Automatic detection of deception in child-produced speech using syntactic complexity features

Maria Yancheva
Division of Engineering Science,
University of Toronto
Toronto Ontario Canada
maria.yancheva@utoronto.ca

Frank Rudzicz
Toronto Rehabilitation Institute; and
Department of Computer Science,
University of Toronto
Toronto Ontario Canada
frank@cs.toronto.edu

Abstract

It is important that the testimony of children be admissible in court, especially given allegations of abuse. Unfortunately, children can be misled by interrogators or might offer false information, with dire consequences. In this work, we evaluate various parameterizations of five classifiers (including support vector machines, neural networks, and random forests) in deciphering truth from lies given transcripts of interviews with 198 victims of abuse between the ages of 4 and 7. These evaluations are performed using a novel set of syntactic features, including measures of complexity. Our results show that sentence length, the mean number of clauses per utterance, and the Stajner-Mitkov measure of complexity are highly informative syntactic features, that classification accuracy varies greatly by the age of the speaker, and that accuracy up to 91.7% can be achieved by support vector machines given a sufficient amount of data.

1 Introduction

The challenge of disambiguating between truth and deception is critical in determining the admissibility of court testimony. Unfortunately, the testimony of maltreated children is often not admitted in court due to concerns about truthfulness since children can be instructed to deny transgressions or misled to elicit false accusations (Lyon and Dorado, 2008). However, the child is often the only witness of the transgression (Undeutsch, 2008); automatically determining truthfulness in such situations is therefore a paramount goal so that justice may be served effectively.

2 Related Work

Research in the detection of deception in adult speech has included analyses of verbal and non-verbal cues such as behavioral changes, facial expression, speech dysfluencies, and cognitive complexity (DePaulo et al., 2003). Despite statistically significant predictors of deception such as shorter talking time, fewer semantic details, and less coherent statements, DePaulo et al. (2003) found that the median effect size is very small. Deception without special motivation (e.g., everyday 'white lies') exhibited almost no discernible cues of deception. However, analysis of moderating factors showed that cues were significantly more numerous and salient when lies were about transgressions.

Literature on deception in children is relatively limited. In one study, Lewis et al. (1989) studied 3-year-olds and measured behavioral cues, such as facial expression and nervous body movement, before and after the elicitation of a lie. Verbal responses consisted of yes/no answers. Results suggested that 3-year-old children are capable of deception, and that non-verbal behaviors during deception include increases in 'positive' behaviors (e.g., smiling). However, verbal cues of deception were not analyzed. Crucially, Lewis et al. (1989) showed that humans are no more accurate in deciphering truth from deception in child speech than in adult speech, being only about 50% accurate.

More recently, researchers have used linguistic features to identify deception. Newman et al. (2003) inferred deception in transcribed, typed, and handwritten text by identifying features of linguistic style such as the use of personal pronouns

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 944–953,
Sofia, Bulgaria, August 4-9 2013.

and exclusive words (e.g., *but*, *except*, *without*). These features were obtained with the Linguistic Inquiry and Word Count (LIWC) tool and used in a logistic regression classifier which achieved, on average, 61% accuracy on test data. Feature analysis showed that deceptive stories were characterized by fewer self-references, more negative emotion words, and lower cognitive complexity, compared to non-deceptive language.

Another recent stylometric experiment in automatic identification of deception was performed by Mihalcea and Strapparava (2009). The authors used a dataset of truthful and deceptive typed responses produced by adult subjects on three different topics, collected through the Amazon Mechanical Turk service. Two classifiers, Naïve Bayes (NB) and a support vector machine (SVM), were applied on the tokenized and stemmed statements to obtain best classification accuracies of 70% (abortion topic, NB), 67.4% (death penalty topic, NB), and 77% (friend description, SVM), where the baseline was taken to be 50%. The large variability of classifier performance based on the topic of deception suggests that performance is context-dependent. The authors note this as well by demonstrating significantly lower results of 59.8% for NB and 57.8% for SVM when cross-topic classification is performed by training each classifier on two topics and testing on the third.

The Mihalcea-Strapparava mturk dataset was further used in a study by Feng et al. (2012) which employs lexicalized and unlexicalized production rules to obtain deep syntactic features. The cross-validation accuracy obtained on the three topics was improved to 77% (abortion topic), 71.5% (death penalty topic), and 85% (friend description). The results nevertheless varied with topic.

Another experiment using syntactic features for identifying sentences containing uncertain or unreliable information was conducted by Zheng et al. (2010) on an adult-produced dataset of abstracts and full articles from BioScope, and on paragraphs from Wikipedia. The results demonstrated that using syntactic dependency features extracted with the Stanford parser improved performance on the biological dataset, while an ensemble classifier combining a conditional random field (CRF) and a MaxEnt classifier performed better than individual classifiers on the Wikipedia dataset.

