Grammar and Lexicon in Individuals With Autism: A Quantitative Analysis of a Large Italian Corpus

Arjuna Tuzzi

Abstract

Statistical and linguistic procedures were implemented to analyze a large corpus of texts written by 37 individuals with autism and 92 facilitators (without disabilities), producing written conversations by means of PCs. Such texts were compared and contrasted to identify the specific traits of the lexis of the group of individuals with autism and assess to what extent it differed from the lexis of the facilitators. The purpose of this research was to identify specific language features using statistical procedures to analyze contingency lexical tables that reported on the frequencies of words and grammatical categories in different subcorpora and among different writers. The results support the existence of lexis and distributional patterns of grammatical categories that are characteristic of the written production of individuals with autism and that are different from those of facilitators.

DOI: 10.1352/1934-9556-47.5.373

Autism is a well-known, pervasive developmental disorder characterized by early onset, usually during the infant or toddler years. The three essential criteria for diagnosing autism include (a) qualitative impairments in social interaction, (b) delays and deficits in language and communication, and (c) restricted, repetitive, and stereotyped behaviors, activities, and interests (i.e., the "triad" of impairment symptoms; American Psychiatric Association, 2000). Among the primary characteristics of autism, language, and communication impairment have been widely acknowledged and investigated in the relevant literature: Over 60 years ago, Kanner (1943) had already noted severe deficits ranging from mutism to impairments in the learning path and in the productive use of language.

The majority of individuals with autism have little or no functional verbal language. Nevertheless, some children with autism increase their communicative attempts and acquire some functional language (either spoken or sign). Consequently, the group of verbal individuals with autism has been widely investigated by researchers through a variety of methodological approaches, and the study of language in individuals with autism has been focused on those individuals in particular (Tager-Flusberg, 2000).

In the field of language, it is generally acknowledged that one of the richest sources of data is provided by spontaneous speech samples, which may be collected in different contexts (e.g., with a parent, with a teacher, with a researcher) and in either open-ended nonstructured or standardized structured approaches. Because of the core deficits and high rates of echolalia, it is difficult for individuals with autism to provide an adequate natural language sample in a context of spontaneous verbal conversational interaction (issues related to spontaneous communication in individuals with autism and approaches to the conceptualization of communicative spontaneity were recently discussed by Chiang & Carter, 2007). Some researchers have questioned whether standardized tests can be used to describe language functioning in individuals with autism. The relationship between the standardized measurement of language and measurement of spontaneous speech is discussed in Condouris, Meyer, and Tager-Flusberg (2003). In this study, natural language samples were collected (through transcription) from 44 children with autism while they interacted with one of their parents. The corpus was collected by selecting a subcorpus for each child of 100 (in some cases less) consecutive,

complete, and intelligible utterances. The children in this study performed lower than age expectations in all measurements and were more impaired compared with reference data in spontaneous speech measurements than in standardized tests. Nevertheless, the correlation between standardized test scores and spontaneous speech measures proved highly significant.

The need for additional research on language was stressed in a recent study by Perkins, Dobbinson, Boucher, Bol, and Bloom (2006) that focused on spoken conversations from a corpus of 70,000 tokens and 7 individuals with autism. This study also highlights the need for more research into the vocabularies of individuals with autism and their individual word use. Similar to Tager-Flusberg's (1985) study, the results in this study also pointed to an imperfect relation between lexical–conceptual understanding and lexical use.

Other studies on verbal language have focused on deficit aspects or nontypical modes of communication in individuals diagnosed with Asperger syndrome or high-functioning autism (Dennis, Lazenby & Lockyer, 2001; Dobbinson, Perkins & Boucher, 2003; Landa & Goldberg, 2005; Martin & McDonald, 2004; Stone & Yoder, 2001). For example, Ghaziuddin et al. (2000) found that individuals with Asperger syndrome showed more complex speech patterns compared with control participants with high-functioning autism. Moreover, grammatical and syntactic features have been found to be useful in differentiating between autism and Asperger syndrome (Bennet et al., 2008).

The study of language in individuals with autism proves a difficult task in general and in nonverbal individuals in particular. The available literature provides important results concerning individuals with autism who are verbal, but a number of more general questions will remain unanswered if methods are not identified to study spontaneous language in individuals with autism without resorting to the collection of spoken language samples.

In noting the most important challenges for future research, Tager-Flusberg (2000) noted the need to use a multimethod perspective, because research on language in autism requires the development of new approaches and new integrated methods (and paradigms) that have developed independently in the field of psychology (and other fields). The *EASIEST* University Project is an Italian interdisciplinary research program involving linguistics, neuropsychiatry, psychology, sociology, statistics, and computer-aided text processing (Bernardi, 2008) that analyzes distinctive features in written language of individuals with autism. In this study I address only the section of the project focused on the quantitative analysis of the data that came from statistics applied to corpus linguistics. In this study I (a) focused on gaining insight into the characteristics of the written language of individuals with autism, (b) proposed an approach based on quantitative analysis typical of statistical analysis of textual data, (c) based the analysis on a large corpus, and (d) adopted a comparative perspective (the texts written by individuals with autism are compared and contrasted with the texts written by facilitators) based on linguistic and lexical data. In brief, in the present study I tried to identify distinctive linguistic and grammatical features in texts written by individuals with autism and by individuals without disabilities in a context of spontaneous written conversation.

To date, few studies have dealt with the lexis and grammar of written language of individuals with autism (Niemi & Kärnä-Lin, 2002, 2003; Saloviita & Sariola, 2003); more important, none of available studies used large collections of texts exceeding several 100,000 tokens. The lack of quantitative analyses of large collections of text is striking also in light of the development of methods such as the computer-aided analysis of corpora (Alexa & Zuell, 2000), corpus statistics (Baayen, 2000; Cortelazzo & Tuzzi, 2008; Oakes, 1998; Strauss, Fan, & Altmann, 2008), text mining (Bolasco, Canzonetti, & Capo, 2005; Sirmakessis, 2004; Sullivan, 2001), and statistical analysis of content (Lebart, Salem, & Berry, 1998; Tuzzi, 2003). However, this may be considered self-explanatory because much of the available literature dealing with language in autism is based on research that was not conducted on large corpora.

