Lexical Features Are More Vulnerable,
Syntactic Features Have More Predictive Power

Jekaterina Novikova[1], **Aparna Balagopalan**[1], **Ksenia Shkaruta**[2] and **Frank Rudzicz**[1,3]

[1]Winterlight Labs, {jekaterina, aparna}@winterlightlabs.com
[2]Georgia Tech, ksenia.shkaruta@gatech.edu
[3]University of Toronto; Vector Institute for Artificial Intelligence, frank@cs.toronto.edu

Abstract

Understanding the vulnerability of linguistic features extracted from noisy text is important for both developing better health text classification models and for interpreting vulnerabilities of natural language models. In this paper, we investigate how generic language characteristics, such as syntax or the lexicon, are impacted by artificial text alterations. The vulnerability of features is analysed from two perspectives: (1) the level of feature value change, and (2) the level of change of feature predictive power as a result of text modifications. We show that lexical features are more sensitive to text modifications than syntactic ones. However, we also demonstrate that these smaller changes of syntactic features have a stronger influence on classification performance downstream, compared to the impact of changes to lexical features. Results are validated across three datasets representing different text-classification tasks, with different levels of lexical and syntactic complexity of both conversational and written language.

1 Introduction

It is important to understand the vulnerability of linguistic features to text alteration because (1) pre-defined linguistic features are still frequently used in health text classification, e.g., detecting Alzheimers disease (AD) (Masrani et al., 2017; Zhu et al., 2018; Balagopalan et al., 2018), aphasia (Fraser et al., 2015), or sentiment from language (Maas et al., 2011); and (2) understanding the importance of syntactic and lexical information separately as well as interactively is still an open research area in linguistics (Lester et al., 2017; Blaszczak, 2019).

Lexical richness and complexity relate to nuances and the intricacy of meaning in language. Numerous metrics to quantify lexical diversity, such as type-token ratio (TTR) (Richards, 1987)

and MLTD (McCarthy, 2005), have been proposed. These metrics capture various dimensions of meaning, quantity and quality of words, such as variability, volume, and rarity. Several of these have been identified to be important for a variety of tasks in applied linguistics (Daller et al., 2003). For example, metrics related to vocabulary size, such as TTR and word-frequencies, have proven to help with early detection of mild cognitive impairment (MCI) (Aramaki et al., 2016), hence are important for early dementia diagnosis. Discourse informativeness, measured via propositional idea density, is also shown to be significantly affected in speakers with aphasia (Bryant et al., 2013). Furthermore, lexicon-based methods have proved to be successful in sentiment analysis (Taboada et al., 2011; Tang et al., 2014).

Syntactic complexity is evident in language production in terms of syntactic variation and sophistication or, in other words, the range and degree of sophistication of the syntactic structures that are produced (Lu, 2011; Ortega, 2003). This construct has attracted attention in a variety of language-related research areas. For example, researchers have examined the developmental trends of child syntactic acquisition (e.g., (Ramer, 1977)), the role of syntactic complexity in treating syntactic deficits in agrammatical aphasia (e.g., (Melnick and Conture, 2000; Thompson et al., 2003)), the relationship between syntactic complexity in early life to symptoms of Alzheimers disease in old age (e.g., (Kemper et al., 2001; Snowdon et al., 1996)), and the effectiveness of syntactic complexity as a predictor of adolescent writing quality (e.g., (Beers and Nagy, 2009)).

Indefrey et al. (2001) reported data on brain activation during syntactic processing and demonstrated that syntactic processing in the human brain happens independently of the processing of lexical meaning. These results were supported

431

by the more recent studies showing that different brain regions support distinct mechanisms in the mapping from a linguistic form onto meaning, thereby separating syntactic agrammaticality from linguistic complexity(Ullman et al., 2005; Friederici et al., 2006). This motivates us to explore the importance of lexical and syntactic features separately.

To our knowledge, there is no previous research in medical text classification area exploring the individual value of lexical and syntactic features with regards to their vulnerability and importance for ML models. Syntactic and lexical feature groups are often used together without specifying their individual value. For example, recent work in text classification for AD detection revealed that a combination of lexical and syntactic features works well (Fraser et al., 2016; Noorian et al., 2017); the same is true for other cognitive disease or language impairment detection (Meteyard and Patterson, 2009; Fraser et al., 2014), as well as sentiment detection in healthy speech and language (Negi and Buitelaar, 2014; Marchand et al., 2013; Pang et al., 2002).

In this paper, we focus on individual value of lexical and syntactic feature groups, as studied across medical text classification tasks, types of language, datasets and domains. As such, the main contributions of this paper are:

- Inspired by the results of neuroscience studies (Indefrey et al., 2001), we explore selective performance of lexical and syntactic feature groups separately.

- We demonstrate, using multiple analysis methods, that there is a clear difference in how lexical features endure text alterations in comparison to the syntactic ones as well as how the latter impact classification.

- We report results on three different datasets and four different classifiers, which allows us to draw more general conclusions.

- We conduct an example-based analysis that explains the results obtained during the analysis.