A meta-analysis of features used in deception detection was performed by Hauch et al. (2012) and revealed that verbal cues based on lexical categories extracted using the LIWC tool show statistically significant, though small, differences between truth- and lie-tellers. Vartapetiance and Gillam (2012) surveyed existing cues to verbal deception and demonstrated that features in LIWC are not indicative of deception in online content, recommending that the features used to identify deception and the thresholds between deception and truth be based on the specific data set.

In the speech community, analysis of deceptive speech has combined various acoustic, prosodic, and lexical features (Hirschberg et al., 2005). Graciarena et al. (2006) combined two independent systems — an acoustic Gaussian mixture model based on Mel cepstral features, and a prosodic support vector machine based on features such as pitch, energy, and duration — and achieved an accuracy of 64.4% on a test subset of the Columbia-SRI-Colorado (CSC) corpus of deceptive and non-deceptive speech (Hirschberg et al., 2005).

While previous studies have achieved some promising results in detecting deception with lexical, acoustic, and prosodic features, syntax remains relatively unexplored compared to LIWC-based features. Syntactic complexity as a cue to deception is consistent with literature in social psychology which suggests that emotion suppression (e.g., inhibition of guilt and fear) consumes cognitive resources, which can influence the underlying complexity of utterances (Richards and Gross, 1999; Richards and Gross, 2000). Additionally, the use of syntactic features is motivated by their successful use on adult-produced datasets for detecting deceptive or uncertain utterances (Feng et al., 2012; Zheng et al., 2010), as well as in other applications, such as the evaluation of changes in text complexity (Stajner and Mitkov, 2012), the identification of personality in conversation and text (Mairesse et al., 2007), and the detection of dementia through syntactic changes in writing (Le et al., 2011).

Past work has focused on identifying deceptive speech produced by adults. The problem of determining validity of child testimony in high-stakes child abuse court cases motivates the analysis of child-produced deceptive language. Further, the use of binary classification schemes in previous work does not account for partial truths often encountered in real-life scenarios. Due to the rarity of real deceptive data, studies typically use arti-

ficially produced deceptive language which falls unambiguously in one of two classes: complete truth or complete deception (Newman et al., 2003; Mihalcea and Strapparava, 2009). Studies which make use of real high-stakes courtroom data containing partial truths, such as the Italian DECOUR corpus analyzed by Fornaciari and Poesio (2012), preprocess the dataset to eliminate any partially truthful utterances. Since utterances of this kind are common in real language, their elimination from the dataset is not ideal.

The present study evaluates the viability of a novel set of 17 syntactic features as markers of deception in five classifiers. Moreover, to our knowledge, it is the first application of automatic deception detection to a real-life dataset of deceptive speech produced by maltreated children. The data is scored using a gradient of truthfulness, which is used to represent completely true, partially true, and completely false statements. Descriptions of the data (section 3) and feature sets (section 4) precede experimental results (section 5) and the concluding discussion (section 6).

3 Data

The data used in this study were obtained from Lyon et al. (2008), who conducted and transcribed a truth-induction experiment involving maltreated children awaiting court appearances in the Los Angeles County Dependency Court. Subjects were children between the ages of 4 and 7 (99 boys and 99 girls) who were interviewed regarding an unambiguous minor transgression involving playing with a toy. To ensure an understanding of lying and its negative consequences, all children passed a preliminary oath-taking competency task, requiring each child to correctly identify a truth-teller and a lie-teller in an object labeling task, as well as to identify which of the two would be the target of negative consequences.

During data collection, a confederate first engaged each child individually in one of four conditions: a) *play*, b) *play and coach*, c) *no play*, and d) *no play and coach*. In the two *play* conditions, the confederate engaged the child in play with a toy house (in the *no play* conditions, they did not); in the two *coach* conditions, the confederate coached the child to lie (i.e., to deny playing if they played with the toy house, or to admit playing if they did not). The confederate then left and the child was interviewed by a second researcher who performed a truth-induction manipulation consisting of one of: a) *control* — no manipulation, b) *oath* — the interviewer reminded the child of the importance of telling the truth and elicited a promise of truth-telling, and c) *reassurance* — the interviewer reassured the child that telling the truth will not lead to any negative consequences.

Each pre- and post-induction transcription may contain explicit statements of up to seven features: looking at toy-house, touching toy-house, playing with toy-house, opening toy-house doors or windows to uncover hidden toys, playing with these hidden toys, spinning the toy-house, and putting back or hiding a toy. All children in the *play* condition engaged in all seven actions, while children in the *no play* condition engaged in none. An eighth feature is the lack of explicit denial of touching or playing with the toy house, which is considered to be truthful in the *play* condition, and deceptive in the *no play* condition (see the examples in the appendix). A transcription is labeled as *truth* if at least half of these features are truthful (53.2% of all transcriptions) and *lie* otherwise (46.8% of transcriptions). Other thresholds for this binary discrimination are explored in section 5.4.