One way to collect large corpora of written language is using the material produced by individuals with autism during sessions of facilitated communication (FC), which is a well-known form of augmentative and alternative communication. FC training entails learning to communicate by typing on a keyboard and involves a combination of physical and emotional support to an individual who has difficulties with speech and with intentional pointing. The person who provides

support is called a facilitator. Texts written during FC sessions may also benefit from the possibilities offered by computer-based technologies.

All the individuals in this research communicated through FC, which inevitably involves the issue of authorship attribution (for a Finnish case study, see Niemi & Kärnä-Lin, 2002; for an Italian case study, see Scopesi, Zanobini, & Cresci, 2003). Although a complex and lively debate is taking place on the FC technique, in this study I did not take into account the scientific controversy that developed in the 1990s on the validity of FC (Biklen & Cardinal, 1997; Cardinal, Hanson, & Wakeham, 1996; Green, 1994; Salomon, Wagner, & Bauman, 1996), was recently revived (Duchan, Calculator, Sonnenmeier, Diehl, & Cumley, 2001; Emerson, Grayson, & Griffiths, 2001; Green, 2005; Mostert, 2002; Smukler, 2005; Wurzburg, 2004), and was left unsettled eventually. To date, the participants in the controversy have not yet agreed on a validation method capable of settling the dispute. For a review, see Jacobson, Mulick, and Schwartz (1995) and Mostert (2001).

In the light of this controversy and of the need for dealing with the issue of authorship attribution (i.e., some studies reported above claim that the texts written by individuals with autism may actually have been written by their facilitators), in the present study, I analyzed two subcorpora of texts written during sessions of FC to check whether distinctive features emerged such as to clearly differentiate between two different modes of communication.

Method

Participants and the Literature Reviewed

The texts considered in this study were produced by 92 facilitators (individuals without disabilities) and 37 individuals who were diagnosed with autism and whose verbal communication was fully absent or greatly impaired (Table 1). Each participant had been diagnosed with autism by neuropsychiatrists in the four accredited FC centers involved in the project and assessed according to criteria from the *Diagnostic and Statistical Manual of Mental Disorders, Text Revision* (*DSM-IV-TR*; American Psychiatric Association, 2000). The sample did not include individuals diagnosed with Asperger syndrome or those who were classified as having high-functioning autism.

Table 1 Distribution of Study Variables for Individuals With Autism Included in the Corpus

Variable	*n*	%
FC center		
Padua	9	24.3
Genoa	9	24.3
Rome	10	27.0
Bari	9	24.3
Gender		
M	29	78.4
F	8	21.6
Age (years)		
Up to 10	3	8.1
11 to 15	11	29.7
16 to 20	11	29.7
21 to 25	7	18.9
Over 25	5	13.5
Start FC (years)		
Up to 7	13	35.1
8 to 15	18	48.6
Over 15	6	16.2
FC training (years)		
Up to 5	10	27.0
6 to 10	23	62.2
Over 10	4	10.8

Note. For all variables, $N = 37$, % = 100%. FC = facilitated communication.

The 37 individuals with autism selected for the research project were assisted in four accredited FC centers and included 29 males and 8 females; their ages at the beginning of the study ranged between 9 and 32 years, with 59.4% of the individuals with autism being between 11 and 20 years old. The majority started practicing FC by the age of 15 (84.8%) and 35.1% by the age of 7. The four accredited FC centers selected the texts of the FC sessions for this study according to a protocol.

The individuals with autism were selected according to the degree of autonomy they had achieved in written communication: Only individuals capable of writing through light facilitation were chosen. *Light facilitation* means that the support provided by facilitators was limited to the contact between the facilitator's hand and the individuals' arm, shoulder, neck, head, back, or leg,

and, in some cases, contact was merely intermittent or occasional (depending on the habit developed by the individual over the years spent practicing with the facilitators). All the individuals with autism involved in this project had reached a high degree of self-sufficiency in written FC.

The selection of both the individuals with autism and their facilitators was based on the degree of familiarity achieved with the writing technique: Only individuals with autism were selected who had undergone a long FC training (they had all been practicing FC for several years: 62.2% for 6–10 years and 10.8% for over 10 years), and all facilitators were professionals (teachers and supervisors) and parents who were specifically trained in this technique. All the individuals with autism involved in this project interacted with at least three different facilitators (typically a professional facilitator, one of their parents, and one of their teachers). Different facilitators may have alternated in different sessions, but each session had only one facilitator. FC sessions took place in different locations (at home, at school, or in FC centers). The corpus totaled over 2,000 sessions complying with the selections criteria.

It is worth mentioning that the FC sessions envisaged real, written, open-ended, nonstructured, nonstandardized, noncompulsory conversations between an individual with autism and a facilitator. They included dialogues, reflections, statements, and observations dealing with a topic proposed by the facilitator or the individual with autism. Text types include written conversations on issues ranging from daily business to private matters, conversations on school subjects, and creative writing (e.g., essays and poems). Conversations were partly educational and partly aimed at communicating and exchanging information. FC sessions contribute to the development of a communication channel for persons whose abilities in terms of spoken communication are absent or severely impaired; for a number of years, these individuals with autism have been using this communication channel with teachers at school, with their parents at home, and so forth.

Because all of the individuals with autism had practiced FC for several years, their conversations achieved satisfactory levels in terms of length and complexity. To minimize the impact of question–answer sequences on the texts, all structured and operational passages composed of multiple-choice questions or short answers were excluded (e.g., multiple-choice tests held in the school environment, mathematics tests, and yes–no sequences were excluded). In addition, the lemmatization of Italian texts allows eliminating many morphological variations needed for the correct syntactic structuring of question–answer sequences (e.g., verb inflections mirroring the presence of six different grammatical persons).