2 Related Work

Prior research reports the utility of different modalities of speech – lexical and syntactic (Bucks et al., 2000; Fraser et al., 2016; Noo-

rian et al., 2017; Zhu et al.) – in detecting dementia. Bucks et al. (2000) obtained a cross-validated accuracy of 87.5% among a sample of 24 participants in detecting AD using eight linguistic features, including part-of-speech (POS) tag frequencies and measures of lexical diversity. A similar feature set was employed by Meilán et al. (2014) in a larger sample, where measures of lexical richness were less useful than features indicative of word finding difficulty (such as pauses and repetitions). Orimaye et al. (2014) obtained F-measure scores up to 0.74 using a combination of lexical and syntactic features on transcripts from a large dataset of AD and controls speech, DementiaBank (see Section 3.1)

Similarly, varying feature sets have been used for detecting aphasia from speech. Researchers have studied the importance of syntactic complexity indicators such as Yngve-depth and length of various syntactic representations for detecting aphasia (Roark et al., 2011), as well as lexical characteristics such as average frequency and the imageability of words used (Bird et al., 2000). Patterns in production of nouns and verbs are also particularly important in aphasia detection (Wilson et al., 2010; Meteyard and Patterson, 2009). Fraser et al. (2014) used a combination of syntactic and lexical features with ASR-transcription for the diagnosis of primary progressive aphasia with a cross-validated accuracy of 100% within a dataset of 30 English-speakers. More recently, Le et al. (2017) proposed methods to detect paraphasia, a type of language output error commonly associated with aphasia, in aphasic speech using phone-level features.

Sentiment analysis methodologies often use lexicon-based features (Taboada et al., 2011; Tang et al., 2014). Syntactic characteristics of text such as proportions of verbs and adjectives, nature of specific clauses in sentences are also salient in sentiment detection (Chesley et al., 2006; Meena and Prabhakar, 2007). Additionally, systems using both syntactic and lexical features have been proposed in prior work (Negi and Buitelaar, 2014; Marchand et al., 2013). For example, Marchand et al. (2013) trained ML models on patterns in syntactic parse-trees and occurrences of words from a sentiment lexicon to detect underlying sentiments from tweets while Negi and Buitelaar (2014) employed syntactic and lexical features for sentence level aspect based sentiment analysis. Pang et al.

		Datasets		
		DemB	IMDBs	AphB
Task nature	Structured	X	X	
	Partially structured			X
Language type	Verbal	X		X
	Written		X	
Lexics	Complex			X
	Medium		X	
	Simple	X		
Syntax	Complex		X	
	Medium	X		
	Simple			X

Table 1: Comparison of the datasets in terms of task nature, type of language used to collect the data, lexical and syntactic complexity.

(2002) showed that unigrams, bigrams and frequencies of parts-of-speech tags such as verbs and adjectives are important for an ML-based sentiment classifier.

3 Method

3.1 Datasets

In the following section, we provide context on each of three similarly-sized datasets that we investigate that differ in the following ways (see also Section 4):

1. Binary text classification task (AD detection, sentiment classification, aphasia detection).

2. Type of language

3. Level of lexical and syntactic complexity.

3.1.1 DementiaBank (DemB)

DementiaBank[1] is the largest publicly available dataset for detecting cognitive impairments, and is a part of the TalkBank corpus (MacWhinney, 2007). It consists of audio recordings of verbal descriptions and associated transcripts of the Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (Becker et al., 1994) from 210 participants aged between 45 to 90. Of these participants, 117 have a clinical diagnosis of AD ($N = 180$ speech recordings), while 93 ($N = 229$ speech recordings) are cognitively healthy. Many participants repeat the task within an interval of a year.

3.1.2 AphasiaBank (AphB)

AphasiaBank[2] (MacWhinney, 2007) is another dataset of pathological speech that consists of

aphasic and healthy control speakers performing a set of standard clinical speech-based tasks. The dataset includes audio samples of speech and associated transcripts. All participants perform multiple tasks, such as describing pictures, storytelling, free speech, and discourse with a fixed protocol. Aphasic speakers have various sub-types of aphasia (fluent, non-fluent, etc.). In total, there are 674 samples, from 192 healthy ($N = 246$ speech samples) and 301 ($N = 428$ speech samples) aphasic speakers.

3.1.3 IMDB Sentiment Extract (IMDBs)

The IMDB Sentiment (Maas et al., 2011) dataset is a standard corpus for sentiment detection that contains typewritten reviews of movies from the IMDB database along with the review-associated binary sentiment polarity labels (positive and negative). This dataset is used in order to extend the range of 'healthy' language and test generalizability of our findings. The core dataset consists of 50,000 reviews split evenly into train and test sets (with equal classes in both train and test). To maintain a comparable dataset size to DemB and AphB, we randomly choose 250 samples from the train sets of each polarity, totalling 500 labeled samples.

All the three datasets cover a breadth of transcripts in terms of presence or absence of impairment, as well as a spectrum of 'healthy' speech.

3.2 Feature Extraction

Following multiple previous works on text classification, we extract two groups of linguistic features – lexical and syntactic.

Lexical features: Features of lexical domain have been recognized as an important construct in a number of research areas, including stylistics, text readability analysis, language assessment, first and second language acquisition, and cognitive disease detection. In order to measure various dimensions of lexical richness in the datasets under comparison, we compute statistics on token/unigram, bigram, and trigram counts. Additionally, we use the Lexical Complexity Analyser (Ai and Lu, 2010) to measure various dimensions of lexical richness, such as lexical density, sophistication, and variation.