Each child's verbal response was recorded twice: at time T_1 (prior to truth-induction), and at time T_2 (after truth-induction). Each child was subject to one of the four confederate conditions and one of the three induction conditions. The raw data were pre-processed to remove subjects with blank transcriptions, resulting in a total of 173 subjects (87 boys and 86 girls) and 346 transcriptions.

4 Methods

Since the data consist of speech produced by 4- to 7-year-old children, the predictive features must depend on the level of syntactic competence of this age group. The "continuity assumption" states that children have a complete system of abstract syntactic representation and have the same set of abstract functional categories accessible to adults (Pinker, 1984). An experimental study with 3- to 8-year-old children showed that their syntactic competence is comparable to that of adults; specifically, children have a productive rule for passive forms which allows them to generalize to previously unheard predicates while following adult-like constraints to avoid over-generalization (Pinker et al., 1987). Recent experiments with syntactic priming showed that children's representations of abstract passive constructions are well-developed as early as age 3 or 4, and young

children are generally able to form passive constructions with both action and non-action verbs (Thatcher et al., 2007). These results suggest that measures of syntactic complexity that are typically used to evaluate adult language could be adapted to child speech, provided that the children are at least 3 or 4 years old.

Here, the complexity of speech is characterized by the length of utterances and by the frequency of dependent and coordinate clauses, with more complex speech consisting of longer utterances and a higher number of subordinate clauses. We segmented the transcriptions into sentences, clauses and *T-units*, which are "minimally terminable units" consisting of a main clause and its dependent clauses (Hunt, 1965; O'Donnell et al., 1967)[1]. Deceptive communication generally has shorter duration and is less detailed than non-deceptive speech (DePaulo et al., 2003), so the length of each type of segment was counted along with frequency features over segments. Here, the frequency of dependent and coordinate clauses per constituent approximate clause-based measures of complexity.

Our approach combines a set of features obtained from a functional dependency grammar (FDG) parser with another (non-overlapping) set of features obtained from a phrase-based grammar parser. We obtained FDG parses of the transcriptions using Connexor's Machinese Syntax parser (Tapanainen and Järvinen, 1997) and extracted the following 5 features:

ARI *Automated readability index*. Measures word and sentence difficulty, $4.71\frac{c}{w}+0.5\frac{w}{s}-21.43$, where c is the number of characters, w is the number of words, and s is the number of sentences (Smith and Senter, 1967).

ASL *Average sentence length*. The number of words over the number of sentences.

COM *Sentence complexity*. The ratio of sentences with $\geq$ 2 finite predicators to those with $\leq$ 1 finite predicator (Stajner and Mitkov, 2012).

PAS *Passivity*. The ratio of non-finite main predicators in a passive construction (@–FMAINV %VP) to the total number of finite (@+FMAINV %VA) and non-finite (@–FMAINV %VA and @–FMAINV %VP) main predicators, including active constructions.

MCU Mean number of clauses per utterance.

Additionally, we searched for specific syntactic patterns in phrase-based parses of the data. We used the Stanford probabilistic natural language parser (Klein and Manning, 2003) for constructing these parse trees, the Stanford Tregex utility (Levy and Andrew, 2006) for searching the constructed parse trees, and a tool provided by Lu (2011) which extracts a set of 14 clause-based features in relation to sentence, clause and T-unit constituents.

4.1 Feature analysis

Analysis of variance (ANOVA) was performed on the set of 17 features, shown in Table 1. A one-factor ANOVA across the *truth* and *lie* groups showed three significant feature variations: average sentence length (ASL), sentence complexity (COM), and mean clauses per utterance (MCU). Dependencies between some feature pairs that are positively correlated are shown in Figure 1.

As expected, the number of clauses (MCU) is dependent on sentence length (ASL) ($r(344) = .92$, $p < .001$). Also, the number of T-units is dependent on the number of clauses: CN/C is correlated with CN/T ($r(344) = .89$, $p < .001$), CP/C is correlated with CP/T ($r(344) = .85$, $p < .001$), and DC/C is correlated with DC/T ($r(344) = .92$, $p < .001$). Other features are completely uncorrelated. For example, the number of passive constructions is independent of sentence length ($r(344) = -.0020$, $p > .05$), the number of complex nominals per clause is independent of clause length ($r(344) = .076$, $p > .05$), and the density of dependent clauses is independent of the density of coordinate phrases ($r(344) = -.027$, $p > .05$).