Because FC uses PC text editors, the transcription of the conversations held by the individuals with autism and their facilitators was directly available in electronic format. The conventional choices made when writing with the computers allowed me to distinguish between the parts written by individuals with autism and those written by facilitators: During an FC session, the facilitator wrote in capital letters, whereas the individual with autism wrote in lowercase letters. These conventions and the data identifying individual writers made it possible to distinguish between the subcorpora referring to either group and individual subcorpora. The text produced during a FC session was matched with a set of data relating to the control variables envisaged by the protocol (e.g., age, gender, education of both the facilitator and the individual with autism, place where the session was held, duration, facilitation level). Text data were processed by means of the TaLTaC[2] dedicated software (Bolasco, Baiocchi, & Morrone, 2008), and the statistical analysis was conducted by means of R (R Development Core Team, 2008).

Some of the texts produced during the FC sessions were collected during the 2-year *EASIEST* Project; other texts were collected over several years and retrieved from archives. This implies that each individual with autism participated in sessions at different ages and different levels of ability in terms of FC training. The protocol implemented determined the choice of texts produced during the sessions, although it also allowed saving time when previous sessions had to be consulted. From this viewpoint, the collection of data was controlled expost and not during the production stage. The age at when the session was held, the training time (how long participants had been practicing FC), the school year attended by the individual with autism, the topics dealt with, and personal life experiences all played a role in the *EASIEST* Project but were dealt with in a different corpus and during a distinct stage of the research (not illustrated in the present article). Those variables were not accounted for in the present analysis because the total text passages

classified according to individual–age–training were too limited in size and did not allow conducting a statistical analysis of textual data.

The goal of this specific stage of the research project was to collect a large corpus of written language that could prove useful in the identification of two communication modes, two language models, and two lexica, by comparing and contrasting two large representative corpora. This research could benefit from both the possibility of working with large corpora and that it included a number of "expert" individuals with autism—the participants with autism had been using FC for a long time, with different facilitators and with a high level of autonomy. In this sense, the corpus proves very interesting from a lexicostatistical perspective but complies with the logic underlying the development of corpora (including marked redundancies) rather than the rationale of random sampling.

In lexical statistics the population is provided by *language* (in the sense of de Saussure's *langue* (cf. Cortelazzo & Tuzzi, 2008; Herdan, 1956; de Saussure, 1968): Statistical units correspond to words, the sample is the *corpus* (i.e., the "observed corpus", because it is an instantiation, a sample of de Saussure's *parole*), and the sample size is measured by *length* (i.e., the number of words in the corpus). From a statistical viewpoint, language is comparable with a theoretical population that is unlimited, nonobservable, and unstable in time and space (Cortelazzo & Tuzzi, 2008; Strauss et al., 2008); consequently, its analysis is a difficult task because words cannot be associated to probability in the statistical sense. Moreover, the sample size determines and affects text measures (e.g., typical text constants and parameters for many word-frequency distribution models). These distinctive features distinguish lexical statistics from most other fields of statistics because an increase in the sample size leads to systematic changes in measures and parameters instead of enhancing the accuracy of estimations (Baayen, 2000). Only empirical measures based on sample size (a *large* corpus must exceed 100,000 words) and lexical richness (Bolasco, 1999) may be used to understand whether a corpus is sufficiently large to allow the implementation of statistical methods.

Procedures

From a computational point of view, the corpus is composed of words (token forms) that are sequences of letters taken from the alphabet and isolated by means of separators: blanks and punctuation marks. A word token (or token) is a particular occurrence of a word type (or type) in a text. A token instantiates a type (e.g., the single word type *the* has many tokens in any English text), but there are also many word types that occur only once in a given corpus. If a word type has only one word token, the item is called *hapax legomena* (or hapax). The number (N) of word tokens is the corpus size in terms of total occurrences. The number (V) of word types measures the size in terms of different words and provides a rough measure of lexical richness. The list of word types and the relevant frequencies is the vocabulary of the corpus.

Because it was conducted on texts written in Italian, the token-form analysis is strongly limited, owing to the contingent nature of some lexical choices (e.g., tenses, plural forms, masculine and feminine forms), which do not depend on the individuals' lexical richness. For example, if the facilitator often asks questions in the second-person singular (e.g., "Would you like…?", i.e., "vorresti?" in Italian) and the individual with autism answers in the first-person singular ("I would like…"/ "vorrei…"), there is a distortion that may systematically magnify the differences between the two subcorpora. To overcome that limit, a transition is needed from the form analysis to the lemma analysis. The lemmatization process associates each token form with a pair including a lemma and a grammatical category (e.g., in English, the token form *thought* is associated with either the lemma *to think* and grammatical category *verb* or the lemma *thought* and category *noun*). In some cases, the same token form leads to different lemma types (as is the case with *thought*) and the number of different lemmas increases, thus reducing ambiguity. In other cases different token forms are associated to the same pair (e.g., in the case of *tooth* and *teeth*, which are both associated with the lemma *tooth* and category *noun*), and the number of different lemmas decreases, thus leading to noise reduction. Normally, the second effect prevails and the number of types decreases. It should be remembered that, more than in other languages, lemmatization plays a major role in Italian in reducing noise and increasing the amount of information conveyed by each lemma type. In addition, owing to the wide range of contingent variations (e.g., masculine, feminine and plural forms, six different persons, verb conjugations, clitic pronouns) in Italian, the

transition from a token-form vocabulary to a lemma-form vocabulary may halve the number of different types.

