP research (Hauser et al. 2019). This might also apply to transcriptions and should therefore be avoided.

However, there are some studies focusing on adult individuals (i.e. Liu et al. (2022)). Their dataset includes 36 participants, overall, there are only 3091 utterances of autistic participants.

Another problem is the female to male ratio of the participants. While most research includes considerably more male participants (Parish-Morris et al. 2016), some do not give any information about the sex of the participants (Prud'hommeaux, van Santen, and Gliner 2017; Liu et al. 2022; Ashwini, Narayan, and Shukla 2023). For example, Ashwini, Narayan, and Shukla (2023) focus on adult indi-

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://www.cdc.gov/autism/data-research/>, accessed February 15, 2025

²<https://www.cdc.gov/autism/data-research/>, accessed February 15, 2025

³<https://icd.who.int/browse11/l-m/en/#/http://id.who.int/icd/entity/437815624>, accessed February 15, 2025

⁴<https://icd.who.int/browse11/l-m/en/#/http://id.who.int/icd/entity/437815624>, accessed February 15, 2025

Search Term	Platform
#asd	TikTok
#asperger	TikTok, Instagram
#actuallyautistic	TikTok, Instagram
#autism	TikTok, Instagram
#autismwomen	TikTok, Instagram
#autisticgirls	TikTok, Instagram
#AuDHD	TikTok, Instagram
#autistiktok	TikTok
#aspergirls	TikTok
#aspiegirl	TikTok, Instagram
#aspiegirl	TikTok, Instagram
autism (in) women	YouTube
autism (in) girls	YouTube
autistic women	YouTube

Table 1: Search terms used for data acquisition

viduals, but fail to give any information on female to male ratio of the participants.

But mostly datasets are not publicly available due to data privacy reasons. Therefore, results are harder to reproduce and new ideas cannot easily be tested.

Often, the data has been collected in diagnostic interviews (i.e. Parish-Morris et al. (2016)). This might influence the data, as this can be a stressful situation for the participants and might not represent their normal speech.

When looking at studies focused on autism, most have much more male participants than females (Probol and Mieskes 2024), reducing the generalizability of the results to females.

3 Data Collecting

This dataset is intended for researchers who want to delve deeper into the linguistic differences between males and females on the autism spectrum.⁵

The data was collected from YouTube, Instagram, and TikTok. Since the videos we collected were created over several years, depending on the specific influencer, the dataset contains diachronic data. To cover as many people as possible, we focus on native English speakers. This also reduces the influence of different languages. The data contains only audio, however, we automatically generated transcriptions with additional information, as described in Section 4.1.

In order to find suitable accounts on social media platforms, we collected hashtags and search terms (see Table 1). While there are specific terms for females on the spectrum there are no common ones for males on the spectrum as they tend to use the more general hashtags, such as #actuallyautistic or #AuDHD. Therefore, while this list seems to lean on one side it actually includes both sides.

To get even more accounts, we followed the recommendations after clicking on the videos found via the mentioned hashtag. We also looked through recommendations made by autistic creators in their videos or linked on their accounts. Some creators use linktrees to make their account available

⁵Our study has the approval of an ethics board, details of which will be published in the accepted version.

on several social media platforms which helps collect data of different lengths from the same creator.⁶ This led to a list of 121 female and 23 male social media users on the spectrum. After contacting all of them and asking for their permission to use their data for research purposes, only few answered and not all agreed. Our final list consists of 18 social media users who agreed to let us use their data.

3.1 Audio data

Our data set contains data from 14 female and 4 male individuals. As shown in Table 2, we collected 2,641 videos from female individuals on the spectrum and 765 videos recorded by males on the spectrum (see Table 3).

This led to more than 282 hours of female data and more than 107 hours of male data. Although there is more female data overall, the average hours per person are higher for males (27 hours per individual) than for females (20 hours per individual).