5 Results

We evaluate five classifiers: logistic regression (**LR**), a multilayer perceptron (**MLP**), naïve Bayes (**NB**), a random forest (**RF**), and a support vector machine (**SVM**). Here, naïve Bayes, which assumes conditional independence of the features, and logistic regression, which has a linear decision boundary, are baselines. The MLP includes a variable number of layers of hidden units, which

[1] T-units include single clauses, two or more phrases in apposition, or clause fragments. Generally, coordinate clauses are split into separate T-units, as are clauses interrupted by discourse boundary markers.

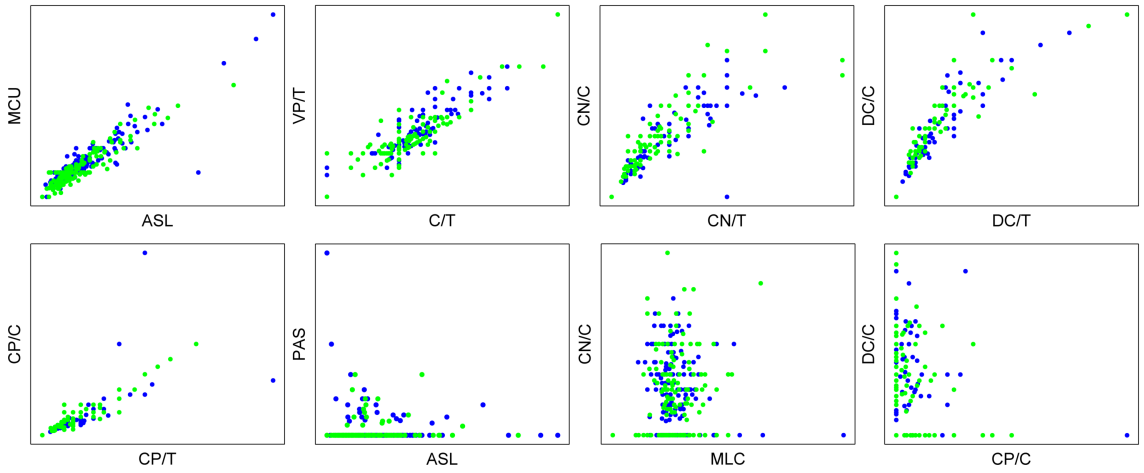

Figure 1: Independent and dependent feature pairs; data points are labeled as *truth* (blue) and *lie* (green).

Feature	$\mathbf{F}_{1,344}$	d
Automated Readability Index (**ARI**)	0.187	0.047
Average Sentence Length (**ASL**)	3.870	**0.213**
Sentence Complexity (**COM**)	10.93	**0.357**
Passive Sentences (**PAS**)	1.468	0.131
Mean Clauses per Utterance (**MCU**)	6.703	**0.280**
Mean Length of T-Unit (**MLT**)	2.286	0.163
Mean Length of Clause (**MLC**)	0.044	-0.023
Verb Phrases per T-Unit (**VP/T**)	3.391	0.199
Clauses per T-Unit (**C/T**)	2.345	0.166
Dependent Clauses per Clause (**DC/C**)	1.207	0.119
Dependent Clauses per T-Unit (**DC/T**)	1.221	0.119
T-Units per Sentence (**T/S**)	3.692	0.208
Complex T-Unit Ratio (**CT/T**)	2.103	0.157
Coordinate Phrases per T-Unit (**CP/T**)	0.463	-0.074
Coordinate Phrases per Clause (**CP/C**)	0.618	-0.085
Complex Nominals per T-Unit (**CN/T**)	0.722	0.092
Complex Nominals per Clause (**CN/C**)	0.087	0.032

Table 1: One-factor ANOVA (**F** statistics and Cohen's d-values, $\alpha = 0.05$) on all features across *truth* and *lie* groups. Statistically significant results are in bold.

apply non-linear activation functions on a linear combination of inputs. The SVM is a parametric binary classifier that provides highly non-linear decision boundaries given particular kernels. The random forest is an ensemble classifier that returns the mode of the class predictions of several decision trees.

5.1 Binary classification across all data

The five classifiers were evaluated on the entire pooled data set with 10-fold cross validation. Table 2 lists the parameters varied for each classifier, and Table 3 shows the cross-validation accuracy for the classifiers with the best parameter settings. The naïve Bayes classifier performs poorly, as could be expected given the assumption of conditional feature independence. The SVM classifier performs best, with 59.5% cross-validation accuracy, which is a statistically significant improvement over the baselines of LR ($t(4) = 22.25, p < .0001$), and NB ($t(4) = 16.19, p < .0001$).