The software tools currently available for the Italian language do not allow the full and correct lemmatization of large corpora through a fully automated process. Lemmatization was conducted on the corpus through a partly manual and partly automatic process. Because of the need for manually disambiguating numerous lemmas by checking the context of occurrence, the lemmatization was not completed fully and the presence of residual ambiguous tokens was tolerated. Because these tokens mostly included grammatical words used and spread throughout all texts, a comparative study is reasonable (out of a total of over 11,780 lemma types, the analysis produced only 133 lexically ambiguous lemma types, 75 grammatically ambiguous lemma types, and 543 hapaxes produced by errors).

The list of lemma types and their frequencies provides the lexicon of the corpus and mirrors the lexical ranges of the writers. Following the lemmatization, the number of word tokens remains unchanged, whereas the resulting number (L) of lemma types is smaller than the number (V) of form types, and the statistics based on L provide a more reliable measure of lexical richness compared with statistics based on V. Moreover, the lemmatization also allows analysis of the distribution of tokens among grammatical categories.

Measures

The corpus was a collection of words, clauses, sentences, and texts subdivided according to grouping criteria: For example, one subcorpus was composed of texts written by the individuals with autism and the other of texts written by facilitators (a dichotomic variable was useful when comparing and contrasting the two subcorpora); an alternative would be arranging the same number of subcorpora and individuals, with each subcorpus being composed of texts written by the same "author" (a politomic variable is useful when comparing individual-level data). Both alternatives are considered in this study.

The approach in this study was mainly quantitative—it did not dwell on the qualitative aspects of language (e.g., semantics, rhetoric, style) and did not consider the psychological, educational, and pedagogic implications of the results.

The lexis of both subcorpora was compared and contrasted by assessing the overlap between both lexicons and computing quantitative indicators: the type–token ratio (TTR), obtained by dividing the number of word types (form types or lemma types) by the total number of word tokens; the hapax percentage, given by the ratio of the number of words (forms or lemmas) appearing only once to the number of word types (form types or lemma types); and the mean frequency of types (mfreq), calculated as the tokens–type ratio (the reciprocal of TTR). A test on proportions was implemented to assess the differences between subcorpora. From the quantitative assessment of lexical richness, such data are useful for comparative purposes because, as shown by Tweedie and Baayen (1998), typical text constants (e.g. measures of lexical richness) become reliable only with large corpora (roughly exceeding 50,000 occurrences).

To assess the role played by grammatical choices, the test on proportions was implemented again to measure the differences of percentage values between subcorpora with reference to the main grammatical categories: nouns, verbs, adjectives, adverbs, and grammatical words. The residual category labeled *other* included proper names, numbers, foreign words, exclamation words, lexical errors, and ambiguities. These comparisons between the two subcorpora were conducted on the whole corpus.

At the individual level, lexical richness was calculated as the number of different words (NDW; Duràn, Malvern, Richards, & Chipere, 2004; McKee, Malvern, & Richards, 2000; Watkins, Kelly, & Harbers, 1995) starting from text chunks, including 1,000 tokens each. More specifically, in the present study the calculation involved lemmas (i.e., the number of different lemmas [NDL]).

To analyze the distribution of grammatical categories among individuals, the percentage values were calculated for each individual with autism and each facilitator and transformed into Z scores. Individual similarities and differences are highlighted on the graph in Figure 1 by the mutual position of individuals with autism (stars) and facilitators (squares) in terms of scores. The differences between the two groups of individuals were also assessed by means of t tests with reference to the main grammatical categories. During this second stage, only individuals who had produced subcorpora including at least 1,000 tokens were considered.

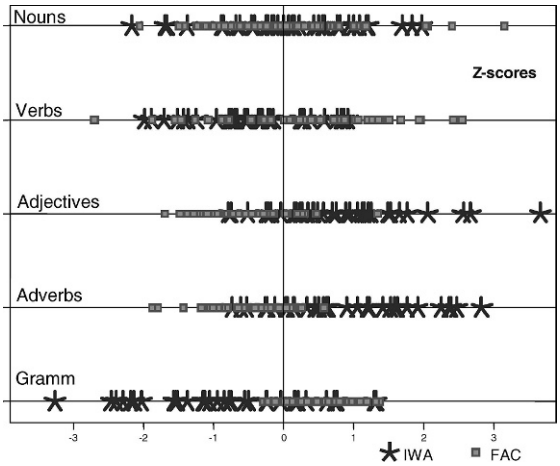

Figure 1 Z scores illustrating grammatical categories distributions among individuals with autism (IWA; denoted by stars; $n = 37$) and facilitators (FAC; denoted by squares; $n = 54$).

Results

The corpus was almost 900 pages long and included slightly less than 300,000 words. Because of its size, studies could be conducted based on textual statistics techniques, the efficiency of which increases according to the size of the corpus available.

The texts of the corpus were subdivided into two subcorpora (Table 2): FAC, including texts produced by facilitators, and IWA, including texts produced by individuals with autism. The two subcorpora were well balanced in terms of the size. Within the corpus, 290,496 word tokens were subdivided into 159,243 tokens assigned to the facilitators and 131,253 tokens assigned to the individuals with autism. Consequently, the total subcorpus produced by the individuals with autism was only slightly smaller than that produced by their facilitators. If the number of word types *V* (and lemma types *L*) contained in both subcorpora were considered instead of that of word tokens, the ratio changed in favor of the individuals with autism. Their subcorpus contained more types and more hapaxes in terms of both forms and lemmas.

Table 3 Comparison of Lexical Items (Lemma Types) Between Subcorpora (FAC Versus IWA)

	Lemmas				
Subcorpora	(L)	Exclusive	%	Common share	%
FAC	7,298	2,308	31.6	4,990	68.4
IWA	9,472	4,482	47.3	4,990	52.7

Note. FAC = texts produced by facilitators; IWA = texts produced by individuals with autism.

The lemmatization process and measurement in terms of different lemma types also allowed understanding to what extent the subcorpora shared the same lexis and evaluating the relative weight of the common share (Table 3). The lexis of individuals with autism overlapped only partially with that of facilitators and was generally broader. The incidence of the shared lemmas was relatively greater in the case of facilitators compared with individuals with autism (Table 4).