This can also be seen in a much shorter median length of 1 : 19 *min* for the videos of the female creators in comparison to a median length of 4 : 58 *min* for the male creators.

While average length of the videos of the female creators is also shorter (6 : 25 *min*) than for the male creators (8 : 26 *min*), the difference is not as large as the median length.

The shortest video is 5 *sec* for both male and female data, however, the longest video is much longer for female creators (2 : 11 : 49 *hours*) than for male creators (1 : 32 : 30 *hours*). Please note that although the shortest videos are only five seconds long, they still contain spoken language.

It is also important to note that the longest videos are re-uploads of livestreams and not scripted videos. Those video might therefore give an even more natural insight into autistic speech.

3.2 Reproducibility

In order to allow reproduction of the described corpus, a list with all used links of the videos can be found on our GitHubRepository.⁷

4 Wordings and semantics

In general, the wordings of female and male individuals, regardless of being on the spectrum or not, differ. Mulac, Lundell, and Bradac (1986) found the usage of first person singular pronouns, judgmental adjectives and references to people to be highly used by male individuals.⁸ Also, the authors found males to use a relatively high amount of syllables per word and vocalize their pauses with "ah" or "uhm". Females, on the other hand, tend to use rhetorical questions, references to emotion, and fillers.⁹ Also, they use a relatively

⁶Some creators make very short videos for platforms like TikTok, however, they also post longer and more elaborate videos on YouTube. This is important as the length of the video might have an impact on the wording.

⁷<https://github.com/Nadine271/female-autism>

⁸The list by Mulac, Lundell, and Bradac (1986) consists of ten indicators for each sex.

⁹Examples given were "okay" or "you know".

#	# videos	overall length	average length	median length	min length	max length	Main platform
1	107	23:33:42	00:13:13	00:12:02	00:02:06	00:40:31	YouTube
2	169	20:53:18	00:07:45	00:02:40	00:00:13	00:55:15	YouTube
3	335	77:06:00	00:13:49	00:12:22	00:02:30	01:51:14	YouTube
4	73	30:34:57	00:25:08	00:24:23	00:10:01	00:55:26	YouTube
5	180	71:42:54	00:23:54	00:14:52	00:03:16	02:11:49	YouTube
6	335	03:33:33	00:00:38	00:00:36	00:00:06	00:04:04	TikTok
7	72	00:54:59	00:00:46	00:00:39	00:00:08	00:04:34	TikTok*
8	278	07:50:37	00:01:42	00:01:17	00:00:06	00:09:52	TikTok
9	228	03:29:53	00:00:55	00:00:58	00:00:05	00:01:29	Instagram
10	362	06:57:07	00:01:09	00:01:07	00:00:06	00:03:51	Instagram
11	263	04:16:26	00:00:59	00:01:02	00:00:12	00:02:02	TikTok
12	71	08:38:02	00:07:18	00:03:49	00:00:15	00:26:30	YouTube*
13	54	10:03:35	00:11:11	00:10:04	00:00:25	00:43:20	YouTube
14	114	11:52:48	00:06:15	00:05:36	00:00:13	00:21:41	YouTube
	2,641	282:25:14	00:06:25	00:01:19	00:00:05	02:11:49	

Table 2: Composition of the data of female autistic individuals in the data set. * means the data was mainly collected from the listed platform but also includes data from one or both other platforms.

#	# videos	overall length	average length	median length	min length	max length	Main platform
1	320	33:06:16	00:06:12	00:05:30	00:01:00	00:51:43	YouTube
2	206	06:01:34	00:01:45	00:01:01	00:00:05	00:14:52	YouTube
3	59	07:28:55	00:07:37	00:03:40	00:00:24	00:49:46	YouTube
4	180	60:50:15	00:20:17	00:16:48	00:00:30	01:32:30	YouTube
	765	107:27:00	00:08:26	00:04:58	00:00:05	01:32:30	

Table 3: Composition of the data of male autistic individuals in the data set

high amount of negations, such as "You don't feel like looking..." (Mulac, Lundell, and Bradac 1986).