	Parameter		Values
LR	R	Ridge value	10^{-10} to 10^{-2}
MLP	L	Learning rate	0.0003 to 0.3
	M	Momentum	0 to 0.5
	H	Number of hidden layers	1 to 5
NB	K	Use kernel estimator	true, false
RF	I	Number of trees	1 to 20
	K	Maximum depth	unlimited, 1 to 10
SVM	K	Kernel	Linear, RBF, Polynomial
	E	Polynomial Exponent	2 to 5
	G	RBF Gamma	0.001 to 0.1
	C	Complexity constant	0.1 to 10

Table 2: Empirical parameter settings for each classifier

5.2 Binary classification by age group

Significant variation in syntactic complexity is expected across ages. To account for such variation, we segmented the dataset in four groups: 44 tran-

	Accuracy	Parameters
LR	0.5347	$R = 10^{-10}$
MLP	0.5838	$L = 0.003$, $M = 0.4$
NB	0.5173	$K =$ false
RF	0.5809	$I = 10$, $K = 6$
SVM	0.5954	Polynomial, $E = 3$, $C = 1$

Table 3: Cross-validation accuracy of binary classification performed on entire dataset of 346 transcriptions.

scriptions of 4-year-olds, 120 of 5-year-olds, 94 of 6-year-olds, and 88 of 7-year-olds. By comparison, Vrij et al. (2004) used data from only 35 children in their study of 5- and 6-year-olds. Classification of truthfulness was performed separately for each age, as shown in Table 4. In comparison with classification accuracy on pooled data, a paired t-test shows statistically significant improvement across all age groups using RF, $t(3) = 10.37, p < .005$.

	Age (years)			
	4	**5**	**6**	**7**
LR	0.6136	0.5333	0.5957*	0.4886
MLP	0.6136†	0.5583	0.6170†	0.5909*
NB	0.6136*	0.5250	0.5426	0.5682
RF	0.6364†	**0.6333***	**0.6383**†	**0.6591**†
SVM	**0.6591**	0.5583	0.6064	0.6250*

Table 4: Cross-validation accuracy of binary classification partitioned by age. The best classifier at each age is shown in bold. The classifiers showing statistically significant incremental improvement are marked: *$p < .05$, $^\dagger p < .001$ (paired t-test, d.f. 4)

5.3 Binary classification by age group, on verbose transcriptions

The length of speech, in number of words, varies widely ($min = 1$, $max = 167$, $\mu = 36.83$, $\sigma = 28.34$) as a result of the unregulated nature of the interview interaction. To test the effect of verbosity, we segment the data by child age and select only the transcriptions with above-average word counts (i.e., ≥ 37 words), resulting in four groups: 12 transcriptions of 4-year-olds, 48 of 5-year-olds, 39 of 6-year-olds, and 37 of 7-year-olds. This mimics the scenario in which some minimum threshold is placed on the length of a child's speech. In this verbose case, 63.3% of transcripts are labeled truth across age groups (using the same definition of truth as in section 3), with no substantial variation between ages; in the non-verbose case, 53.2% are marked truth. Fisher's exact test on this contingency table reveals no significant difference between these distributions ($p = 0.50$). Classification results are shown in Table 5. The size of the training set for the youngest age category is low compared to the other age groups, which may reduce the reliability of the higher accuracy achieved in that group. The other three age groups show a growing trend, which is consistent with expectations — older children exhibit greater syntactic complexity in speech, allowing greater variability of feature values across truth and deception. Here, both SVM and RF achieve 83.8% cross-validation accuracy in identifying deception in the speech of 7-year-old subjects.

	4	**5**	**6**	**7**
LR	0.7500†	0.5417	0.6667†	0.7297†
MLP	0.8333†	**0.6250**†	0.6154	0.7838†
NB	0.6667†	0.4583	0.4103	0.7297*
RF	0.8333†	0.5625	**0.7179**†	**0.8378**†
SVM	**0.9167***	**0.6250**†	0.6154*	**0.8378**†

Table 5: Cross-validation accuracy of binary classification performed on transcriptions with above average word count (136 transcriptions), by age group. Rows represent classifiers, columns represent ages. The best classifier for each age is in bold. The classifiers showing statistically significant incremental improvement are marked: *$p < .05$, $^\dagger p < .001$ (paired t-test, d.f. 4)

5.4 Threshold variation

To study the effect of the threshold between the *truth* and *lie* classes, we vary the value of the threshold, τ, from 1 to 8, requiring the admission of at least τ truthful details (out of 8 possible details) in order to label a transcription as *truth*. The effect of τ on classification accuracy over the entire pooled dataset for each of the 5 classifiers is shown in Figure 2. A one-factor ANOVA with τ as the independent variable with 8 levels, and cross-validation accuracy as the dependent variable, confirms that the effect of the threshold is statistically significant ($F_{7,40} = 220.69$, $p < .0001$) with $\tau = 4$ being the most conservative setting.