With reference to the token distribution among grammatical categories, the subcorpus produced by individuals with autism had a greater percentage of adjectives and adverbs and a smaller percentage of verbs and grammatical words than the subcorpus produced by facilitators. The analysis did not show highly significant differences in the use of nouns (Table 5).

The texts of the corpus were also subdivided into 129 (37 + 92) subcorpora referring to individual writers. The 37 individuals with autism

Table 2 Size of Subcorpora (FAC Versus IWA) for Word Tokens and Word Types (Forms and Lemmas)

		Forms (V)					Lemmas (L)				
Subcorpora	Tokens (*N*)	V	TTR %	Hapax	Hapax %	mfreq	L	TTR%	Hapax	Hapax %	mfreq
FAC	159,243	12,359	7.8	6,301	51.0	12.9	7,298	4.6	3,212	44.0	21.8
IWA	131,253	14,875	11.3	8,373	56.3	8.8	9,472	7.2	4,692	49.5	13.9
Corpus	290,496	20,166	6.9	10,055	49.9	14.4	11,780	4.1	5,053	42.9	24.7

Note. FAC = texts produced by facilitators; IWA = texts produced by individuals with autism; TTR = type-token ratio; Hapax = word type with only one word token; mfreq = mean frequency of types.

Table 4 Comparison of Hapax Percentage and Common Share Between Subcorpora (FAC Versus IWA)

Hapax type	Proportion test (*Z*)	*p*
Hapax (forms)	−8.751	***
Hapax (lemmas)	−7.104	***
Common share (lemmas)	20.526	***

Note. FAC = texts produced by facilitators; IWA = texts produced by individuals with autism.
***$p < .001$

produced texts comparable in length, ranging from 1,162 word tokens at a minimum, to 8,587 maximum word tokens, with a mean value of 3,547 word tokens each. In terms of the 92 facilitators, text length varied to a greater extent, ranging from a minimum of 100 word tokens to over 13,000, with a mean value of 1,731.

During this stage, only the individuals who produced subcorpora including at least 1,000 tokens were considered (i.e., all 37 individuals with autism and 54 facilitators out of 92). To compare and contrast lexical richness, the NDL of samples of 1,000 tokens was calculated for each individual. The results showing a greater amount of different words and hapaxes in the group including individuals with autism compared with the facilitators were confirmed by the test results (Table 6) and illustrated in the box plots of Figure 2. In addition, individuals with autism showed a richer vocabulary.

To observe the distribution of grammatical categories among individuals in terms of Z scores, similarities and differences are highlighted by the mutual position of individuals with autism (stars) and facilitators (squares) on the graph in Figure 1. Moreover, *t* tests were carried out on individual-level percentages. Only individual subcorpora including at least 1,000 tokens were considered (Table 7). The grammatical analysis at the individual level produced the same results obtained by comparing and contrasting the two subcorpora: The individuals with autism used more adjectives and adverbs and less verbs and grammatical words compared with their facilitators. In addition, in this case, no significant differences emerged in the use of nouns. Additional tests (e.g., nonparametric tests) produced the same results.

Discussion

Research is still at an early stage in this area, and caution should be taken in interpreting these data because the available corpus was wide, the texts covered a broad range of topics, and the participants differed significantly (e.g., in terms of gender, age, experiences, habits). However, tentative conclusions may be drawn because results were achieved that do not agree with the results of other studies.

Lexical Richness

In the light of this comparative analysis of a lemmatized large corpus in Italian, the written production of individuals with autism showed higher lexical richness compared with written production of facilitators. Comparing and contrast-

Table 5 Distribution of the Tokens Among Grammatical Categories Between Subcorpora (FAC Versus IWA)

	FAC		IWA			
Category	*N*	%	*N*	%	Proportion test (*Z*)	*p*
Nouns	26,070	16.4	21,239	16.2	1.367	
Verbs	40,881	25.7	30,806	23.5	13.684	***
Adjectives	7,507	4.7	10,064	7.7	−33.238	***
Adverbs	8,213	5.2	11,883	9.1	−41.189	***
Grammatical words	66,610	41.8	48,628	37.0	26.191	***
Other	9,975	6.3	8,633	6.6		
Total	159,256	100.0	131,253	100.0		

Note. FAC = texts produced by facilitators; IWA = texts produced by individuals with autism.
***$p < .001$.

Table 6 Comparison of Lexical Richness (NDL and No. of Hapaxes Per 1,000 Tokens) in Individuals With Autism (n = 37) and in Facilitators (n = 54)

	FAC (n = 54)			IWA (n = 37)				
Variable	*M*	*SD*	Range	*M*	*SD*	Range	*t*	*p*
NDL	326.5	42.9	213–461	378.2	55.3	264–479	−5,019	***
Hapax	192.3	41.7	100–336	250.3	52.7	137–346	−5.852	***

Note. NDL = number of different lemmas; FAC = texts produced by facilitators; IWA = texts produced by individuals with autism.
***$p < .001$.

ing the lists of lemma types with standard frequency lexicons might yield an explanation (i.e., reference models for the Italian language). The number of lemmas that did not belong to the models was relatively higher in the case of individuals with autism compared with the facilitators, which indicates a greater incidence of unusual words in the subcorpus produced by individuals with autism. From a qualitative viewpoint, lexical richness also resulted from both the use of words not attributable to the common lexis of Italian in both ordinary and unusual contexts and the emergence of a certain amount of neologisms, mostly hapaxes, coined according to the word-formation rules of the Italian language.

The higher lexical richness observed in this study remains difficult to explain and needs further research because language development is often limited in individuals with autism, and their difficulties in learning new words has been described in detail (Frith, 1989). In contrast, this research provides a more thorough qualitative analysis showing the use of high-register words by individuals with autism (e.g., "to emulate/emulare," "to dare/ardire," "to yearn/anelare," "to oppress/vessare," "to soothe/lenire") in adults as well as in children younger than 10 years of age ("to distress/angustiare," "to be eager/bramare," "to elude/eludere").