Following Oraby et al. (2018), Dušek et al. (2019), and Jagfeld et al. (2018), we also use Shannon entropy (Manning and Schtze, 2000,

---

[1]https://dementia.talkbank.org
[2]https://aphasia.talkbank.org

p. 61ff.) as a measure of lexical diversity in the texts:

$$H(\text{text}) = - \sum_{x \in \text{text}} \frac{\text{freq}(x)}{\text{len}(\text{text})} \log_2 \left(\frac{\text{freq}(x)}{\text{len}(\text{text})} \right) \tag{1}$$

Here, x stands for all unique tokens/n-grams, *freq* stands for the number of occurrences in the text, and *len* for the total number of tokens/n-grams in the text. We compute entropy over tokens (unigrams), bigrams, and trigrams.

We further complement Shannon text entropy with n-gram conditional entropy for next-word prediction (Manning and Schtze, 2000, p. 63ff.), given one previous word (bigram) or two previous words (trigram):

$$H_{cond}(\text{text}) = - \sum_{(c,w) \in \text{text}} \frac{\text{freq}(c,w)}{\text{len}(\text{text})} \log_2 \left(\frac{\text{freq}(c,w)}{\text{freq}(c)} \right) \tag{2}$$

Here, (c, w) stands for all unique n-grams in the text, composed of c (context, all tokens but the last one) and w (the last token). Conditional next-word entropy gives an additional, novel measure of diversity and repetitiveness: the more diverse text is, the less predictable is the next word given the previous word(s) is; on the other hand, the more repetitive the text, the more predictable is the next word given the previous word(s).

Syntactic Features: We used the D-Level Analyser (Lu, 2009) to evaluate syntactic variation and complexity of human references using the revised D-Level Scale (Lu, 2014).

We use the L2 Syntactic Complexity Analyzer (Lu, 2010) to extract 14 features of syntactic complexity that represent the length of production units, sentence complexity, the amount of subordination and coordination, and the frequency of particular syntactic structures. The full list of lexical and syntactic features is provided in Appendix A.

3.3 Classification Models

We benchmark four different machine learning models on each dataset with 10-fold cross-validation. In cases of multiple samples per participant, we stratify by subject so that samples of the same participant do not occur in both the train and test sets in each fold. This is repeated for each text alteration level. The minority class is over-sampled in the training set using SMOTE (Chawla et al., 2002) to deal with class imbalance.

Feature subgroup	Feature	DemB	IMDBs	AphB
Lexical richness	distinct tokens occuring once, %	0.58	**0.64**	0.32
	distinct bigrams occuring once, %	0.89	**0.95**	0.83
	distinct trigrams occuring once, %	0.96	**0.99**	0.92
Lexical complexity	unigram entropy	5.42	6.53	**6.70**
	bigram entropy	6.4	7.46	**8.68**
	trigram entropy	6.55	7.54	**9.19**
	bigram conditional entropy	1.01	0.95	**1.99**
	trigram conditional entropy	0.16	0.09	**0.51**
	lexicon complexity	1.33	**1.47**	1.32
Length of production unit	Mean length of clause	7.45	**9.24**	5.42
	Mean length of sentence	8.77	**21.42**	6.01
	Mean length of T-unit	11.85	**18.99**	6.15
Sentence complexity	Clauses per sentence	1.21	**2.35**	1.08
	D-level 0	0.63	0.26	**0.74**
	D-level 1-4	**0.23**	0.21	0.14
	D-level 5-7	0.14	**0.52**	0.11
Amount of subordination	Clauses per T-unit	1.62	**2.07**	1.12
	Complex T-units per T-unit	0.19	**0.55**	0.14
	Dependent clauses per T-unit	0.68	**1.00**	0.19
Amount of coordination	Coordinate phrases per clause	0.11	**0.22**	0.10
	Coordinate phrases per T-unit	0.17	**0.44**	0.11
	T-units per sentence	0.77	**1.13**	0.95
Particular structures	Complex nominals per clause	0.64	1.09	0.33
	Complex nominals per T-unit	1.03	**2.28**	0.38
	Verb phrases per T-unit	1.93	**2.64**	1.19

Table 2: Lexical complexity and richness, and syntactic complexity of the three datasets. Counts for n-grams appearing only once are shown as proportions of the total number of respective n-grams. Highest values on each line are typeset in bold.

We consider Gaussian naive Bayes (with equal priors), random forest (with 100 estimators and maximum depth 5), support vector Machine (with RBF kernel, penalty $C = 1$), and a 2-hidden layer neural network (with 10 units in each layer, ReLU activation, 200 epochs and Adam optimizer) (Pedregosa et al., 2011). Since the datasets have imbalanced classes, we identify F1 score with macro averaging as the primary performance metric.

3.4 Altering Text Samples

There can be three types of language perturbations at the word level: insertions, deletions, and substitutions on words. (Balagopalan et al., 2019) showed that deletions are more affected (significantly) than insertions and substitutions, so we likewise focus on deletions. Following Balagopalan et al. (2019), we artificially add deletion errors to original individual text samples at predefined levels of 20%, 40%, 60%, and 80%. To add the errors, we simply delete random words from original texts and transcripts at a specified rate.

3.5 Evaluating Change of Feature Values

In order to evaluate the change of feature values for different levels of text alterations, z-scores are used. We calculate z-scores of each individual feature in the transcripts with each level of alteration,

with relation to the value of that feature in the original unaltered transcript.

$$Z^x_{feat} = (feat^x - \mu_{no-alteration})/\sigma_{no-alteration}, (3)$$

where $feat^x$ refers to a given syntactic or lexical feature extracted from a transcript with an alteration level of $x = 20..80$, μ and σ are computed over the entire original unaltered dataset.

