Computational Linguistics in the Netherlands Journal 8 (2018) 3–23 Submitted 05/2018; Published 12/2018

Stylometric Text Analysis for Dutch-speaking Adolescents with Autism Spectrum Disorder

Luna De Bruyne* LUNA.DEBRUYNE@UGENT.BE
Ben Verhoeven** BEN.VERHOEVEN@UANTWERPEN.BE
Walter Daelemans** WALTER.DAELEMANS@UANTWERPEN.BE

*LT3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium

**CLiPS, University of Antwerp
Prinsstraat 13, 2000 Antwerp, Belgium

Abstract

One of the main characteristics of individuals with autism spectrum disorder (ASD) is a deficit in social communication. The effects of ASD on both verbal and non-verbal communication are widely researched in this respect. In this exploratory study, we investigate whether texts of Dutch-speaking adolescents with ASD (aged 12-18 years) are (automatically) distinguishable from texts written by typically developing peers. First, we want to reveal whether specific characteristics can be found in the writing style of adolescents with ASD, and secondly, we examine the possibility to use these features in an automated classification task. We look for surface features (word and character n-grams, and simple linguistic metrics), but also for deep linguistic features (namely syntactic, semantic and discourse features). The differences between the ASD group and control group are tested for statistical significance and we show that mainly syntactic features are different among the groups, possibly indicating a less dynamic writing style for adolescents with ASD. For the classification task, a Logistic Regression classifier is used. With a surface feature approach, we could reach an F-score of 72.15%, which is much higher than the random baseline of 50%. However, a pure n-gram-based approach very much relies on content and runs the risk of detecting topics instead of style, which argues the need of using deeper linguistic features. The best combination in the deep feature approach originally reached an F-score of just 62.14%, which could not be boosted by automatic feature selection. However, by taking into account the information from the statistical analysis and merely using the features that were significant or trending, we could equal the surface-feature performance and again reached an F-score of 72.15%. This suggests that a carefully composed set of deep features is as informative as surface-feature word and character n-grams. Moreover, combining surface and deep features resulted in a slight increase in F-score to 72.33%.

1. Introduction

1.1 Background

Language can reveal more about a speaker or writer than one would suspect. Besides the actual content, texts contain a lot of information about the speaker or writer, reflected in the writing style. In the field of computational stylometry, writing styles are automatically analyzed for authorship attribution or author profiling. For the task of author profiling, systems are developed to detect sociological (age, gender, education level, etc.) or psychological properties of the author of the text (e.g. personality or mental health), which can be applied for forensic purposes, in literary science, sociolinguistics, etc. (Daelemans 2013, Pennebaker 2011).

Another field in which computational stylometry can be useful, is medical diagnosis. Especially in research on Alzheimer's disease and related forms of dementia, text analytic approaches prove to be promising (Croisile et al. 1996, Snowdon et al. 1996, Riley et al. 2005, Baldas et al. 2010, Le et al.

2011, Hirst and Wei Feng 2012). The challenge here is not only to distinguish texts from individuals with a certain cognitive or developmental disorder from individuals without that disorder, but also to support the formulation of hypotheses for future research. Such follow-up studies could deal with the explanation of possible differences in writing style and can provide a better understanding of how specific disorders affect language. This approach, where computational stylometry does not restrict itself to classification but also aids explanation, was already suggested by Daelemans (2013) and Regneri and King (2016).

In this study, we will use such an approach for the analysis of Dutch texts written by Flemish, high-functioning adolescents with autism spectrum disorder (ASD). In DSM-5, the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (American Psychiatric Association 2013), the term ASD is used to indicate disorders that show two types of symptoms:

1. Deficits in social communication and social interaction;

2. Restricted, repetitive patterns of behavior, interests or activities.

Regarding social communication, DSM-5 focuses mainly on deficits in non-verbal communication, like badly integrated verbal and non-verbal communication, difficulties in understanding gestures and body language, and a lack of eye contact. Formal competences like vocabulary and grammar are not necessarily affected. However, a lot of people with ASD do show affected language: in some cases, speech is completely absent or its development can be delayed. Apart from that, DSM-5 does not report any linguistic characteristics of individuals with ASD.

1.2 Aims and research questions

This is an exploratory study, the aim of which is twofold: on the one hand we want to reveal whether there are specific features in the written language of people with ASD, and how we can interpret those features. On the other hand, we want to investigate the possibility of using these features in an automated classification task.

Our work differs from most other works on language and ASD in various ways: firstly, we examine written language, while most (but not all) studies focus on oral narrations. Secondly, we develop a method for automatically extracting features for Dutch texts. To our knowledge, only van Halteren and op de Weegh (2012), investigated such approaches for this language in the context of ASD. Also, we use a notably larger dataset than most other studies (which is possible since features are not annotated manually in this study, but they are extracted automatically). Finally, we use machine learning techniques to automatically classify the analyzed texts. Most work only focuses on finding differences in language use, but not on using these differences for identifying ASD.

The aspects that are examined comprise a wide range of features that can be subsumed under two types: surface features and deep features. Surface features do not need a deep linguistic understanding and include word and character n-grams and simple linguistic features like average word and sentence length. Deep features are linguistically more complex and include syntactic features (like part of speech tag frequencies, patterns of constituents, etc.), semantic features (frequency of word categories, propositional idea density) and discourse features (with a main focus on cohesion). Using two-sample t-tests, we want to find out which features show significant differences between the ASD and control group

For the classification task, we compare three feature set-ups: one where only surface features are used, one where only deep features are used, and finally the combination of deep and surface features. The largest part of the surface features consists of n-grams. Counting word and character n-grams is a common technique in computational stylometry and has, despite its simplicity, proven to be very informative. We thus suspect that these features will already give relatively good results. However, in this study we are particularly interested in the relevance of deep features, as they, unlike n-grams, do not (or less) rely on content.

We based our work on a dataset which we gathered ourselves, consisting of 140 Dutch texts: 70 texts were written by high-functioning adolescents with autism spectrum disorder, and 70 texts were written by a control group without ASD, matched by age, gender and education. All texts were school assignments and have a rather formal register. The texts are not genre or topic restricted, but there are some consistencies in the topic choice between schools, which has some consequences for the writing style and increases the risk of topic detection instead of style detection. However, considering the lack of existing datasets and the difficulty of obtaining a dataset that is restricted in topic and still large enough for our purposes, we regard this dataset as the best option.

The paper is organized as follows: Section 2 discusses related work on language and communication in ASD (2.1), and on computational approaches for the classification of ASD (2.2). In Section 3, we discuss how we collected (3.1) and prepared the data (3.2), and how we extracted features from these data (3.3). Section 4 discusses the results of the statistical analysis (4.1), and describes the experimental set-up and results of the classification task (4.2). Section 5 is devoted to the discussion, where we give a summary of our findings (5.1) and consider some limitations of this study (5.2). An overall conclusion is given in Section 6.

2. Related Work

2.1 Language and communication in ASD

DSM-5 does not report any specific characteristics of the (spoken or written) language of people with ASD, except for the absence of speech in some individuals or a delay in the development of language. In fact, this delay is the main cue for the distinction between Asperger syndrome and other autistic disorders: children with Asperger do not show delays in language or cognition (American Psychiatric Association 2013). Slower language development is thus not sufficient for a diagnosis of ASD, although most children in this group do show such regressions. Moreover, there is a notable heterogeneity in the language abilities of individuals with ASD (Tager-Flusberg 2004). In what follows, we will only discuss language of individuals with high-functioning ASD (i.e. not language impaired and with normal cognition).