Grammatical Categories

All grammatical categories (e.g., nouns, verbs, adjectives, adverbs) were present in this study (i.e., the texts produced by individuals with autism contained tokens referring to all the grammatical categories). From a grammatical viewpoint, how-

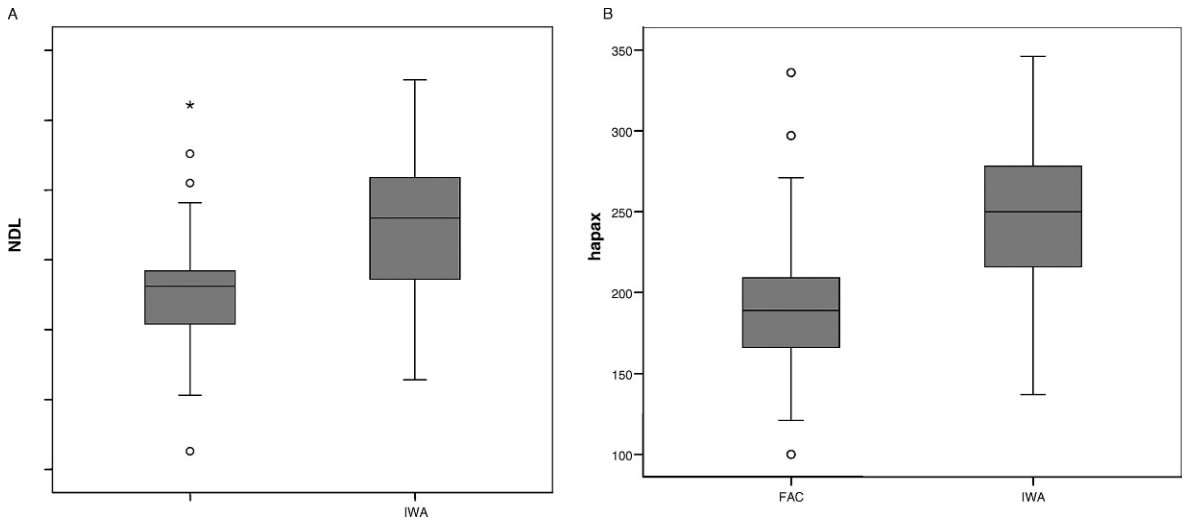

Figure 2 Box plot illustrating (A) the distribution of number of different lemmas (NDL) and (B) number of hapaxes per 1,000 tokens in individuals with autism (IWA; n = 37) and facilitators (FAC; n = 54).

Table 7 Distribution of Tokens Among Grammatical Categories in Individuals With Autism ($n = 37$) and Facilitators ($n = 54$)

	FAC ($n = 54$)			IWA ($n = 37$)				
Categories	*M*	*SD*	Range	*M*	*SD*	Range	*t*	*p*
Nouns	16.3	2.9	10.7–25.3	16.6	2.7	10.4–22.0	−0.436	
Verbs	25.6	3.3	16.2–32.7	23.3	2.4	18.4–27.5	3.572	***
Adjectives	4.7	1.3	2.1–9.1	7.9	2.2	4.2–14.5	−8.535	***
Adverbs	5.3	1.3	1.8–8.4	9.2	2.6	4.9–14.6	−9.502	***
Grammatical words	41.6	1.9	38.2–45.5	36.2	5.0	25.2–42.5	7.189	***

Note. FAC = texts produced by facilitators; IWA = texts produced by individuals with autism.
***$p < .001$

ever, the written production of individuals with autism showed a different token distribution among grammatical categories.

The most obvious difference compared with the texts written by facilitators was a greater share of adjectives and adverbs and the tendency to decrease verbs and omit grammatical words. Albeit less clear, the resulting sentences remained fully understandable. The individuals with autism showed a preference for linguistic resources that qualified objects (i.e., adjectives) and actions (i.e., adverbs). Both adjectives and adverbs were modifiers because they added information modifying the meaning of the nouns or verbs to which they referred, by making it more or less specific (e.g., a *black* cat, a *cunning* cat; to love *very much*, to love *intensely*).

However, the function and rationale behind the high frequency of modifiers in the language of individuals with autism remain difficult to explain. Although this is worth investigating and could lead to a number of speculations on the frequent use of modifiers (both adjectives and verbs but also including high-register choices), a more complex issue remains unexplained: Compared with the studies highlighting lexical poverty, the present data point to unexpected communicative skills in individuals with autism.

Authorship Attribution

With reference to the debate on authorship attribution, the results of these analyses supported the hypothesis of the existence of very specific lexis in the written production of individuals with autism. Consequently, the results supported the authenticity of the texts produced by the individuals with autism in this sample, contributing positive support to the ongoing controversy involving FC.

The differences in terms of grammatical choices and typical lexical patterns of the individuals with autism made it plausible to attribute the texts to them. If the facilitators could have influenced the writers with autism, even by suggesting topics and stylistic solutions, the language of individuals with autism would have overlapped with the language of facilitators; the analysis seems to prove the opposite. In the interactive context of the FC sessions, the facilitators would have needed to be able to switch their communication registers during the transition from a conversational turn to be attributed to them and the following conversational turn to be attributed to the individual with autism. As already explained by Niemi and Kärnä-Lin (2003) in the case of Mr. Alatalo and in the light of the new results, the hypothesis that a number of different facilitators manage to imitate such a specific style is difficult to support.