Then, we average the individual z-scores across all the features within each feature group (syntactic and lexical) to get a z-score per feature group.

$$Z^x_{syntactic} = \frac{1}{N_{syn}} \sum_{i=1}^{N_{syn}} Z^x_{feat_i} \quad (4)$$

$$Z^x_{lexical} = \frac{1}{N_{lex}} \sum_{i=1}^{N_{lex}} Z^x_{feat_i}, \quad (5)$$

where N_{syn} and N_{lex} refer to the total number of syntactic and lexical features, respectively.

3.6 Evaluating change of feature predictive power

We extract $\Delta F1_x$, or change in classification F1 macro score, with $x\%$ alteration with respect to no alteration, for $x = 20, 40, 60, 80$, i.e,

$$\Delta F1_x = F1_{x\%alteration} - F1_{no-alteration}. \quad (6)$$

To identify the relative importance of syntactic or lexical features on classification performance, we estimate coefficients of effect for syntactic and lexical features. These coefficients are obtained by regressing to F1 deltas using the syntactic and lexical feature z-scores described in Section 3.5 for each alteration level. Thus, the regression equation can be expressed as:

$$\Delta F1 = \alpha Z_{syntactic} + \beta Z_{lexical}. \quad (7)$$

The training set for estimating α and β consists of $\Delta F1_x; (Z^x_{syntactic}, Z^x_{lexical})$ for $x = 20, 40, 60, 80$.

4 Comparing datasets

Three datasets used in our exploration represent different dimensions of lexical and syntactic complexity, and are unique in the nature of the tasks they involve and their type of language, as shown in Tab.1. AphB is the only dataset that includes speech samples of unstructured speech, while

Dataset	Level (%) of alterations	Lexical features (z-score)	Syntactic features (z-score)
DemB	20	**0.35**	0.30
	40	**0.75**	0.62
	60	**1.21**	0.94
	80	**1.85**	1.31
AphB	20	0.10	**0.15**
	40	**0.26**	0.26
	60	**0.51**	0.37
	80	**0.93**	0.51
IMDBs	20	**0.29**	0.13
	40	**0.61**	0.25
	60	**1.00**	0.35
	80	**1.61**	0.31

Table 3: Change of feature values, per dataset and per level of text alterations.

IMDBs is unique as it contains samples of written language, rather than transcripts of verbal speech.

In terms of lexical and syntactic complexity, it is interesting to note that AphB contains samples that are most lexically complex, while at the same time it is the most simple from the syntactic point of view. We associate this with the fact that AphB data come from partially unstructured tasks, where free speech increases the use of a more complex and more diverse vocabulary. IMDB is the most lexically rich dataset (see Table 2), with the highest ratio of uni-, bi-, and trigrams occuring only once.

IMDB is the most complex according to various measures of syntactic complexity: it has the highest scores with metrics associated with length of production unit, amount of subordination, coordination, and particular structures, and it also has the highest amount of complex sentences (sentences of D-level 5-7, as shown in Table 2). This may be explained by the fact it is the only dataset based on typewritten language. AphB has the lowest level of syntactic complexity, containing the highest amount of the simplest sentences (D-level 0), and lowest scores in other subgroups of syntactic features (see Table 2).

Next, we analyse if these variously distinct datasets have any common trend with regards to the vulnerability and robustness of lexical and syntactic feature groups.

5 Results and discussion

5.1 Feature vulnerability

Following the method described in Section 3.5, we analyse if any of the feature groups (lexical or syntactic) is influenced more by text alter-
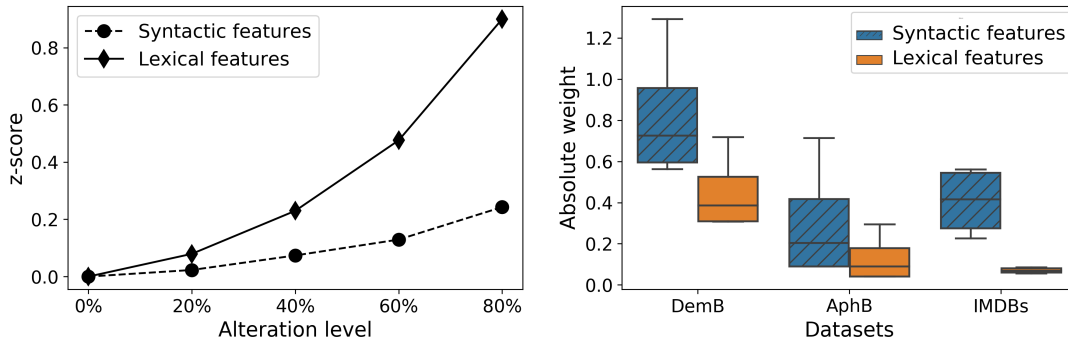
Figure 1: Left: Change of syntactic and lexical feature values at different alteration levels, averaged across three datasets. Right: Impact of syntactic and lexical features on classification for DementiaBank, AphasiaBank and IMDBsentiment datasets, averaged across fours classifiers.

ations. As shown in Figure 1, the values of lexical features are, on average, influenced significantly more than syntactic ones (Kruskal-Wallis test, $p <$0.05). Such a difference is observed in all three datasets individually (see Table 3).