Delfos (2011) emphasizes that people with ASD mainly struggle with a communication problem and not necessarily with a language problem. Apparent problems in the language of a person with ASD are supposedly effects of the underlying communication deficit. Also Happé and Frith (1996) come to this conclusion and according to them, the problem is not in the development of phonology or syntax, but in the act of 'communicative sharing'.

However, several studies have shown some discrepancies between the language of people with ASD and neurotypical controls. Most studies have been performed on oral utterances of children with ASD, and only little research has been done on their writing style. Yet, there is evidence that deficits in oral language also have an impact on the production of written texts (Berninger et al. 2006, Mackie and Dockrell 2004, Wagner et al. 2011). Below, we describe some previously examined characteristics of the language of children and adolescents with ASD, either in their written or in their spoken language.[1]

2.1.1 General peculiarities in language and communication

As mentioned before, some people with ASD never develop speech. Yet, the opposite can be true as well: some individuals with ASD show hyperlexia (Delfos 2011). They produce a flood of words and sideline their conversation partner.

A characteristic aspect of peculiar speech in ASD is the occurrence of echolalia (Tager-Flusberg et al. 2005), the repetition of someone else's utterances, with similar intonation. It is a classic

1. The characteristics mentioned in reference works like DSM-5 are supposed to be language independent. Empirical studies cited in this section all investigated texts or narrations of English-speaking persons.

symptom of autism, already described in 1946 by Kanner (1946), but it is not present in all children with autism, nor is it a symptom only restricted to ASD (Yule and Rutter 1987).

Also paralinguistic aspects such as intonation, stress patterns and voice quality can be deviant in individuals with ASD (Rutter et al. 1992). Monotonous intonation is often associated with ASD, but in some cases even singsong patterns are observed (Fay and Schuler 1980). Pronovost et al. (1966) identified voice disorders like extraordinary high fundamental frequency levels, hoarseness, harshness, hypernasality and poor control of volume with deviant fluctuations.

2.1.2 Word use

A lot of individuals with ASD show abnormal use of words and phrases (Rutter 1970). This can reveal itself in neologisms or idiosyncratic language – 'metaphorical language', as it was called by Kanner (1946). Here, words are used in an unusual way (often by modifying the ordinary word root), but still make sense (e.g. 'bluesers' for 'bruises'). In the case of Asperger, extremely formal and distant language is characteristic (Delfos 2011).

Several studies state that high-functioning individuals with ASD do not differ much in their word use compared to neurotypical peers. For example, Tager-Flusberg (1985) states that children with ASD use semantic groupings in a non-deviant way. However, other studies show that they use some word classes to a lesser extent, namely mental state terms and social-emotional terms (Tager-Flusberg 1992, Tager-Flusberg and Sullivan 1994, Storoschuk et al. 1995).

2.1.3 Syntax and morphology

Children with ASD generally are considered to have intact morphological and syntactic development (Tager-Flusberg 2000, Tager-Flusberg et al. 2005). Indeed, Dockrell et al. (2014) investigated texts of students with ASD, and did not find any grammatical or spelling problems. Also Myles et al. (2003) compared the writing of students with ASD with neurotypical controls. They did not find any differences on the TOWL-3 (Test of Written Language) scores (Hammill and Larsen 1996), but other variables in their study showed that students with ASD produced shorter texts with lower syntactic complexity. Also Brown and Klein (2011) found that people with ASD wrote shorter, less complex texts, and studies on (oral) narratives by children with ASD showed lower syntactic complexity in their stories, compared to controls (Tager-Flusberg and Sullivan 1995, Diehl et al. 2006, King et al. 2014).

In a study on past tense by Bartolucci and Albers (1974), children with autism performed significantly weaker than the control group. The same was observed by Tager-Flusberg (1989). In addition to this, Bartolucci et al. (1980) found that children with ASD were more likely to omit articles, auxiliary verbs, copula verbs, third-person present tense and ing-forms.

2.1.4 Discourse

Whereas some studies on syntactic complexity are still contradictory, most researchers agree on the fact that individuals with ASD are impaired on discourse-related aspects of language.

One problem of pragmatic nature is that people with ASD tend to take words or expressions too literally. Therefore, they have problems with ambiguities and indirect language (irony, metaphors), but also with vague words and phrases like 'sometimes', 'often' or 'maybe tomorrow' (Delfos 2011). The latter is related to a problem with deixis (Tager-Flusberg et al. 2005) (i.e. references to the linguistic situation itself, where contextual information is needed for understanding). Deixis is marked by pronouns, but also by some words that express time and place (e.g. 'this', 'there', 'now').

Other problems with discourse are related to cohesion and coherence. The structure or logical flow of a text is referred to as coherence, while cohesion concerns the way sentences and clauses are linked together linguistically (Karmiloff-Smith 1985). In order to tell a coherent story, both cohesion and coherence need to be integrated. Especially in research on narratives by children and

adolescents with ASD, it was found that the stories of ASD groups were significantly less coherent than the stories of controls (Loveland and Tunali 1993, Capps et al. 2000, Losh and Capps 2003, Losh and Capps 2006, Diehl et al. 2006). However, Tager-Flusberg and Sullivan (1995) did not find such differences.

2.2 Computational approaches for the classification of ASD

The language difficulties discussed above have been analyzed through various approaches. Hand-coding linguistic features (which is the most frequently used approach) is labor-intensive and causes difficulties in consistency. Therefore, automated approaches are desirable. Moreover, automated systems can be used for the classification and detection of ASD.

Studies that use computational approaches for the analysis of English texts of children or adolescents with ASD, but do not make use of classifiers are for example those of Regneri and King (2015, 2016). In the study of 2015, general language competence features (proportions of low-frequency words and pronoun use), topic coherence features (by measuring tf-idf), and features related to the expression of sentiment are analyzed. In the study of 2015, the feature set consists of coreference related measures, using Stanford CoreNLP (Manning et al. 2014). Losh and Gordon (2014) use Latent Semantic Analysis (LSA) for measuring semantic quality in narratives of children with ASD and typically developing controls. They showed that the semantic content was similar among both groups when they used a picture book for their narrations, but that semantic quality decreased for the ASD group in a narrative recall task. Yet, they did not perform a classification task.

Prud'hommeaux et al. (2011) and Rouhizadeh et al. (2013) did introduce a classification model for English ASD texts. In the former, spontaneous speech of 50 children (4-8 years) from three diagnostic groups was analyzed: ASD, typical development, and developmental language disorder. They built a model with classification and regression trees (CART (Breiman 1984)), used for diagnostic classification. Features included n-gram cross entropy features, surprisal-based features and features measuring syntactic complexity, and they reached an F1-score of 67%. In the latter, distributional semantic models are used to automatically identify unexpected words in retellings of children with ASD. Those unexpected words can distinguish the ASD narratives from the narratives of neurotypical controls with an accuracy of 83.4%.

We are aware of one study concerning an ASD classification tasks for Dutch: van Halteren and op de Weegh (2012) trained a word n-gram model (uni, bi- and trigrams) on Dutch Twitter messages written by 11 users diagnosed with ASD. The test set consisted of tweets from 15 users with ASD and control tweets of 1158 other users. There was a false reject rate for users with ASD of 7%, and a false accept rate of again 7% for control users. The most indicative words and word sequences were content words related to ASD – e.g. 'autisme' ('autism'), 'diagnose' ('diagnosis') – but also the less frequent use of hedging words and words related to personal feelings and plans was indicative.