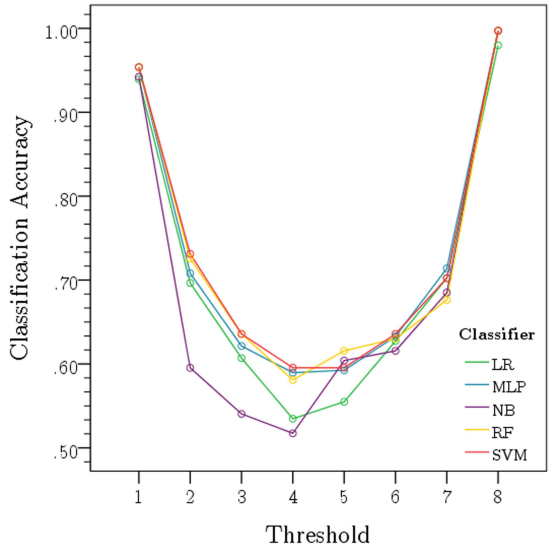

Figure 2: Effect of threshold and classifier choice on cross-validation accuracy. Threshold $\tau = 0$ is not present, since all data would be labeled *truth*.

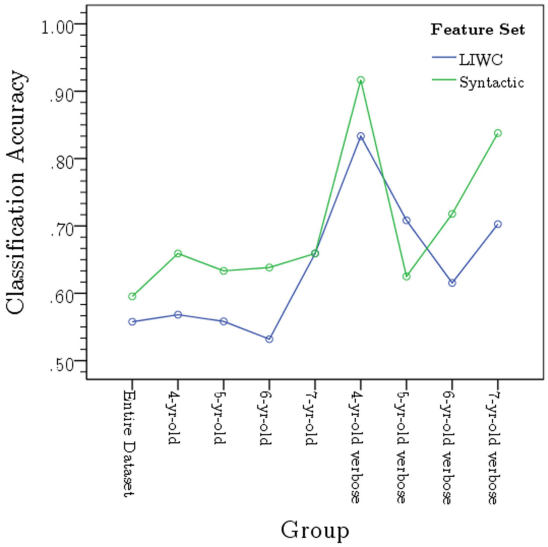

Figure 3: Effect of feature set choice on cross-validation accuracy.

5.5 Linguistic Inquiry and Word Count

The Linguistic Inquiry and Word Count (LIWC) tool for generating features based on word category frequencies has been used in deception detection with adults, specifically: first-person singular pronouns (FP), exclusive words (EW), negative emotion words (NW), and motion verbs (MV) (Newman et al., 2003). We compare the performance of classifiers trained with our 17 syntactic features to those of classifiers trained with those LIWC-based features on the same data. To evaluate the four LIWC categories, we use the 86 words of the Pennebaker model (Little and Skillicorn, 2008; Vartapetiance and Gillam, 2012). The performance of the classifiers trained with LIWC features is shown in Table 6.

The set of 17 syntactic features proposed here result in significantly higher accuracies across classifiers and experiments ($\mu = 0.63$, $\sigma = 0.10$) than with the LIWC features used in previous work ($\mu = 0.58$, $\sigma = 0.09$), as shown in Figure 3 ($t(53) = -0.0691$, $p < .0001$).

6 Discussion and future work

This paper evaluates automatic estimation of truthfulness in the utterances of children using a novel set of lexical-syntactic features across five types of classifiers. While previous studies have favored word category frequencies extracted with LIWC (Newman et al., 2003; Little and Skillicorn, 2008; Hauch et al., 2012; Vartapetiance and Gillam, 2012; Almela et al., 2012; Fornaciari and Poesio, 2012), our results suggest that the set of syntactic features presented here perform significantly better than the LIWC feature set on our data, and across seven out of the eight experiments based on age groups and verbosity of transcriptions.

Statistical analyses showed that the average sentence length (ASL), the Stajner-Mitkov measure of sentence complexity (COM), and the mean number of clauses per utterance (MCU) are the features most predictive of truth and deception (see section 4.1). Further preliminary experiments are exploring two methods of feature selection, namely forward selection and minimum-Redundancy-Maximum-Relevance (mRMR). In forward selection, features are greedily added one-at-a-time (given an initially empty feature set) until the cross-validation error stops decreasing with the addition of new features (Deng, 1998). This results in a set of only two features: sentence complexity (COM) and T-units per sentence (T/S). Features are selected in mRMR by minimizing redundancy (i.e., the average mutual information between features) and maximizing the relevance (i.e., the mutual information between the given features and the class) (Peng et al., 2005). This approach selects five features: verb phrases per T-unit (VP/T), passive sentences (PAS), coordinate phrases per clause (CP/C), sentence complexity (COM), and complex nominals per clause (CN/C). These results confirm the predictive strength of sentence complexity. Further, preliminary classi-