The differences of z-scores between lexical and syntactic feature groups are higher for the IMDBs dataset, which suggests that the difference is most visible either in healthy or in written language.

These results suggest that lexical features are more vulnerable to simple text alterations, such as introduced deletion errors, while syntax-related features are more robust to these modifications. However, stronger changes of raw feature values do not necessarily mean that the resulting modified features become more or less important for classifiers. This leads us to inspect the impact of text alteration on feature predictive power.

5.2 Feature significance and the impact of alterations on feature predictive power

A simple method to understand the potential predictability of a feature is by looking at how different the feature value is between classes and whether this difference is statistically significant. This method was previously used in studies assessing automatic speech recognition for Alzheimer's (Zhou et al., 2016) and aphasia detection (Fraser et al., 2013).

We rank the p-values obtained, in each condition, from a two tailed non-parametric Kruskal-Wallis test performed on each feature between the two classes (healthy vs unhealthy in the DB and AphB datasets, and positive vs negative in IMDBs) and assign rank to each feature. It is interesting to note that lexical features occupy the overwhelming majority of first places across all

Dataset	Classifiers			
	NN	**SVM**	**RF**	**NB**
DemB	1.82	1.83	1.98	1.80
IMDBs	5.22	6.39	7.15	3.74
AphB	2.22	2.44	2.28	2.17

Table 4: Ratio of coefficients, calculated as $Importance_{syntactic}/Importance_{lexical}$. Ratio higher than one indicates that syntactic features are more important for a classifier than lexical ones.

datasets, showing that lexical features are significantly different between classes. We further analyse, following (Brunato et al., 2018), how the rank of each feature changes when different levels of text alterations are introduced. The maximum rank increase is higher on average for lexical features than for syntactic (see Figure 2 for details of rank changes in DemB dataset) across all datasets. The ratio of features that become insignificant after text alteration is also higher for lexical features rather than in syntactic on average across all datasets. As Figure 2 shows, the features with increased rank are those that were not initially significantly different between classes. The combination of these results suggest that not so important lexical features become more and more important with addition of text alterations, which may decrease the performance of classification.

The above method of calculating p-values is analogous to feature selection performed as a pre-processing step before classification. Although this step may provide some initial insights into feature significance, it does not guarantee the most significant features will be those having the most predictive power in classification.

436

Lexical features	20%	40%	60%	80%
slextypes	-14	-45	-31	-24
swordtypes	-21	-37	-29	-26
rttr	-14		-35	-20
cttr	-12		-33	-18
lextypes		-20	-4	0
slextokens		-23	-15	-14
wordtypes		-16	-6	-1
ndw		-16	-6	-1
distinct_bigrams		-28	-15	-7
distinct_tokens		-10	-1	4
ls2	0	-31	-43	-45
swordtokens		-24	-27	-28
uber	8	-14	-37	
entropy_2gram		-8	4	7
entropy_1gram		-16	-8	4
lextokens		4	12	13
distinct_trigrams		-11	-3	2
bigrams		-8	0	6
entropy_3gram		4	14	15
trigrams		1	4	7
ndwerz			-19	4
ls1	10	-19	-30	

Syntactic features	20%	40%	60%	80%
T/S	-33	-22	-36	-57
C/S	-34	-27	-24	-42
C		-52	-58	-56
CT		-42	-45	-46
T			-49	-49
MLS	-6	-1	-1	-1
CP	-20			
CN		-40	-36	-29
VP		-32	-34	-35
CT/T		-34	-35	-37
C/T	-22	-11	-17	-30
DC		-16	-26	-26
S				-46
CP/T	-21	-45	-51	-55
CP/C	-14	-35	-46	-50
MLT	-8	-5	-2	-2
CN/T		-44	-43	-36
DC/T	-24	-17	-28	-32
CN/C	-25		-47	-44
DC/C		-21	-30	-33
MLC	-18	-18	-19	-27
VP/T	0	0	2	-4

Figure 2: Change of lexical (left) and syntactic (right) feature rank when text alterations of different levels are introduced. Negative numbers denote decrease in rank, and positive numbers are an increase of rank. Blue cell colours denote the highest increase in rank, red (the highest decrease) and yellow (a smaller level of increase or decrease). Features are ranked based on p-values with the lowest p-value at the top. White cells show that features were not significantly different between classes in the original text samples, based on DemB dataset.

We use the method described in Section 3.6 to evaluate the impact of text alteration on the features predictive power. The results in Table 4 show that syntactic features have more predictive power than lexical features. The lowest ratio is observed with DemB, and the AphB results are very close, suggesting that syntactic features are approximately twice as important than lexical features in predicting pathological speech. In healthy written language, the difference is even higher and reaches 7.15 for the random forest classifier.

In summary, the predictive power of syntactic features is much stronger than that of lexical features across three datasets and four main classifiers, which suggest the results can be generalizable across several different tasks and domains.

5.3 Example-based Analysis

As shown in previous sections, values of lexical features are on average more influenced by text alterations but this change does not affect classification as much as smaller value changes in syntactic features. Table 5 provides examples of two features, one lexical and one syntactic, their value

changes when text samples are modified, and the associated change of the classifier's predictions.