Looking at the wording of autistic individuals, it was found to be very formal and direct (Hosseini and Molla 2023). This goes as far as having no regards to avoid being offensive to others.

Begeer et al. (2014) examined differences in coming up with words in between neurotypical (NT) and ASD individuals. They concluded that while ASD individuals come up with larger clusters of words they have fewer topic switches. This means that ASD individuals can list more words from just one subcategory, whereas NT individuals tend to switch more often between subcategories. A possible reason for this could be special interests in some topics. Asperger (1944) described ASD individuals as having narrow and pedantic interests. This is particularly important when looking at the dataset created in this paper, as the topics that autistic people talk about online are likely to be a special interest. While stereotypical behaviour is mostly seen as an impairment, Begeer et al. (2014) hypothesize that it can be an asset. Using it to create content on the Internet could be such an asset.

Another aspect of ASD is a certain creativity in language use. Asperger (1944) observed patients using uncommon words that did not match the environment of the participant. This is in line with the findings of Wing (1981) and Ghaziuddin and Gerstein (1996), who also observed the tendency of individuals with ASD to use complicated and uncommon words.

Asperger (1944) also observed his patients to form completely new words or transforming already existing ones, if they felt that existing words were not suitable. However,

while these words can be fitting, there might be situations in which these newly formed words seem absurd.¹⁰ This is generally associated with higher structural language skills (Luyster, Zane, and Wisman Weil 2022).

Parish-Morris et al. (2016) compiled a list of words that are potential indicators of ASD. While they found words such as "mhm", "uh", or "eh" to be "ASD-like", the authors observed "um" to be an indicator against being on the spectrum. Generally, Parish-Morris et al. (2016) collected words without a lexical counterpart, e.g. imitative or expressive noises, to be "ASD-like", as well as words, that show stuttering-like disfluency.¹¹ More unassuming words they found to be indicators of ASD are "know", "well", or "actually".¹² Words indicating to be not on the spectrum are "like", "if", or "them".

4.1 Transcriptions

In order to get transcriptions as universally usable as possible, we use the time-stamped version¹³ of the open source Speech-To-Text tool *Whisper*¹⁴ from *OpenAI*. This gives us not only the complete transcription of the audio, but also segments and the single utterances, which are each marked with a timestamp at the start and the end. Additionally, we use a specific version of the Whisper-model, called Whisper-

¹⁰Examples can be found in Asperger (1944)

¹¹For example "uh" or "w-".

¹²This is just part of the list. The complete list can be found in Parish-Morris et al. (2016)

¹³<https://github.com/linto-ai/whisper-timestamped>

¹⁴<https://openai.com/research/whisper>

Stutter	Laugh
Pause	Cry
Repetition	Sigh
Blunder	Sing
Lisp	Tremors
Emphasize	Whisper
Clicking sounds	Background noises

Table 4: List of tags which were manually added.

SV model by SebLih, which is fine-tuned on the Common Voice dataset.¹⁵ We decided on this specific model, as it produces more correct transcriptions of the videos than the original Whisper-model.

We then further annotate the automatic transcriptions. The list of tags can be found in Table 4.

As Echolalia is a known symptom of ASD, the repetition of words is marked in the transcriptions. In order to track problems with certain words or possible connections between blunders and speaking rate, we add a tag for blundering. If words are emphasized or are prolonged more than normal, we tag them with "Emphasize". Often, the influencers were singing parts of their texts, which we tagged as well. Tagged background noises include bird singing and door clapping, which are not directly linked to the content of the audio. Background music is tagged separately.

We will use the transcriptions with manually added tags from Table 4 to train another Speech-To-Text tool for the rest of our data. As of this time, we are still working on the manually added tags.