3. Data

3.1 Data collection

For the purpose of this study, we collected 140 Dutch texts, of which 70 were written by adolescents diagnosed with ASD and 70 by neurotypical controls, matched by age, gender and education.

We first collected texts of adolescents diagnosed with ASD, by contacting schools in the Flemish special secondary education. In Flanders, children with ASD are free to choose between mainstream education and special education. The special education structure consists of four education forms and nine types, of which Type 9 is intended for children with ASD.[2] The four education forms

2. Type 9 of special education, intended for children with ASD, is a very recent development in the special education typology and only exists since 2015. Most children with ASD who did not go to regular schools, were mainly

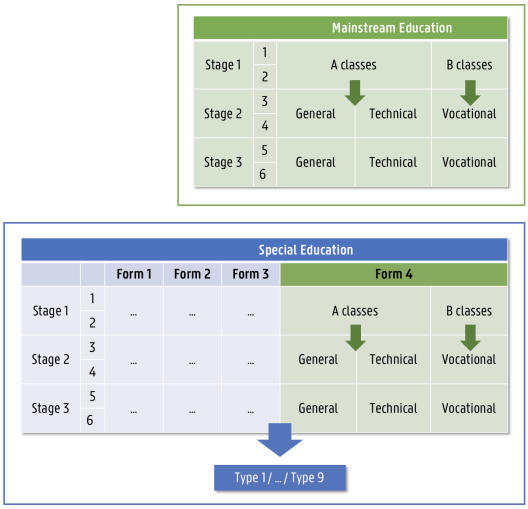

Figure 1: Simplified schematic representation of mainstream and special education system in Flanders.

determine the level of education, where Education Form 4 is the highest level and offers the same program as mainstream education, but with special assistance for the needs of the student (see Figure 1 for a simplified, schematic overview of the Flemish education system).

We contacted all schools that offer Type 9 in Education Form 4 and asked if we could use some texts written by their students. All texts are school assignments, but the instructions were different among schools and assignments. Mostly, the students had been asked to write in a specific genre or about a specific topic (e.g. their internship experience, the ending of a fairy tale, a book report), but sometimes they were free to write something about their interests or life. The texts were written individually, but the circumstances varied: the texts could either have been written in class or at home (which has implications for the access of tools like dictionaries and Internet) and could have been written by hand or typewritten (probably word-processed, implying that there may have been spelling, style or grammar correction).

We were able to gather 70 texts, written by students from age 12 to 18. Only 6 out of 70 students were girls.[3] Students from year 1 and 2 (from now on called Stage 1) came from A classes (in the Belgium school system, this is the education with the aim of moving up to General or Technical Secondary Education) and B classes (preparation for Vocational Secondary Education). Students from year 3 and 4 (Stage 2) came from General and Technical Secondary Education, and students from the two last years (Stage 3) all did courses in Technical Education.

When the data of the ASD group was gathered, we started looking for control data in the mainstream education. For each subject in the ASD group, we matched a control subject by age, gender and education (however, not by assignment topic, seeing the wide variation of topics and assignment instructions), in order to get a balanced dataset. As a result, the control group also consisted of 70 texts, again with 6 girls and 64 boys (see Table 1).[4]

included in Type 1 (for children with a light mental disability) or Type 7 (for children with an auditory impairment, or speech or language impairment).

3. This is in line with observations that ASD affects males at a higher frequency than females, with a 4:1 ratio. For high-functioning adolescents, it is stated that the differences are even larger (6 males for 1 female) (Fombonne 1999).
4. Although the data was anonymized, we did not get permission to make the data publicly available.

Year	Boys	Girls
Stage 1	9	0
Stage 2	38	5
Stage 3	17	1
Total	64	6

Table 1: Distribution of the 70 texts of both groups, regarding gender and education stage. For the total number of texts in the dataset, all numbers need to be doubled, since the distribution is exactly the same for the ASD and control group.

3.2 Data preparation

Most of the assignments were typewritten, but some were written by hand. We transcribed all handwritten texts, and converted all typed assignments to plain text files. We also removed images and headings. Every text was manually checked for misspellings. We assume that a lot of the misspellings were due to the young age of some of the participants, and as most of our features rely on correct orthography, all misspellings were removed. However, grammar mistakes were not corrected.

All plain text files were processed by the language tool Frog (van den Bosch et al. 2007), which is a memory-based morphosyntactic tagger and parser for Dutch. For each plain text file, Frog returns an output file with morphological and syntactic information for each word (e.g. lemma, morphological segmentation, POS tag, etc.). These 'frogged files' will be used for the feature engineering part (see Section 3.3).

The mean number of words per text is 338 for the ASD group and 382 for the control group, with a standard deviation of 189 for the former and of 236 for the latter. Table 2 also shows the mean number of words per text for each education stage. A two-sample t-test showed no significant difference for the mean text length between the ASD and control group (p=.391).

	ASD			Control		
Year	mean	min-max	SD	mean	min-max	SD
Stage 1	289	108-813	231	429	109-1275	406
Stage 2	318	131-952	180	321	129-1004	149
Stage 3	410	169-958	180	505	188-1226	257
Total	338	108-958	189	382	109-1275	236

Table 2: Average text length, minimum and maximum length and standard deviation per group and education stage.

3.3 Feature extraction

The data are represented by a set of 69,890 features, consisting of five categories originating from two main types: n-grams and simple linguistic features (surface features), and syntactic features, semantic features and discourse features (deep features). Table 3 gives an overview of these features, and in what follows we discuss them more thoroughly.

3.3.1 Surface: n-grams

For this category, we use tri- and tetragrams for characters and uni- and bigrams for words.

3.3.2 Surface: Simple linguistic features

Since ordinary word and character n-grams are highly sensitive to content, we extracted unigrams of function words (pronouns, prepositions, conjunctions, articles, numerals and adverbs) as simple linguistic features. Other features in this category are average word length, average sentence length and average paragraph length. The latter is given both in number of words and sentences.

3.3.3 Deep: Syntactic features

Syntactic features include POS-tag frequencies, the frequency of auxiliary verbs and the frequency of function words. We look for syntactic patterns by extracting POS n-grams and constituent n-grams, both with n equal to 2 and 3. We also take the POS-tag of the first token in each sentence in a text as a feature, because this gives an idea of syntactic complexity and variation. For each text, we calculate the proportions of all POS-tags that appear in the first position in a sentence. As additional features, we provide the skewness of those tags, and the proportion of the POS-tag that occurs most often at the first position. These last two features give an idea of syntactic diversity: if words with one specific POS-tag occur almost always at the first position of the sentence (high proportion), then this is a sign of little syntactic diversity. Also a skewed distribution of POS-tags of such 'first words' can be a sign of low diversity. In contrast, a symmetric distribution indicates that there is a better variation in sentence structure.

Other features for measuring syntactic complexity are the average number of phrases (constituents) and average number of (finite) clauses per sentence. The number of phrases was calculated by counting all B-tags (represents the beginning of the phrase) in the phrase chunk. For the number of clauses, we just counted all finite verbs. Of course, only finite clauses are captured in that way, and non-finite clauses are ignored.

Surface features	**Deep features**
N-grams	**Syntactic**
word n-grams (40,100)	POS-tag freq. (12)
char n-grams (27,147)	auxiliary freq. (1)
	function word freq. (1)
Simple	POS n-grams (142)
function word unigrams (1,759)	phrase n-grams (614)
average word length (1)	POS first token (13)
average sentence length (1)	phrases per sentence (1)
average paragraph length (2)	clauses per sentence (1)
	Semantic
	LIWC word categories (68)
	idea density (1)
	Discourse
	discourse connectives (5)
	lexical chain measures (5)
	average similarity (10)
	word overlap (7)

Table 3: Overview of the feature groups per category.