Group	Accuracy	Best Classifier	Parameters
Entire dataset	0.5578	RF	$I = 20$, $K = unlimited$
4-yr-olds	0.5682	MLP	$L = 0.005$, $M = 0.3$, $H = 1$
5-yr-olds	0.5583	RF	$I = 5$, $K = unlimited$
6-yr-olds	0.5319	MLP	$L = 0.005$, $M = 0.3$, $H = 1$
7-yr-olds	0.6591	RF	$I = 5$, $K = unlimited$
4-yr-olds, verbose	0.8333	SVM	PolyKernel, $E = 4$, $C = 10$
5-yr-olds, verbose	0.7083	SVM	NormalizedPolyKernel, $E = 1$, $C = 10$
6-yr-olds, verbose	0.6154	MLP	$L = 0.09$, $M = 0.2$, $H = 1$
7-yr-olds, verbose	0.7027	MLP	$L = 0.01$, $M = 0.5$, $H = 3$

Table 6: Best 10-fold cross-validation accuracies achieved on various subsets of the data, using the LIWC-based feature set.

fication results across all classifiers suggest that accuracies are significantly higher given forward selection ($\mu = 0.58$, $\sigma = 0.02$) relative to the original feature set ($\mu = 0.56, \sigma = 0.03$); $t(5) = -2.28$, $p < .05$ while the results given the mRMR features are not significantly different.

Generalized cross-validation accuracy increases significantly given partitioned age groups, which suggests that the importance of features may be moderated by age. A further incremental increase is achieved by considering only transcriptions above a minimum length. O'Donnell et al. (1967) examined syntactic complexity in the speech and writing of children aged 8 to 12, and found that speech complexity increases with age. This phenomenon appears to be manifested in the current study by the extent to which classification increases generally across the 5-, 6-, and 7-year-old groups, as shown in Table 5. Future examination of the effect of age on feature saliency may yield more appropriate age-dependent features.

While past research has used logistic regression as a binary classifier (Newman et al., 2003), our experiments show that the best-performing classifiers allow for highly non-linear class boundaries; SVM and RF models achieve between 62.5% and 91.7% accuracy across age groups — a significant improvement over the baselines of LR and NB, as well as over previous results. Moreover, since the performance of human judges in identifying deception is not significantly better than chance (Lewis et al., 1989; Newman et al., 2003), these results show promise in the use of automatic detection methods.

Partially truthful transcriptions were scored using a gradient of 0 to 8 truthful details, and a threshold τ was used to perform binary classification. Extreme values of τ lead to poor F-scores despite high accuracy, since the class distribution of transcriptions is very skewed towards either class. Future work can explore the effect of threshold variation given sufficient data with even class distributions for each threshold setting. When such data is unavailable, experiments can make use of the most conservative setting ($\tau = 4$, or an equivalent mid-way setting) for analysis of real-life utterances containing partial truths.

Future work should consider measures of confidence for each classification, where possible, so that more ambiguous classifications are not treated on-par with more certain ones. For instance, confidence can be approximated in MLPs by the entropy across continuous-valued output nodes, and in RFs by the number of component decision trees that agree on a classification. Although acoustic data were not provided with this data set (Lyon and Dorado, 2008) (and, in practice, cannot be assured), future work should also examine the differences in the acoustics of children across truth conditions.

Acknowledgments

The authors thank Kang Lee (Ontario Institute for Studies in Education, University of Toronto) and Angela Evans (Brock University) for sharing both this data set and their insight.

Appendix

The following is an example of evasive deceptive speech from a 6-year-old after no truth induction (i.e., the *control* condition in which the interviewer merely states that he needs to ask more questions):

> *... Yeah yeah ok, I'm a tell you. We played that same game and I won and he won. I'm going to be in trouble if I tell you. It a secret. It's a secret 'cuz we're friends. ...*

Transcription excerpt labeled as *truth* by a threshold of $\tau = 1$: 7-year-old child's response (*play, no coach* condition), in which the child does not explicitly deny playing with the toy house, and admits to looking at it but does not confess to any of the other six actions:

> *...I was playing, I was hiding the coin and I was trying to find the house... trying to see who was in there...*

Transcription excerpt labeled as *truth* by a threshold of $\tau = 4$: 7-year-old child's response (*play, no coach* condition), in which the child does not explicitly deny playing, and admits to three actions:

> *...me and him was playing with it... we were just spinning it around and got the toys out...*

References

Angela Almela, Rafael Valencia-Garcia, and Pascual Cantos. 2012. Seeing through deception: A computational approach to deceit detection in written communication. *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection*, April 23-27, 2012, Avignon, France, 15-22.

Ethem Alpaydin. 2010. Introduction to Machine Learning. Cambridge, MA: *MIT Press*.