4.2 Topics

The topics covered mostly focus on how to deal with the autism of the creator or other people. This includes problems due to ASD, tricks and tips to improve life and help to understand NT reactions to their behaviour. However, there are also videos of the "get ready with me" type, in which the creators talk about various things in their lives and/or of importance to them, while getting ready for their day. Additionally, some talk about hobbies like sports or are producing gaming content.

4.3 Qualitative Analysis

In our qualitative analysis we take a preliminary look at representative videos of the creators.

In many videos, we were adding a singing tag. Interestingly, the creators often start singing in the middle of a sentence or when saying their channel's name. Singing is normally not a separate part within in the videos but integrates into the rest of the text. This shows the importance of adding our tags mentioned in Table 4 not by sentence but by single words or utterances.

Other anomalies such as echolalia or general repetitions are not very frequent. This might be due to the nature of the videos. Most of the videos are pre-recorded and not livestreamed and therefore cut before being uploaded. However, tagging these anomalies is important if the dataset is

supposed to be extended at some point with more spontaneous speech such as livestreams.

5 Conclusion & Future Work

While our corpus contains fewer participants than most other datasets used for autism detection tasks in NLP, it focuses on data created by autistic individuals in a more natural environment. The participants are therefore much more relaxed than when using data collected in a diagnostic interview. Though the general number of participants is lower, our corpus contains much more data (hours) overall. Also, we collected a higher amount of female data than male data.

Contrary to the other studies and official statistics, during our search we observed that the number of female influencers on the spectrum is higher than the number of male influencers.

When collecting the data, getting the consent of the participants was a big problem. Even though we found a much larger number of influencers on the spectrum who would have been able to provide suitable data, many do not give an email address for contact. The majority of all contacted influencers did not even react in negative but ignored our request completely. Then, we had to cut out all possible influencers under the age of 18. This made it difficult to collect a large amount of data.

The next step in our research is to create another and more balanced dataset. The aim of that new dataset is, to reach a larger pool of participants. To get an even larger group of participants, we will set up a Prolific study. The participants will be asked to give some information on themselves, including sex, age, and diagnosis, give some monologues on certain topics. Additionally, the participants are going to be asked to speak specific sentences. Finally, the label preference of the participants is going to be asked.

Ethical Statement

Even though we look into identifiers for ASD in voice, speech and language, it is important to note, that we do not intend to say that these findings can be used to automatically classify the disorder. Our findings should therefore not be used in any way to replace a professional diagnosis, but rather the described indicators of ASD might be of use to support a diagnosis.

Limitations

Although we tried to include data from as many different backgrounds as possible, this survey is not able to include all existing cultural or ethnic groups. Therefore, it is not possible to generalise the findings of this paper to all individuals on the spectrum. Additionally, we rely on the influencers to truthfully state their diagnosis. We did not back check each individually with a diagnostician.

Acknowledgements

A huge thank you to all the social media creators who allowed us the usage of their data. We would also like to thank the reviewers for their helpful comments.

¹⁵<https://huggingface.co/SebLih/whisper-SV>

References

Ashwini, B.; Narayan, V.; and Shukla, J. 2023. SPASHT: Semantic and Pragmatic Speech Features for Automatic Assessment of Autism. In *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, Jun 4th – 10th 2023, 1–5. IEEE.

Asperger, H. 1944. Die Autistischen Psychopathen” im Kindesalter. *Archiv für Psychiatrie und Nervenkrankheiten*, 117(1): 76–136.

Begeer, S.; Wierda, M.; Scheeren, A. M.; Teunisse, J.-P.; Koot, H. M.; and Geurts, H. M. 2014. Verbal fluency in children with autism spectrum disorders: Clustering and switching strategies. *Autism*, 18(8): 1014–1018.

Breddemann, A.; Schilbach, L.; Kunerl, E.; Witzmann, M.; and Schuwerk, T. 2023. Geschlechtsunterschiede in der Autismusdiagnostik. *Psychiatrische Praxis*.