About the Role of FC

Issues calling for further investigation involve if, how, and to what degree such data can be related to the specific communicative situation deriving from FC. Studies are available that were inspired by this line of research: For example, an experiment on pragmatic abilities in children with autism showed that methods that focus on the communicative task improve the performance and suggested that FC allows individuals with autism to overcome the attention deficits (Bara, Bucciarelli, & Colle, 2001). It should be remembered that over 20 years ago Bernard-Opitz (1982) had already systematically studied the effect of the context on the

frequency and quality of communicative acts in children with autism and found that in a highly familiar setting with highly familiar interlocutors, these children talked more and produced more advanced language than in other social contexts. In the opinion of Tager-Flusberg (2000), this means that laboratory-based studies may not always provide the best instrument to assess linguistic skills. It may be the case that, because of a long training period, FC determines a more favorable and less threatening context of interaction, such as to encourage individuals with autism to use more elaborate and effective language, at least in the written mode.

Limitations of the Study and Future Work

From the viewpoint of corpus linguistics, any conclusion about lexical richness and distribution among grammatical categories becomes less clear-cut when working with small corpora. This is the main reason for the collection of large corpora, which include numerous sessions that produce a broad range of texts yet do not allow detail. Because in this study the sessions were recorded over a period of years for the individuals involved, variables such as age, length of FC training, and so forth, could not be accounted for (e.g., according to a multilevel analysis approach).

For the purpose of this study, texts written by individuals with autism were compared and contrasted with texts written by facilitators to contribute to the ongoing debate on authorship attribution. However, the comparison group was not ideal for the purpose of an overall comparison between the language of individuals with autism and the language of individuals without disabilities. Two additional corpora are available, both from the *EASIEST* research group. In the first control-case corpus, 6 individuals with autism and 6 individuals without disabilities wrote a text on the same topic ("Write about a moment/fact which was important to you") by means of the same computer, the same facilitator, and the same contact on their arms/shoulders. In this situation, the facilitator's interventions were limited to those strictly necessary (i.e., title, task, and greetings at the beginning/end of the session). The second corpus was collected for historical purposes and included a limited number of individuals who have participated in FC sessions for long periods of time (i.e., 3 years of practice). Those two corpora can provide the material for a future study on the impact of age, the skills developed during FC training–practice, and other factors linked to individual characteristics, life histories, and text contents.

Further data analyses (e.g., discriminant analysis, correspondence analysis, cluster analysis) are necessary to understand what factors contribute to the description of written language. The linguistic–statistical method implemented in this study appeared effective in highlighting differences among the groups and would be suitable for an interdisciplinary approach to the study of language in autism. From a more qualitative viewpoint, there is the opportunity to address semantic aspects and move on from the analysis of formal linguistic aspects to the study of contents. The data produced by the present study may be considered preliminary, but the resulting trend can be confirmed by other studies and from different perspectives—by the EASIEST research team, in a volume edited by Bernardi (2008), as well as by other ongoing research.

References

Alexa, M., & Zuell, C. (2000). Text analysis software: Commonalities, differences and limitations: The results of a review. *Quality & Quantity*, *34*, 299–321.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders, text revision* (4th ed.). Washington, DC: Author.

Baayen, H. R. (2000). *Word frequency distributions. Exploring quantitative aspects of lexical structure*. Dordrecht, The Netherlands: Kluwer.

Bara, B. G., Bucciarelli, M., & Colle, L. (2001). Communicative abilities in autism: Evidence for attentional deficits. *Brain and Language*, *77*, 216–240.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Bennet, T., Szatmari, P., Bryson, S., Volden, J., Zwaigenbaum, L., Vaccarella, L., Duku, E., & Boyle, M. (2008). Differentiating autism and Asperger syndrome on the basis of language delay or impairment. *Journal of Autism and Developmental Disorders*, *38*, 616–625.

Bernardi, L. (Ed.). (2008). *Il delta dei significati* [*The delta of meanings*]. Rome: Carocci.

Biklen, D. (Ed.). (2005). *Autism and the myth of the person alone*. New York: New York University Press.

Biklen, D., & Cardinal, D. (1997). *Contested words, contested science*. New York: Teachers College Press.

Bolasco, S. (1999). *Analisi multidimensionale dei dati* [*Multidimensional data analysis*]. Rome: Carocci.

Bolasco, S., Baiocchi, F., & Morrone, A. (2008). TaLTaC^2: Trattamento automatico Lessicale e Testuale per l'analisi del Contenuto di un Corpus (Version 2.8.0.2) [Software] [Lexical and textual automatic treatment for content analysis of a corpus]. Rome: Author. Retrieved October 27, 2008, from http://www.taltac.it

Bolasco, S., Canzonetti, A., & Capo, F. M. (Eds.). (2005). *Text mining*. Rome: Centro Informazione e Stampa Universitaria.

Cardinal, D. N., Hanson, D., & Wakeham, J. (1996). investigation of authorship in facilitated communication. *Mental Retardation*, *34*, 231–242.

Chiang, H.-M., & Carter, M. (2007). Spontaneity of communication in individuals with autism. *Journal of Autism and Developmental Disorders*, *38*, 693–705.

Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003), The relationship between standardized measures of language and measures of spontaneous speech in children with autism. *American Journal of Speech and Language Pathology*, *12*, 349–358.

Cortelazzo, M., & Tuzzi, A. (2008). *Metodi statistici applicati all'italiano* [*Statistical methods applied to Italian*]. Bologna, Italy: Zanichelli.

Dennis, M., Lazenby, A. L., & Lockyer, L. (2001). inferential language in high-function children with autism. *Journal of Autism and Developmental Disorders*, *31*, 47–54.

de Saussure, F. (1968). *Cours de linguistique généralle* [*Course of general linguistics*]. Paris: Payot.

Dobbinson, S., Perkins, M. R., & Boucher, J. (2003). The interactional significance of formulas in adult autistic language. *Clinical Linguistics and Phonetics*, *17*, 299–307.

Duchan, J., Calculator, S., Sonnenmeier, R., Diehl, S., & Cumley, G. (2001). A framework for managing controversial practices. *Language Speech and Hearing Services in Schools*, *32*, 133–141.

Duràn, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*, 220–242.