3.3.4 Deep: Semantic features

An important semantic feature is the frequency of various word categories. These categories are based on the work of Pennebaker et al. (2001) – the Linguistic Inquiry and Word Count (LIWC) – and adapted to Dutch by Zijlstra et al. (2004). Examples of such 'LIWC' categories are 'negation', 'family', 'religion', 'sport' and 'motion'. All categories can be grouped into five bigger classes: 'linguistic dimensions', 'psychological processes', 'relativity', 'personal matters' and 'other'. These bigger classes can be of particular interest to validate whether individuals with ASD use fewer mental-state terms and social-emotional terms (the 'psychological processes' class).

Another semantic feature we investigate is propositional idea density. This gives the number of propositions or ideas in a text, normalized by text length. We used a computerized propositional idea density tool for Dutch, as developed by Marckx (2017) for the detection of Alzheimer's disease.

3.3.5 Deep: Discourse features

This category of features mainly focuses on text cohesion and coherence, since previous research revealed that individuals with ASD have difficulties with producing coherent texts. Surprisingly, most of these studies do not rely on objectively quantifiable features, but use holistic, manually annotated ratings.

Two studies do give a suggestion for features that can be automatically quantified: Regneri and King (2015) use tf-idf to detect unusual off-topic words, and Regneri and King (2016) extract measures from coreference chains, using Stanford CoreNLP. Although the last suggestion seems promising, implementation for Dutch texts is very difficult at the moment, as existing Dutch coreference resolution tools do not yet reach sufficient performance for our purposes.

For quantifying text cohesion, we used discourse connectives for measuring cohesion on sentence and paragraph level. Based on the explicit connectives in the Penn Discourse Treebank (Miltsakaki et al. 2004), we made a list of Dutch connectives and tagged them with one of four semantic classes: temporal, contingency, comparison or expansion For example, 'ondertussen' ('meanhwile') was tagged with temporal, 'bijgevolg' ('consequently') with contingency, 'in tegenstelling tot' ('in contrast with') with comparison and 'also' with expension.[5] We use the total proportion of connectives in a text and the proportion per semantic class as features.

We also calculated cohesion by using lexical chains. These are chains of words that are related to each other, or express the same topic. We extract lexical chains on the basis of (cosine) similarity scores between word embeddings in the text, and use the number of chains, average chain length, minimum length and maximum length as features. We also compute a 'cohesion score', by adding up the similarity scores within the chains, and dividing that by the total number of chains. That way, a text with few but long chains gets a high cohesion score (because there are only a few topics, but they are discussed extensively), but when there are many short chains, the cohesion score is low (due to fragmentation of topics which are not discussed thoroughly).

We present two additional approaches for capturing cohesion, one based on word overlap and one based on word similarity. For the latter, we used word embeddings to construct sentence vectors, and by comparing all sentence vectors with each other, we get a measure for average (cosine) similarity.

For the detection of problems with deixis, we rely on the the POS-tag frequency for pronouns (included in the category of syntactic features) and two LIWC categories of 'relativity', namely 'time' and 'space' (semantic features).

5. We did not take into account ambiguity, but this is something we would like to further optimize in future research.

4. Experiments & Results

4.1 Statistical analysis

First, we want to get a clearer view on the data and on how the data are represented by the features. We compare means between the ASD and control group for all features except for n-gram features. For word categories, we made a selection of those categories that are the most likely to be distinctive according to previous studies ('affective processes', 'cognitive processes', 'perceptual processes', 'social processes' and the categories 'space' and 'time', see Section 2.1). We perform two-tailed independent sample t-tests at a 5% significance level.[6]

For simple linguistic features, we found that the mean word length was 4.7 characters for both groups, the mean sentence length 17 words for the control group and 18 words for the ASD group, and mean paragraph length 6 sentences or approximately 100 words (both groups). These differences were not significantly different among the groups (p=.99 for average word length; p=.41 for average sentence length; p=.79 for average paragraph length in sentences; p=.49 for average paragraph length in words).

Regarding the syntactic features, 10 out of 29 features were significantly different between the ASD and control group. Especially the use of lexical classes (POS-tags) turned out to be distinctive. Adolescents from the ASD group used fewer adjectives (p=.034) and articles (p=.02), but more verbs (p=.043), pronouns (p=.037) and conjunctions (p=.004). Moreover, they used more auxiliaries (p=.01) and function words (p=.049) than the control group. Adolescents with ASD used fewer numerals, but this difference did not reach significance (p=.051). In what follows, we will call features that are not significant but have have a p-value under 0.10 (like numerals) 'trending' features'. The use of nouns, prepositions, adverbs and interjections did not differ between groups.

Regarding syntactic complexity, we see no significant difference in the number of phrases and clauses per sentence (p=.17 and p=.13), but we do see a trend in the variation of sentence structure: when we look at the lexical classes that occur at the beginning of a sentence, we see that the frequency of the lexical class that occurs most often at the first position is higher for the ASD group than for controls (p=.08). This could be an indication that there is less variation in sentence structure for adolescents with ASD. We also looked at the distribution of the lexical classes that occurred at the first position of sentences and saw that the distribution was more skewed for the ASD group (p=.08). This suggests that there is less variation in their texts. However, these measures for syntactic variation are merely a trend and do not reach significance. We did find a significant difference in the use of pronouns, articles and conjunctions as the first word of a sentence. Adolescents from the ASD group were more likely to begin a sentence with a pronoun (p<0.01) or conjunction (p=.01) than controls, whereas neurotypical peers started their sentences more often with an article than the ASD group (p<.01).

In the category of semantic features, we tested for differences in propositional idea density and for different use of some of the LIWC word categories. Idea density was similar for the texts of the ASD group and the control group p=.46). However, we did see some trends in the frequency of word categories. Adolescents from the ASD group used fewer affective words (p=.07) and (surprisingly) more words related to cognitive processes (p=.07). For the use of words referring to time and space and social words, the differences between both groups were not large enough to be distinctive (p=.17; p=.19; p=.85).

We investigated four subgroups of discourse features. For the first one, discourse connectives, there were no significant differences in the use of specific types of connectives (temporal, contingency, expansion, comparison) between the groups (p=.74; p=.15; p=.88; p=.45). However, when connectives were analyzed in general, students with ASD seemed to use them to a lesser extent than neurotypical peers. Yet, this was only a trend and was not significant (p=.07). Regarding word overlap, we could not find any differences (.45<p<.73). We also tried to measure cohesion in terms

6. As this is an exploratory study, we do not adjust the p-values with Bonferroni correction.

of average similarity (using word embeddings), but this did not provide any significant results either ($.13<p<.99$). Our last method, lexical chains, was more promising: the maximum chain length was significantly lower for adolescents with ASD (10 words) as compared to controls (13 words) ($p<.01$). Also the average chain length and lexical cohesion score were lower in the ASD group, but only average chain length was trending ($p=.06$ and $p=.17$).

In sum, only 11 out of 67 investigated features show a significant difference in mean, and 7 features were trending. Most of them are syntactic features. We could say that students with ASD have a more formal, less dynamic writing style, since they use more function words, but for example fewer adjectives. There is also an indication (although no significant evidence) that there is less variation in their sentence structure. Analysis of lexical chains suggests that students with ASD write less coherent texts, but judging from these particular data, we have to be careful with such claims.