The value of lexical feature *cond_entropy_3gram*, showing conditional entropy calculated for trigrams, decreases by more than 50% when the text sample is modified by only 20%. This change is much higher than the associated absolute change of the syntactic feature *C/S* (that shows the number of clauses per sentence) that increases by 11% only on the same level of alteration. The prediction made by a classifier in the case of the lexical feature, however, is the same as the prediction of original transcript. Only when the general level of alteration reaches 60% and the value of the lexical feature decreases by more than 85%, the prediction becomes incorrect. In the case of syntactic features, the prediction already changes to incorrect with the general level of alteration of 20%, although the feature value is still quite close to the original one.

Consider this sentence in the original transcript:

Table 5 content:

Alteration level	Text sample	Feature	Feature value / △	Prediction	Dataset
original	&uh the boy is reaching into the cookie jar. he's falling off the stool. the little girl is reaching for a cookie. mother is drying the dishes. the sink is running over. mother's getting her feet wet. they all have shoes on. there's a cup two cups and a saucer on the sink. the window has draw withdrawn drapes. you look out on the driveway. there's kitchen cabinets. oh what's happening. mother is looking out the window. the girl is touching her lips. the boy is standing on his right foot. his left foot is sort of up in the air. mother's right foot is flat on the floor and her left she's on her left toe. &uh she's holding the dish cloth in her right hand and the plate she is drying in her left. I think I've run out of. yeah.	lexical (cond_entropy_3gram)	0.24 / -	Correct (healthy)	DemB
20%	&uh the boy reaching the cookie jar. he's falling off the stool. the little girl is reaching for cookie. mother is the dishes. the sink is over. mother's getting her feet. all have shoes. there's cup two cups a saucer on sink. window has draw withdrawn drapes. you look out on driveway. there's kitchen cabinets. oh what's happening. mother out the window. the girl is lips. the boy standing on. his left foot is sort of up in the air. mother's right foot is flat on the floor and left she's on her left toe. &uh she's holding the cloth in right hand the plate she drying in her left. think I've run out of.	lexical (cond_entropy_3gram)	0.11 / 0.48	Correct (healthy)	DemB
40%	&uh reaching the jar. he's falling the stool. the little is reaching a cookie. mother drying the dishes. the sink is running over. mother's her wet. all have shoes on. a two and a sink. the. you look driveway. there's kitchen. oh what's happening. mother out the window. the is her. is his foot. his left foot is sort of up air. foot is flat floor and she's her toe. &uh she's holding the dish cloth in right the she is drying in left. I think of.	lexical (cond_entropy_3gram)	0.07 / 0.28	Correct (healthy)	DemB
60%	&uh is cookie. falling stool. for cookie. the dishes. the. mother's feet wet. they have. a two cups a sink. the has withdrawn drapes. the. there's. oh. mother the window. the lips. the boy right. is sort of. right foot is flat on floor on her left. &uh cloth right hand and the she is in her left. yeah.	lexical (cond_entropy_3gram)	0.03 / 0.14	Incorrect (AD)	DemB
original	okay. well in the first place the the mother forgot to turn off the water and the water's running out the sink. and she's standing there. it's falling on the floor. the child is got a stool and reaching up into the cookie jar. and the stool is tipping over. and he's sorta put down the plates. and she's reaching up to get it but I don't see anything wrong with her though. yeah that's it. I can't see anything.	syntactic (C/S)	1.1 / -	Correct (healthy)	DemB
20%	well the first the the mother forgot to turn off the water the water's out the sink. and standing there. it's falling floor. it's got a stool and into the cookie jar. and the stool is tipping. and he's sorta down the plates. and she's reaching to get it but I don't see anything wrong with her though. that's it. I can't see anything.	syntactic (C/S)	1.22 / 1.11	Incorrect (AD)	DemB
40%	okay. well in the forgot the water the water's out the sink. and she's standing there. it's on the. the is got a stool and reaching up the. the is tipping. and he's sorta the. and she's reaching up to get but I her. yeah that's. I can't.	syntactic (C/S)	1.0 / 0.91	Incorrect (AD)	DemB
60%	okay. in water's out the sink. falling. the got stool the cookie jar. and the stool is over. and he's down the plates. and she's up but don't wrong. can't see anything.	syntactic (C/S)	1.0 / 0.91	Incorrect (AD)	DemB

Table 5: Examples of two features, *cond_entropy_3gram* and *C/S*, their value change when text samples are modified on the level of 20%, 40% and 60%, and associated classifier's predictions. Examples are provided using the DemB transcript samples and feature values.

She's holding the dish cloth in her right hand and the plate she is drying in her left.

With 20% of errors it is converted to the following:

She's holding the cloth in right hand the plate she drying in her left.

It is clear that lexical features based on the frequency of uni-, bi- and trigrams are affected by this change, because quite a few words disappear in the second variant. In terms of syntactic structures, however, the sentence is not damaged much, as we still can see the same number of clauses, coordinate units, or verb phrases. Such an example helps explain the results in the previous sections.

6 Conclusions and Future Research

This paper shows that linguistic features of text, associated with syntactic and lexical complexity, are not equal in their vulnerability levels, nor in their predictive power. We study selective performance of these two feature aggregations on three distinct datasets to verify the generalizability of observations.

We demonstrate that values of lexical features are easily affected by even slight changes in text, by analysing z-scores at multiple alteration levels. Syntactic features, however, are more robust to such modifications. On the other hand, lower changes of syntactic features result in stronger effects on classification performance. Note that these patterns are consistently observed across different datasets with different levels of lexical and syntactic complexity, and for typewritten text and transcribed speech.