Emerson, A., Grayson, A., & Griffiths, A. (2001). Can't or won't? Evidence relating to authorship in facilitated communication. *International Journal of Language & Communication Disorders*, *36*(Suppl.), 98–103.

Frith, U. (1989). *Autism: Explaining the enigma*. Oxford, England: Blackwell.

Ghaziuddin, M., Thomas, P., Napier, E., Kearney, G., Tsai, L., Welch, K., & Fraser, W. (2000). Brief report: Brief syntactic analysis in Asperger syndrome: A preliminary study. *Journal of Autism and Developmental Disorders*, *30*, 67–70.

Green, G. (1994). Facilitated communication: Mental miracle or sleight of hand? *Skeptic*, *2*, 68–76.

Herdan, G. (1956). *Language as choice and chance*. Groningen, The Netherlands: Noordhoff.

Jacobson, J. W., Mulick, J. A., & Schwartz, A. A. (1995). A history of facilitated communication: Science, pseudoscience, and antiscience: Science Working Group on Facilitated Communication. *American Psychologist*, *50*, 750–765.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, *2*, 217–250.

Landa, R. J., & Goldberg, M. C. (2005). Language, social, and executive functions in high functioning autism: A continuum of performance. *Journal of Autism and Developmental Disorders*, *35*, 557–573.

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring textual data*. Boston: Kluwer.

Martin, I., & McDonald, S. (2004). An exploration of causes of non-literal language problems in individuals with Asperger syndrome. *Journal of Autism and Developmental Disorders*, *34*, 311–328.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*, 323–337.

Mostert, M. P. (2001). Facilitated communication since 1995: A review of published studies. *Journal of Autism and Developmental Disorders*, *31*, 287–320.

Mostert, M. P. (2002). Letter to the editor: Teaching the illusion of facilitated communication. *Journal of Autism and Developmental Disorders*, *32*, 239–240.

Niemi, J., & Kärnä-Lin, E. (2002). Grammar and lexicon in facilitated communication: A linguistic authorship analysis of a Finnish case. *Mental Retardation*, *40*, 347–357.

Niemi, J., & Kärnä-Lin, E. (2003). Four vantage points to the language performance and capacity of human beings: Response to Saloviita and Sariola. *Mental Retardation*, *41*, 380–385.

Oakes, M. P. (1998). *Statistics for corpus linguistics*, Edinburgh: Edinburgh University Press.

Perkins, M. R., Dobbinson, S., Boucher, J., Bol, S., & Bloom P. (2006). Lexical knowledge and lexical use in autism. *Journal of Autism and Developmental Disorders*, *36*, 795–805.

R Development Core Team. (2008). R: A language and environment for statistical computing (Ver. 2.8.1) [Software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved December 22, 2008, from http://www.R-project.org

Salomon, M. J., Wagner, S. H., & Bauman, M. L. (1996). A validated case study of facilitated communication. *Mental Retardation*, *34*, 220–230.

Saloviita, T., & Sariola, H. (2003). Authorship in facilitated communication: A re-analysis of a case of assumed representative authentic writing. *Mental Retardation*, *41*, 374–379.

Scopesi, A., Zanobini, M., & Cresci, L. R. (2003). Aspetti semantici e stilistici della produzione di un bambino autistico in situazione di comunicazione facilitata [*Semantic and stylistic aspects of the production of an autistic child in a facilitated communication environment*]. *Rivista di Psicolinguistica Applicata*, *2–3*, 83–105.

Sirmakessis, S. (2004). *Text mining and its applications*. Heidelberg, Germany: Springer-Verlag.

Smukler, D. (2005). Unauthorized minds: How "theory of mind" theory misrepresents autism. *Mental Retardation*, *43*, 11–24.

Stone, W. L., & Yoder, P. J. (2001). Predicting spoken language level in children with autism spectrum disorders. *Autism*, *5*, 341–361.

Strauss, U., Fan, F., & Altmann, G. (2008), *Problems in quantitative linguistics*. Lüdenscheid, Germany: RAM-Verlag.

Sullivan, D. (2001). *Document warehousing and text mining. Techniques for improving business operations, marketing and sales*. New York: Wiley.

Tager-Flusberg, H. (1985). The conceptual basis for referential word meaning in children with autism. *Child Development*, *56*, 1167–1178.

Tager-Flusberg, H. (2000). The challenge of studying language development in children with autism. In L. Menn & N. B. Ratner (Eds.), *Methods for studying language production* (pp. 313–332). Mahwah, NJ: Erlbaum.

Tweedie, F. J., & Baayen, H. R. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*, 323–352.

Watkins, R.V., Kelly, D. J., & Harbers, H. M. (1995). Measuring children's lexical diversity. Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research*, *38*, 1349–1355.

Received 6/1/07, first decision 12/18/07, accepted 12/15/08.

Editor-in-Charge: Steven J. Taylor

The present study was included in the activities conducted within the frame of the EASIEST University Project, which focus on the written language of individuals with autism. The EASIEST Project is an interdisciplinary study involving linguistics, neuropsychiatry, psychology, sociology, statistics, and computer-aided text processing (Bernardi, 2008) funded by the University of Padua for the 2005–2006 period (Scientific Coordinator: Lorenzo Bernardi, Department of Statistical Sciences, University of Padua).

The text collection was coordinated by Vittoria Cristoferi Realdon, child neuropsychiatrist, and conducted in four accredited FC centers in Italy: Centro Studi e Ricerca in Neuroriabilitazione CNAPP in Rome, Centro Studi sulla Comunicazione Facilitata - W.O.C.E. in Zoagli (Genoa), Istituto M.P.P. Padri Trinitari A. Quarto di Palo in Andria (Bari), and Centro Sperimentale per i Disturbi dello Sviluppo e della Comunicazione in Padua.

Author:
Arjuna Tuzzi, PhD (E-mail: arjuna.tuzzi@unipd.it), Associate Professor, Department of Sociology, University of Padua, Padua, 35123 Italy.