Even when there are no or only few clear significant correlations between linguistic features and the ASD or neurotypical groups, it may still be the case that Machine Learning methods find predictive patterns in the data. We investigate this in the next section.

4.2 Classification task

4.2.1 Main experiments

We create a model to detect texts written by adolescents with ASD. Since the number of features (69,890 in total) is large in comparison with the number of data instances (only 140), we employ a Logistic Regression classifier, implemented by scikit-learn (Pedregosa et al. 2011).[7] The performance of the system is evaluated by 10-fold cross validation, using the default parameters.

We test different combinations of all feature categories: n-grams (N), simple linguistic features (Si), syntactic features (Sy), semantic features (Se) and discourse features (D), as discussed in Section 3.3. We try the features of all categories separately, every combination of two, three and four categories, and the combination of all five feature categories, which makes a total of 31 different combinations. The results are shown in Table 4, where we also show the distinction between surface features, deep features and combinations of both.

	1 cat.		2 cat.		3 cat.		4 cat.		5 cat.	
	Feat.	F1	Feat.	F1	Feat.	F1	Feat.	F1	Feat.	F1
Surf.	N	**71.57**	N, Si	70.13						
	Si	55.00								
Deep	Sy	53.57	Sy, Se	58.59	Sy, Se, D	59.29				
	Se	58.57	Sy, D	52.86						
	D	55.04	**Se, D**	**62.14**						
Combi			**N, Sy**	**72.33**	N, Si, Sy	69.53	N, Si, Sy, Se	69.53	N, Si, Sy	68.85
			N, Se	71.57	N, Si, Se	70.13	N, Si, Sy, D	68.85	Se, D	
			N, D	71.57	N, Si, D	69.45	N, Si, Se, D	69.29		
			Si, Sy	60.00	**N, Sy, Se**	**72.33**	N, Sy, Se, D	71.65		
			Si, Se	57.86	N, Sy, D	71.65	Si, Sy, Se, D	62.15		
			Si, D	55.00	N, Se, D	71.57				
					Si, Se, D	57.14				
					Si, Sy, Se	60.00				
					Si, Sy, D	62.15				

Table 4: F1-scores (%) of different combinations of the feature groups. The best F-score with surface features is shown in yellow. The best scores when only deep features are used are shown in red, and the best scores for combinations of surface and deep features are shown in blue.

7. We also did some experiments with other classifiers. As expected, SVM with linear kernel performed rather similarly, yet Logistic Regression seemed to slightly outperform SVM.

Since we use a balanced dataset, the statistical baseline (both weighted random baseline and majority baseline) is 50%. When we train a model using n-gram features, this baseline is outperformed significantly, namely with an F-score of 71.57%. This is the highest F-score we could achieve when only one feature category was used (see first column Table 4). Experiments with other isolated categories of features give F-scores between 53.57% and 58.57%. These results are much lower than when n-grams are used, but they are still higher than the statistical baseline. Apart from n-grams, semantic features give the best results in isolation, and syntactic features the weakest. These results do not completely correspond to the findings of the statistical analysis, where syntactic features were shown to be the most distinctive. However, all features in the particular set —both the significant and the non-significant ones—are used here, explaining why the syntactic features do not necessarily perform best.

Discourse	**Syntactic**	
discourse connectives (total)	pos first token: article	pos: article
average lexical chain length	pos first token: conjunction	pos: numeral
max lexical chain length	pos first token: pronoun	pos: conjunction
	pos first token: verb	pos: pronoun
	max frequency pos first token	pos: verb
	skewness pos first token	frequency auxiliaries
	pos: adjective	frequencie function words
	Semantic	
LIWC Category Achieve	LIWC Category Friends	LIWC Category Pronoun
LIWC Category Affect	LIWC Category Humans	LIWC Category Relig
LIWC Category Anx	LIWC Category I	LIWC Category School
LIWC Category Article	LIWC Category Insight	LIWC Category See
LIWC Category Body	LIWC Category Leisure	LIWC Category Self
LIWC Category Cause	LIWC Category Music	LIWC Category Sexual
LIWC Category Certain	LIWC Category Negate	LIWC Category Space
LIWC Category Cogmech	LIWC Category Number	LIWC Category Sports
LIWC Category Comm	LIWC Category Occup	LIWC Category Swear
LIWC Category Discrep	LIWC Category Optim	LIWC Category TV
LIWC Category Down	LIWC Category Physcal	LIWC. Category Time
LIWC Category Eating	LIWC Category Posemo	LIWC Category Up
LIWC Category Feel	LIWC Category Posfeel	LIWC Category You
LIWC Category Hear	LIWC Category Present	

Table 5: Final set of stylometric features, after selection of the significant features.

We looked at the effect of combining surface features with deep features, and saw that some feature group combinations equaled the n-gram F-score of 71.57%, namely when semantic or discourse features were combined with n-grams. This result was even outperformed by the combination of syntactic features and n-grams, although to a limited extent (72.33% F-score). In all combinations however, simple linguistic features decrease the F-score. Indeed, the statistical analysis indicated that there were no simple linguistic features that showed significant or trending differences between the two groups (although we did not test for function word unigrams, which we did include in the classification model). In contrast, for all categories in the deep feature type (syntactic, semantic and discourse features), we did find some trends and significant differences. In spite of the informativeness of deep features, we observe that combinations of the three categories, together with the n-grams, could not outperform the F-score of the n-gram–syntax combination. This can again be explained by the inclusion of non-significant features. However, when only the significant features are combined with n-grams, the performance did not increase either. Moreover, the slightly boosting

effect of adding syntactic features is equivalent to running a feature selection algorithm on the word and character n-grams (by selecting the 10% best features according to the ANOVA F-value scoring function). Both the best feature composition in the surface feature set-up and the best composition from the combination set-up reach an F-score of 72.15%.

Although deep features did not seem to boost the system's performance significantly, we also want to explicitly compare set-ups from the deep feature types with surface features, as deep features do not (or less) rely on content. We saw that there was no combination of deep features that can compete with the performance of surface (in particularly n-gram) features. The highest F-score we achieved with deep features only, was 62.14%, namely in the combination of semantic and discourse features. However, both significant and non-significant features are present in these combination. Therefore, we tried a combination with only significant and trending features, and added LIWC features. Since only a few LIWC categories were tested for statistical significance, we started by adding all LIWC features, and then reducing the set by automatic feature selection (using Anova F-value as scoring function) in scikit-learn.[8] In this set-up, we could achieve an F-score of 72.15%, which suggests that deep features are as informative as the surface word and character n-grams, on the condition that the feature set is carefully composed. The final feature set of this set-up is shown in Table 5.

Model		Precision	Recall	F1-score
surface	Total	72.15	72.14	72.15
	ASD	72	73	72
	Control	72	71	72
deep	Total	72.15	72.14	72.15
	ASD	72	71	72
	Control	72	73	72
combi	Total	72.52	72.14	72.33
	ASD	70	79	74
	Control	75	66	70

Table 6: Results (%) of the best systems in the three different feature set-ups. For the surface features, these are the 10% best word and character n-grams (67,247 features); the deep-feature set includes only the manually selected features (based on the statistical analysis; 58 features), and combi is the combination of all word and character n-grams and all syntactic features (68,033 features).