Several methods to detect and correct syntactic (Ma and McKeown, 2012) and lexical errors (Klebesits and Grechenig, 1994) as a post-processing step for output from machine translation or ASR systems have been proposed in prior work. Since our analysis indicates that error-affected syntactic features have a stronger effect on classification performance, we suggest imposing higher penalties on detecting and correcting syntactic errors than lexical errors in medical texts. A limitation in our study is that we focused on text alterations of a specific type, and the results were only tested on relatively small datasets. In future work, we will extend the analysis to other simple text alterations such as substitutions as well as adversarial text attacks (Alzantot et al., 2018). In addition, we will extend the current work to see how state-of-the-art neural network models, such as Bert, can handle text alterations as they capture lexical, syntactic and semantic features of the input text in different layers. Finally, note that the datasets considered in this study are fairly small (between 500 and 856 samples per domain). Efforts to release larger and more diverse data sets through multiple channels (such as challenges) in such domains as Alzheimer's or aphasia detection, and depression detection (Valstar et al., 2016; MacWhinney, 2007; Mozilla, 2019) need to be reinforced.

References

Haiyang Ai and Xiaofei Lu. 2010. A web-based system for automatic measurement of lexical complexity. In *27th Annual Symposium of the Computer-Assisted Language Instruction Consortium. Amherst, MA*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896.

Eiji Aramaki, Shuko Shikata, Mai Miyabe, and Ayae Kinoshita. 2016. Vocabulary size in speech may be an early indicator of cognitive impairment. *PloS one*, 11(5):e0155195.

Aparna Balagopalan, Jekaterina Novikova, Frank Rudzicz, and Marzyeh Ghassemi. 2018. The Effect of Heterogeneous Data for Alzheimer's Disease Detection from Speech. In *Proceedings of the Machine Learning for Health (ML4H) Workshop at NeurIPS 2018*.

Aparna Balagopalan, Ksenia Shkaruta, and Jekaterina Novikova. 2019. Impact of ASR on Alzheimer's Disease Detection: All Errors are Equal, but Deletions are More Equal than Others. *arXiv:1904.01684*.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Scott F Beers and William E Nagy. 2009. Syntactic complexity as a predictor of adolescent writing quality: Which measures? which genre? *Reading and Writing*, 22(2):185–200.

Helen Bird, Matthew A Lambon Ralph, Karalyn Patterson, and John R Hodges. 2000. The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and language*, 73(1):17–49.

Joanna Blaszczak. 2019. *Investigation into the Interaction between the Indefinites and Negation*, volume 51. Walter de Gruyter GmbH & Co KG.

Dominique Brunato, Lorenzo De Mattei, Felice DellOrletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.

Lucy Bryant, Elizabeth Spencer, Alison Ferguson, Hugh Craig, Kim Colyvas, and Linda Worrall. 2013. Propositional idea density in aphasic discourse. *Aphasiology*, 27(8):992–1009.

Romola S Bucks, Sameer Singh, Joanne M Cuerden, and Gordon K Wilcock. 2000. Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini K Srihari. 2006. Using verbs and adjectives to automatically classify blog sentiment. *Training*, 580(263):233.

Helmut Daller, Roeland Van Hout, and Jeanine Treffers-Daller. 2003. Lexical richness in the spontaneous speech of bilinguals. *Applied linguistics*, 24(2):197–222.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2019. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Journal of Computer Speech & Language*, 59:123–156.

Kathleen Fraser, Frank Rudzicz, Naida Graham, and Elizabeth Rochon. 2013. Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of the fourth workshop on speech and language processing for assistive technologies*, pages 47–54.

Kathleen C Fraser, Naama Ben-David, Graeme Hirst, Naida Graham, and Elizabeth Rochon. 2015. Sentence segmentation of aphasic speech. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 862–871.

Kathleen C Fraser, Jed A Meltzer, Naida L Graham, Carol Leonard, Graeme Hirst, Sandra E Black, and Elizabeth Rochon. 2014. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *cortex*, 55:43–60.

Kathleen C Fraser, Frank Rudzicz, and Graeme Hirst. 2016. Detecting late-life depression in alzheimers disease through analysis of speech and language. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 1–11.

Angela D Friederici, Christian J Fiebach, Matthias Schlesewsky, Ina D Bornkessel, and D Yves Von Cramon. 2006. Processing linguistic complexity and grammaticality in the left frontal cortex. *Cerebral Cortex*, 16(12):1709–1717.

Peter Indefrey, Peter Hagoort, Hans Herzog, Rüdiger J Seitz, and Colin M Brown. 2001. Syntactic processing in left prefrontal cortex is independent of lexical meaning. *Neuroimage*, 14(3):546–555.

Glorianna Jagfeld, Sabrina Jenne, and Ngoc Thang Vu. 2018. Sequence-to-Sequence Models for Data-to-Text Natural Language Generation: Word- vs. Character-based Processing and Output Diversity. In *Proceedings of the 11th International Conference on Natural Language Generation*, Tilburg, The Netherlands. ArXiv: 1810.04864.

Susan Kemper, Marilyn Thompson, and Janet Marquis. 2001. Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. *Psychology and aging*, 16(4):600.

K Klebesits and Th Grechenig. 1994. Lexical error correction using contextual linguistic expectations. In *Computer Science 2*, pages 337–345. Springer.