Kan Deng. 1998. OMEGA: On-line memory-based general purpose system classifier. Doctoral thesis, School of Computer Science, Carnegie Mellon University

Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin*, 129(1):74-118.

Song Feng, Ritwik Banerjee and Yejin Choi. 2012. Syntactic stylometry for deception detection. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, July 8-14, 2012, Jeju, Republic of Korea, 171-175.

Tommaso Fornaciari and Massimo Poesio. 2012. On the use of homogeneous sets of subjects in deceptive language analysis. *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection*, April 23-27, 2012, Avignon, France, 39-47.

Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Frank Enos, Julia Hirschberg, Sachin Kajarekar. 2006. Combining prosodic lexical and cepstral systems for deceptive speech detection. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages I-1033-I-1036.

Valerie Hauch, Jaume Masip, Iris Blandon-Gitlin, and Siegfried L. Sporer. 2012. Linguistic cues to deception assessed by computer programs: A meta-analysis. *Proceedings of the EACL Workshop on Computational Approaches to Deception Detection*, pages 1-4.

Julia Hirschberg, Stefan Benus, Jason M. Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, Bryan Pellom, Elizabeth Shriberg, and Andreas Stolcke. 2005. Distinguishing deceptive from non-deceptive speech. *Proceedings of Eurospeech 2005*, pages 1833-1836.

Kellogg W. Hunt. 1965. Grammatical structures written at three grade levels. *NCTE Research Report No. 3*.

Dan Klein and Christopher Manning. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423-430.

Xuan Le, 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):435-461.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. *5th International Conference on Language Resources and Evaluation*.

Michael Lewis, Catherine Stanger, and Margaret W. Sullivan. 1989. Deception in 3-year-olds. *Developmental Psychology*, 25(3):439-443.

A. Little and D. B. Skillicorn. 2008. Detecting deception in testimony. *Proceedings of IEEE International Conference of Intelligence and Security Informatics (ISI 2008)*, June 17-20, 2008, Taipei, Taiwan, 13-18.

Xiaofei Lu. 2011. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474-496.

Thomas D. Lyon and J. S. Dorado. 2008. Truth induction in young maltreated children: the effects of oath-taking and reassurance on true and false disclosures. *Child Abuse & Neglect*, 32(7):738-748.

Thomas D. Lyon, Lindsay C. Malloy, Jodi A. Quas, and Victoria A. Talwar. 2008. Coaching, truth induction, and young maltreated children's false allegations and false denials. *Child Development*, 79(4):914-929.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457-500.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: explorations in the automatic recognition of deceptive language. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, August 4, 2009, Suntec, Singapore, 309-312.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: predicting deception from linguistic styles. *PSPB*, 29(5):665-675.

Roy C. O'Donnell, William J. Griffin, and Raymond C. Norris. 1967. A transformational analysis of oral and written grammatical structures in the language of children in grades three, five, and seven. *PSPB*, 29(5):665-675.

Steven Pinker. 1984. *Language learnability and language development*, Cambridge, MA: Harvard University Press.

Steven Pinker, David S. Lebeaux, and Loren Ann Frost. 1987. Productivity and constraints in the acquisition of the passive. *Cognition*, 26:195-267.

Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226-1238.

J. M. Richards and J. J. Gross. 1999. Composure at any cost? The cognitive consequences of emotion suppression. *PSPB*, 25(8):1033-1044.

J. M. Richards and J. J. Gross. 2000. Emotion regulation and memory: the cognitive costs of keeping one's cool. *Journal of Personality and Social Psychology*, 79:410-424.

E. A. Smith and R. J. Senter. 1967. Automated readability index. Technical report, Defense Technical Information Center. United States.

Sanja Stajner and Ruslan Mitkov. 2012. Diachronic changes in text complexity in 20th century English language: an NLP Approach. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1577-1584.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64-71.

Katherine Thatcher, Holly Branigan, Janet McLean, and Antonella Sorace. 2007. Children's early acquisition of the passive: evidence from syntactic priming. *Child Language Seminar, University of Reading*.

Udo Undeutsch. 2008. Courtroom evaluation of eyewitness testimony. *Applied Psychology*, 33(1):51-66.

Anna Vartapetiance and Lee Gillam. 2012. "I don't know where he is not": does deception research yet offer a basis for deception detectives? *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection*, April 23-27, 2012, Avignon, France, 5-14.

Aldert Vrij, Lucy Akehurst, Stavroula Soukara, and Ray Bull. 2004. Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. *Human Communication Research*, 30(1):8–41

Yi Zheng, Qifeng Dai, Qiming Luo, and Enhong Chen. 2010. Hedge classification with syntactic dependency features based on an ensemble classifier. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, July 15-16, 2010, Uppsala, Sweden, 151-156.