4.2.2 Error analysis

Precision, recall and F-score of the best systems for the surface features, deep features and combined set-up are summarized in Table 6. It shows that the model performs similarly when surface features are used as when only deep features are used. For these feature set-ups, the model gives similar results for precision and recall and for the ASD and control group. When surface and deep features are combined, the system performs slightly better for the ASD group than for the controls, with a higher recall for ASD, but a higher precision for controls. If the goal is providing an aid for detecting

8. For the automatic feature selection, only the information of the training folds are used, but since our statistical analysis used all data, the manual feature selection takes the information of all instances into account. Automatic feature selection on the deep features, aiming at the same number of features as selected through manual feature selection, could not boost performance when chi square was used as scoring function (drop to 56.45%), and increased the F-score only to a limited extent with ANOVA F-value as scoring function (60.00%). The biggest difference between the manual selection and automatic feature selection is that the automatic selection chooses fewer LIWC features and more syntactic features, particularly POS and phrase n-grams.

		surface		deep		combi	
Group	Stage	#	%	#	%	#	%
ASD	Stage 1	4	44.4	3	33.3	3	33.3
	Stage 2	11	25.6	12	28.0	8	18.6
	Stage 3	4	22.2	5	27.8	5	27.8
Control	Stage 1	5	55.6	4	44.4	6	66.7
	Stage 2	11	25.6	12	28.0	13	30.2
	Stage 3	5	27.8	3	16.7	4	22.2

Table 7: Error analysis: misclassifications in absolute and relative numbers.

ASD, then these results look promising, as we should be more tolerant for false positives than for false negatives in a first screening for ASD.

When we take a closer look at the errors, we see that the texts from the first stage of secondary education are misclassified the most frequently (in relative proportions) in all feature set-ups. Especially the first stage control texts are often classified as ASD texts. In fact, they are more often misclassified than correctly classified (see Table 7). The performance of the surface-feature system is rather similar for second stage as for third stage texts. For deep features and the combined model, third stage control texts are more often correctly classified than second stage texts, but for the ASD groups, the misclassification rate varies in these feature set-ups.

5. Discussion

5.1 Summary of findings

Our findings from the statistical analysis provided only partial support for the hypothesis that the written language of high-functioning adolescents with ASD has some distinctive characteristics compared to neurotypical peers. We did not find any differences for word, sentence or paragraph length. Also the overall text length was not significantly different among groups, although this can be due to the fact that we explicitly asked for texts with sufficient length.

These findings suggest that adolescents with ASD do not show abnormalities in text production. However, we have to emphasize that this is only true for the sample we investigated. Special education schools that also offer Education Form 3 stated that most of their students had problems with producing written texts. Yet, it is disputable whether these students belong to the high-functioning group in the autism spectrum.

Previous studies suggested that adolescents with ASD would write less syntactically complex sentences. We did not find evidence for that regarding the number of phrases and clauses per sentence: students with ASD even tended to use more phrases and clauses, but this did not reach significance. We did find some significant and trending differences in the use of lexical classes (POS-tags), both in general and when we looked at the POS-tag of the first word in sentences. Overall, the ASD group used significantly more function words and fewer adjectives, which can be interpreted as a sign of a less dynamic writing style.

There were no significant differences found for semantic features, only a few trends in the frequency of various LIWC categories (words related to affect and cognitive processes). For discourse features, we did not find significant differences in word overlap or average similarity. We did see a trend in the use of discourse connectives and found that the texts of students with ASD contained shorter lexical chains, which could be an indication of less cohesion.

We have to keep in mind that the present study only focused on written text. Some researchers argue that the cognitive task of writing differs from speaking (Grabowski 2010), although this is contradicted by others (Shanahan 2006). The circumstances in which an oral or written narrative is established do differ, particularly in that writing is less spontaneous and less affected by time pressure

than speaking, and that oral utterances are often produced in social contexts, while this is not the case for writing products. This suggests the hypothesis that more characteristics indicative of ASD can be found in spoken language, as these indicators can be corrected or compensated in written language. For future research, it would be interesting to compare our results with experiments on (transcribed) speech data instead of written data.

In the second part of this study we performed classification experiments to examine whether linguistic characteristics can be used for detecting ASD. One of the feature set-ups we examined was a set-up of surface features, which mainly consisted of word and character n-grams. An n-gram-based method is commonly used in computational linguistics and classification tasks and has proven to be very informative, but it is highly sensitive to systematic differences in content. As the texts from our dataset were assignments that differed among schools, a word and character n-gram approach runs the risk of classifying on topics instead of on ASD-related features. Therefore, we performed experiments to explicitly test the relevance of deep features in capturing the writing style of people with neurological and neurodevelopmental disorders.

We reached 72.15% for the surface-feature set-up (with automatic feature selection), which is a lot higher than the statistical baseline of 50%. When only deep features were used, our initial model achieved an F-score of only 62.14%. However, after manual feature selection, we could obtain an F-score of 72.15%. The combination of surface and deep features gave an F-score of 72.33%. This leads us to conclude that in these experiments, deep features are as informative as surface features, but that the performance could not be boosted notably by combining them. Although n-grams are commonly used and easy to extract, we see a surplus value in deep features, since they are less topic sensitive and can reveal writing style characteristics more clearly.

Error analysis showed that the misclassification rate was not equally distributed over ages. Mainly texts written by adolescents from the first two years of secondary school were misclassified. This can be explained by the small number of texts in this particular subset of the data, but it is also possible that the writing style of first stage students shows less differences between children with and without ASD because both groups still have some writing problems (e.g. regarding syntactic complexity, cohesion, etc.) and thus are harder to distinguish. Maybe, neurotypically developing individuals learn to deal with these challenges when they grow older, while these characteristics keep being present in the texts of students with ASD (which corresponds to the lower misclassification rate of third stage control texts). Of course, this is only a hypothesis, and could be an interesting topic for further research.

5.2 Limitations

Some methodological matters have to be considered in order to objectively interpret the results from the previous sections.

5.2.1 Dataset

We discuss some limitations of our dataset. Although our dataset is already larger than most of the datasets used in previous research, a collection of 140 texts is still rather small. Additionally, our dataset consists mainly of texts written by students from the third and fourth year of secondary education, and only a limited number of first and third stage texts.

Another limitation is that the texts in our dataset consist of a variety of genres and, in particular, of topics. Yet, this can be seen as an advantage too, since it gives us the opportunity to investigate whether our model also works without using surface features that largely depend on content. Moreover, this is a better representation of the everyday situation, since 'texts in the wild' are not all about the same topic. On the other hand, when texts do treat the same topic, they are easier to compare. After all, word n-grams are not only useful for topic classification, but can also be predictive for word use within the same topic. When different words can be used to say the same thing, some groups can systematically prefer one word to another.

The texts from our dataset were not written in controlled circumstances. We have no knowledge about how long it took the subjects to write their assignments, and whether they used tools like dictionaries, etc. Some adolescents wrote their assignments at school, which probably gave them less access to such tools, and others wrote them at home. Also the manner of text production differed among subjects: some wrote their texts by hand, while others typed them on a computer (probably using word-processors).

A clear limitation is that we do not have any information about the presence of other disorders or learning disabilities in our dataset. Dyslexia and dysorthography for example, have an impact on writing ability, and are common among both individuals with and without ASD. It has been stated that ASD has a high comorbidity rate: more than 70% of children with ASD have co-occurring conditions (Simonoff et al. 2008). The influence of these comorbid disorders remains unclear.