Ghaziuddin, M.; and Gerstein, L. 1996. Pedantic speaking style differentiates Asperger syndrome from high-functioning autism. *Journal of autism and developmental disorders*, 26(6): 585–595.

Hauser, M.; Sariyanidi, E.; Tunc, B.; Zampella, C.; Brodtkin, E.; Schultz, R. T.; and Parish-Morris, J. 2019. Using natural conversations to classify autism with limited data: Age matters. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019)*, Minneapolis, Minnesota, Jun 6th 2019, 45–54.

Hosseini, S. A.; and Molla, M. 2023. *Asperger Syndrome*. StatPearls Publishing, Treasure Island (FL).

Liu, D.; Liu, Z.; Yang, Q.; Huang, Y.; and Prud’hommeaux, E. 2022. Evaluating the Performance of Transformer-based Language Models for Neuroatypical Language. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*, Gyeongju, Republic of Korea, Oct 12th – 17th 2022, 3412–3419. International Committee on Computational Linguistics.

Luyster, R. J.; Zane, E.; and Wisman Weil, L. 2022. Conventions for unconventional language: Revisiting a framework for spoken language features in autism. *Autism & Developmental Language Impairments*, 7: 23969415221105472.

Mulac, A.; Lundell, T. L.; and Bradac, J. J. 1986. Male/female language differences and attributional consequences in a public speaking situation: Toward an explanation of the gender-linked language effect. *Communications Monographs*, 53(2): 115–129.

Parish-Morris, J.; Liberman, M.; Ryant, N.; Cieri, C.; Bateman, L.; Ferguson, E.; and Schultz, R. T. 2016. Exploring Autism Spectrum Disorders Using HLT. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2016)*, San Diego, CA, USA, Jun 16th 2016, 74–84. Association for Computational Linguistics.

Probol, N.; and Mieskes, M. 2024. Autism Detection in Speech – A Survey. In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 1115–1125. St. Julian’s, Malta: Association for Computational Linguistics.

Prud’hommeaux, E.; van Santen, J.; and Gliner, D. 2017. Vector space models for evaluating semantic fluency in autism. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL 2017)*, Vancouver, Canada, Jul 30th – Aug 4th 2017, 32–37. Association for Computational Linguistics.

Wing, L. 1981. Asperger’s syndrome: a clinical account. *Psychological medicine*, 11(1): 115–129.

A Appendix

Table 5 shows all the social media creators who allowed us to use their data for research purposes. The creator name is embedded in the link. Their accounts are linked in the second column. The first part are the female creators in the second part, the male creators are listed. Some creators do have more than one account and on different platforms. The listed link is the largest one or the one we found first. Additional accounts are usually listed on their account in our list, however, a list of all used videos from all different platforms can be found on our GitHub Repository.¹⁶

Link

<https://www.youtube.com/@MomontheSpectrum>
<https://www.youtube.com/@imautisticnowwhat>
<https://www.youtube.com/@IndigoChina>
<https://www.youtube.com/@WoodshedTheory/about>
<https://www.youtube.com/@StephanieBethany>
https://www.tiktok.com/@kaelynn_vp
<https://www.tiktok.com/@tesslathey>
<https://www.tiktok.com/@autisticayla>
<https://www.instagram.com/myfavouritejo/>
<https://www.instagram.com/ashralouisa/>
<https://www.instagram.com/elenacarr0ll/>
<https://www.instagram.com/michelevision/>
<https://www.youtube.com/@Aneva>
<https://www.youtube.com/@NeurodiverJENnt>

<https://www.youtube.com/@TheAspieWorld/videos>
<https://www.youtube.com/@GenericArtDad>
<https://www.youtube.com/@TheCakeIsNotaVlog>
<https://www.youtube.com/@ThomasHenley>

Table 5: Social media creators that allowed the usage of their data.

¹⁶<https://github.com/Nadine271/female-autism>