We assumed that all participants had normal intelligence, seeing that the adolescents with ASD followed education in Education Form 4, and our control group was gathered in mainstream education, either in General or Technical Education. We think this is a fair assumption. Yet, it would be interesting to have more detailed insight in the cognitive abilities of our participants.

5.2.2 Feature extraction

We tried to measure some characteristics of Dutch texts automatically with computational methods. Overall, we think that the metrics give a good indication of both surface and deeper linguisitic features in texts.

However, still a few characteristics were not measured in this study. The presence of metaphorical language for example, something that a lot of people with ASD have problems with, is very difficult to trace without human judgment. Also grammaticality and misspelling rates were not included as features. For grammaticality, a commonly used measure is the number of correct word sequences or CWS (Gansle et al. 2006), but to our knowledge, an automatic CWS rater has not been developed yet.

Regarding misspellings, we manually corrected all texts, but we did not take the number of corrections into account as a feature. This was a conscious choice, since we had no information about the participants' learning disabilities and because misspelling rate is influenced by the way the texts are produced (written by hand or typewritten).

It would be useful to investigate to what extent the discourse features match with human judgments of text coherence. We did perform some tests to evaluate our features, but these were rather superficial. Also, we recommend to add coreference chain measures as features. This approach was introduced by Regneri and King (2016) and shows promising results. However, before we can implement this, the existing coreference resolution methods for Dutch need to be optimized.

5.2.3 General limitations

In this study we compared texts of (high-functioning) adolescents with ASD with texts of typically developing peers. Yet, we can not generalize our results to all adolescents with ASD, since we only included students from Education Form 4 in our study. Moreover, ASD is a collection of disorders, including classic autism and Asperger syndrome, which could possibly affect language in a different way. We do not know how these subtypes of ASD are represented in our data, which makes generalization harder. Additionally, we are not sure whether the features that distinguish adolescents with ASD from neurotypical peers are specific to ASD, or if they are related to developmental difficulties in general. Therefore, it is recommended to use more comparison groups with other impairments in future research. Also, it would be interesting to let humans judge our text collection and compare it with the performance of the automated classification system. Now, we only compared the system's performance to a statistical baseline.

6. Conclusion

In this exploratory study, we wanted to look for characteristic features in the written language of adolescents with ASD, and examined whether these features could be used in an automated classification task. Apart from surface features, we managed to define some deeper linguistic features that could be automatically extracted from the texts, namely syntactic, semantic and discourse features. Statistical analysis showed that only 11 out of 67 investigated features were significantly different in mean between the ASD and control group, and 7 features were trending. These were mainly syntactic, POS-tag related features, possibly indicating a less dynamic writing style. The deep features seemed useful in the classification task, seen that they (after selecting only the most distinctive ones) could equal the performance of a (largely on content relying) n-gram based approach. However, to use these features for practical diagnostic purposes, our model has to be further optimized in future research. We also recommend to not use a system of the kind on its own, but to see it as a preliminary diagnostic test.

References

American Psychiatric Association (2013), *Diagnostic and statistical manual of mental disorders (DSM-5)*, APA, Washington, DC.

Baldas, Vassilis, Charalampos Lampiris, Christos Capsalis, and Dimitrios Koutsouris (2010), Early diagnosis of Alzheimer's type dementia using continuous speech recognition, *International Conference on Wireless Mobile Communication and Healthcare*, Springer, pp. 105–110.

Bartolucci, Giampiero and Robert J Albers (1974), Deictic categories in the language of autistic children, *Journal of Autism and Developmental Disorders* **4** (2), pp. 131–141, Springer.

Bartolucci, Giampiero, Sandra J Pierce, and David Streiner (1980), Cross-sectional studies of grammatical morphemes in autistic and mentally retarded children, *Journal of Autism and Developmental Disorders* **10** (1), pp. 39–50, Springer.

Berninger, Virginia W, Robert D Abbott, Janine Jones, Beverly J Wolf, Laura Gould, Marci Anderson-Youngstrom, Shirley Shimada, and Kenn Apel (2006), Early development of language by hand: Composing, reading, listening, and speaking connections; three letter-writing modes; and fast mapping in spelling, *Developmental neuropsychology* **29** (1), pp. 61–92, Taylor & Francis.

Breiman, Leo (1984), *Classification and Regression Trees*, Routledge, New York.

Brown, Heather M and Perry D Klein (2011), Writing, Asperger syndrome and theory of mind, *Journal of autism and developmental disorders* **41** (11), pp. 1464–1474, Springer.

Capps, Lisa, Molly Losh, and Christopher Thurber (2000), "the frog ate the bug and made his mouth sad": Narrative competence in children with autism, *Journal of abnormal child psychology* **28** (2), pp. 193–204, Springer.

Croisile, Bernard, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet (1996), Comparative study of oral and written picture description in patients with Alzheimer's disease, *Brain and language* **53** (1), pp. 1–19, Elsevier.

Daelemans, Walter (2013), Explanation in computational stylometry, *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, pp. 451–462.

Delfos, Martine France (2011), *Een vreemde wereld: over autismespectrumstoornissen (ASS): voor ouders, partners, hulpverleners, wetenschappers en de mensen zelf*, SWP.

Diehl, Joshua J, Loisa Bennetto, and Edna Carter Young (2006), Story recall and narrative coherence of high-functioning children with autism spectrum disorders, *Journal of abnormal child psychology* **34** (1), pp. 83–98, Springer.

Dockrell, Julie E, Jessie Ricketts, Tony Charman, and Geoff Lindsay (2014), Exploring writing products in students with language impairments and autism spectrum disorders, *Learning and Instruction* **32**, pp. 81–90, Elsevier.

Fay, Warren H and Adriana Luce Schuler (1980), *Emerging language in autistic children*, Vol. 5, Hodder Arnold.

Fombonne, Eric (1999), The epidemiology of autism: a review, *Psychological medicine* **29** (4), pp. 769–786, Cambridge University Press.

Gansle, Kristin A, Amanda M VanDerHeyden, George H Noell, Jennifer L Resetar, and Kashunda L Williams (2006), The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students, *School Psychology Review* **35** (3), pp. 435, National Association of School Psychologists.

Grabowski, Joachim (2010), Speaking, writing, and memory span in children: Output modality affects cognitive performance, *International Journal of Psychology* **45** (1), pp. 28–39, Taylor & Francis.

Hammill, Donald D and Stephen C Larsen (1996), *TOWL-3: Test of written language*, Pro-Ed.

Happé, Francesca and Uta Frith (1996), The neuropsychology of autism, *Brain* **119** (4), pp. 1377–1400, Oxford University Press.

Hirst, Graeme and Vanessa Wei Feng (2012), Changes in style in authors with Alzheimer's disease, *English Studies* **93** (3), pp. 357–370, Taylor & Francis.

Kanner, Leo (1946), Irrelevant and metaphorical language in early infantile autism, *American journal of Psychiatry* **103** (2), pp. 242–246, Am Psychiatric Assoc.

Karmiloff-Smith, Annette (1985), Language and cognitive processes from a developmental perspective, *Language and cognitive processes* **1** (1), pp. 61–85, Taylor & Francis.

King, Diane, Julie Dockrell, and Morag Stuart (2014), Constructing fictional stories: a study of story narratives by children with autistic spectrum disorder, *Research in developmental disabilities* **35** (10), pp. 2438–2449, Elsevier.

Le, Xuan, Ian Lancashire, Graeme Hirst, and Regina Jokel (2011), Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists, *Literary and Linguistic Computing* **26** (4), pp. 435–461, Oxford University Press.

Losh, Molly and Lisa Capps (2003), Narrative ability in high-functioning children with autism or Asperger's syndrome, *Journal of autism and developmental disorders* **33** (3), pp. 239–251, Springer.

Losh, Molly and Lisa Capps (2006), Understanding of emotional experience in autism: insights from the personal accounts of high-functioning children with autism., *Developmental psychology* **42** (5), pp. 809, American Psychological Association.

Losh, Molly and Peter C Gordon (2014), Quantifying narrative ability in autism spectrum disorder: A computational linguistic analysis of narrative coherence, *Journal of autism and developmental disorders* **44** (12), pp. 3016–3025, Springer.

Loveland, Katherine and Belgin Tunali (1993), Narrative language in autism and the theory of mind hypothesis: A wider perspective, *Understanding other minds: Perspectives from autism* pp. 247–266, Oxford University Press Oxford, UK.

Mackie, Clare and Julie E Dockrell (2004), The nature of written language deficits in children with sli, *Journal of Speech, Language, and Hearing Research* **47** (6), pp. 1469–1483, ASHA.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky (2014), The Stanford CoreNLP Natural Language Processing toolkit., *ACL (System Demonstrations)*, pp. 55–60.

Marckx, Silke (2017), *Propositional Idea Density in Patients with Alzheimer's Disease: An Exploratory Study*, Master's thesis, University of Antwerp, Antwerp.

Miltsakaki, Eleni, Rashmi Prasad, Aravind K Joshi, and Bonnie L Webber (2004), The Penn discourse treebank., *LREC*.

Myles, Brenda Smith, Abigail Huggins, Maleia Rome-Lake, Taku Hagiwara, Gena P Barnhill, and Deborah E Griswold (2003), Written language profile of children and youth with asperger syndrome: From research to practice, *Education and Training in Developmental Disabilities* pp. 362–369, JSTOR.

Pedregosa, F., 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, and E. Duchesnay (2011), Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, pp. 2825–2830, Microtome Publishing.

Pennebaker, James W (2011), *The secret life of pronouns*, Bloomsbury Press, New York.

Pennebaker, James W, Martha E Francis, and Roger J Booth (2001), *Linguistic inquiry and word count: LIWC 2001*, Mahwah: Lawrence Erlbaum Associates.

Pronovost, Wilbert, M Phillip Wakstein, and D Joyce Wakstein (1966), A longitudinal study of the speech behavior and language comprehension of fourteen children diagnosed atypical or autistic, *Exceptional children* **33** (1), pp. 19–26, SAGE Publications Sage CA: Los Angeles, CA.

Prud'hommeaux, Emily T, Brian Roark, Lois M Black, and Jan Van Santen (2011), Classification of atypical language in autism, *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, Association for Computational Linguistics, pp. 88–96.

Regneri, Michaela and Diane King (2015), Automatically evaluating atypical language in narratives by children with autistic spectrum disorder, *Natural Language Processing and Cognitive Science: Proceedings 2014* p. 173, Walter de Gruyter GmbH & Co KG.

Regneri, Michaela and Diane King (2016), Automated discourse analysis of narrations by adolescents with autistic spectrum disorder, *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pp. 1–9.

Riley, Kathryn P, David A Snowdon, Mark F Desrosiers, and William R Markesbery (2005), Early life linguistic ability, late life cognitive function, and neuropathology: findings from the nun study, *Neurobiology of aging* **26** (3), pp. 341–347, Elsevier.

Rouhizadeh, Masoud, Emily Prud'Hommeaux, Brian Roark, and Jan Van Santen (2013), Distributional semantic models for the evaluation of disordered language, *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, Vol. 2013, NIH Public Access, p. 709.

Rutter, M, L Mawhood, and P Howlin (1992), Language delay and social development, *in* Fletcher, Paul and David MB Hall, editors, *Specific speech and language disorders in children*, Whurr Publishers, Ltd., pp. 63–78.

Rutter, Michael (1970), Autistic children: infancy to adulthood., *Seminars in psychiatry*, Vol. 2, p. 435.

Shanahan, Timothy (2006), *Relations among oral language, reading, and writing development*, Guilford Press, New York, NY, US, pp. 171–183.

Simonoff, Emily, Andrew Pickles, Tony Charman, Susie Chandler, Tom Loucas, and Gillian Baird (2008), Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample, *Journal of the American Academy of Child & Adolescent Psychiatry* **47** (8), pp. 921–929, Elsevier.

Snowdon, David A, Susan J Kemper, James A Mortimer, Lydia H Greiner, David R Wekstein, and William R Markesbery (1996), Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the nun study, *Jama* **275** (7), pp. 528–532, American Medical Association.

Storoschuk, S, C Lord, and S Jaedicke (1995), Autism and the use of mental verbs, *Biennial Meeting of the Society for Research in Child Development, Indianapolis*.

Tager-Flusberg, Helen (1985), The conceptual basis for referential word meaning in children with autism, *Child Development* pp. 1167–1178, JSTOR.

Tager-Flusberg, Helen (1989), A psycholinguistic perspective on language development in the autistic child, *Autism: Nature, diagnosis, and treatment* pp. 92–115, Guilford Press New York.

Tager-Flusberg, Helen (1992), Autistic children's talk about psychological states: Deficits in the early acquisition of a theory of mind, *Child Development* **63** (1), pp. 161–172, Wiley Online Library.

Tager-Flusberg, Helen (2000), Understanding the language and communicative impairments in autism, *International review of research in mental retardation* **23**, pp. 185–205, Elsevier.

Tager-Flusberg, Helen (2004), Strategies for conducting research on language in autism, *Journal of Autism and Developmental Disorders* **34** (1), pp. 75–80, Springer.

Tager-Flusberg, Helen and Kate Sullivan (1994), A second look at second-order belief attribution in autism, *Journal of Autism and Developmental Disorders* **24** (5), pp. 577–586, Springer.

Tager-Flusberg, Helen and Kate Sullivan (1995), Attributing mental states to story characters: A comparison of narratives produced by autistic and mentally retarded individuals, *Applied Psycholinguistics* **16** (3), pp. 241–256, Cambridge University Press.

Tager-Flusberg, Helen, Rhea Paul, and Catherine Lord (2005), Language and communication in autism, *in* Volkmar, Fred R, Rhea Paul, Ami Klin, and Donald J Cohen, editors, *Handbook of Autism and Pervasive Developmental Disorders, Volume 1, Third Edition*, Wiley Online Library, chapter 12, pp. 335–364.

van den Bosch, Antal, Bertjan Busser, Sander Canisius, and Walter Daelemans (2007), An efficient memory-based morphosyntactic tagger and parser for Dutch, *LOT Occasional Series* **7**, pp. 191–206, LOT, Netherlands Graduate School of Linguistics.

van Halteren, Hans and Maarten op de Weegh (2012), Clues for autism in Dutch tweet production.

Wagner, Richard K, Cynthia S Puranik, Barbara Foorman, Elizabeth Foster, Laura Gehron Wilson, Erika Tschinkel, and Patricia Thatcher Kantor (2011), Modeling the development of written language, *Reading and writing* **24** (2), pp. 203–220, Springer.

Yule, William and Michael Rutter (1987), *Language development and disorders*, MacKeith.

Zijlstra, Hanna, T van Meerveld, Henriet van Middendorp, James Pennebaker, and Rinie Geenen (2004), De nederlandse versie van de Linguistic Inquiry and Word Count (LIWC), een gecomputeriseerd tekstanalyseprogramma, Tijdstroom.