Duc Le, Keli Licata, and Emily Mower Provost. 2017. Automatic paraphasia detection from aphasic speech: A preliminary study. In *Interspeech*, pages 294–298.

Nicholas Lester, Laurie Feldman, and Fermín Moscoso del Prado Martín. 2017. You can take a noun out of syntax...: Syntactic similarity effects in lexical priming.

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *Tesol Quarterly*, 45(1):36–62.

Xiaofei Lu. 2014. *Computational methods for corpus annotation and analysis*. Springer.

Wei-Yun Ma and Kathleen McKeown. 2012. Detecting and correcting syntactic errors in machine translation using feature-based lexicalized tree adjoining grammars. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 17, Number 4, December 2012-Special Issue on Selected Papers from ROCLING XXIV*, 17(4).

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

Brian MacWhinney. 2007. The talkbank project. In *Creating and digitizing language corpora*, pages 163–180. Springer.

Christopher D. Manning and Hinrich Schtze. 2000. *Foundations of statistical natural language processing*, 2nd printing edition. MIT Press, Cambridge, MA, USA.

Morgane Marchand, Alexandru Ginsca, Romaric Besançon, and Olivier Mesnard. 2013. [lvic-limsi]: Using syntactic features and multi-polarity words for sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 418–424.

Vaden Masrani, Gabriel Murray, Thalia Field, and Giuseppe Carenini. 2017. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In *BioNLP 2017*, pages 232–237, Vancouver, Canada,. Association for Computational Linguistics.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Arun Meena and TV Prabhakar. 2007. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *European Conference on Information Retrieval*, pages 573–580. Springer.

Juan José G Meilán, Francisco Martínez-Sánchez, Juan Carro, Dolores E López, Lymarie Millian-Morell, and José M Arana. 2014. Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.

Kenneth S Melnick and Edward G Conture. 2000. Relationship of length and grammatical complexity to the systematic and nonsystematic speech errors and stuttering of children who stutter. *Journal of Fluency Disorders*, 25(1):21–45.

Lotte Meteyard and Karalyn Patterson. 2009. The relation between content and structure in language production: An analysis of speech errors in semantic dementia. *Brain and language*, 110(3):121–134.

Mozilla. 2019. Common Voice dataset.

Sapna Negi and Paul Buitelaar. 2014. Insight galway: syntactic and lexical features for aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 346–350.

Zeinab Noorian, Chloé Pou-Prom, and Frank Rudzicz. 2017. On the importance of normative data in speech-based assessment. *arXiv preprint arXiv:1712.00069*.

Shereen Oraby, Lena Reed, and Shubhangi Tandon. 2018. Controlling Personality-Based Stylistic Variation with Neural Natural Language Generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, Melbourne, Australia.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for alzheimers disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 78–87.

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied linguistics*, 24(4):492–518.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Andrya LH Ramer. 1977. The development of syntactic complexity. *Journal of Psycholinguistic Research*, 6(2):145–161.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090.

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

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1555–1565.

Cynthia K Thompson, Lewis P Shapiro, Swathi Kiran, and Jana Sobecks. 2003. The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia. *Journal of Speech, Language, and Hearing Research*.

Michael T Ullman, Roumyana Pancheva, Tracy Love, Eiling Yee, David Swinney, and Gregory Hickok. 2005. Neural correlates of lexicon and grammar: Evidence from the production, reading, and judgment of inflection in aphasia. *Brain and Language*, 93(2):185–238.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10. ACM.

Stephen M Wilson, Maya L Henry, Max Besbris, Jennifer M Ogar, Nina F Dronkers, William Jarrold, Bruce L Miller, and Maria Luisa Gorno-Tempini. 2010. Connected speech production in three variants of primary progressive aphasia. *Brain*, 133(7):2069–2088.

Luke Zhou, Kathleen C Fraser, and Frank Rudzicz. 2016. Speech recognition in alzheimer's disease and in its assessment. In *INTERSPEECH*, pages 1948–1952.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. Detecting cognitive impairments by agreeing on interpretations of linguistic features. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, pages 1431–1441.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. 2018. Deconfounding age effects with fair representation learning when assessing dementia. *arXiv:1807.07217*.

A List of Linguistic Features

Lexical Feature	Description
distinct_tokens	Number of distinct tokens
distinct_tokens_ratio	Number of distinct tokens occuring once
bigrams	Number of distinct bigrams
distinct_bigrams	Number of distinct bigrams occuring once
distinct_bigrams_ratio	Ratio of distinct bigrams occuring once
trigrams	Number of distinct trigrams
distinct_trigrams	Number of distinct trigrams occuring once
distinct_trigrams_ratio	Ratio of distinct trigrams occuring once
entropy_1gram	Unigram entropy
entropy_2gram	Bigram entropy
entropy_3gram	Trigram entropy
cond_entropy_2gram	Conditional bigram entropy
cond_entropy_3gram	Conditional trigram entropy
wordtypes	Number of word types
swordtypes	Number of sophisticated word types
lextypes	Number of lexical types
slextypes	Number of sophisticated lexical word types
wordtokens	Number of word tokens
swordtokens	Number of sophisticated word tokens
lextokens	Number of lexical tokens
slextokens	Number of sophisticated lexical tokens
ld	Lexical density
ls1	Lexical sophistication I
ls2	Lexical sophistication II
vs1	Verb sophistication I
vs2	Verb sophistication II
cvs1	Corrected VS1
ndw	Number of different words
ndwz	NDW (first 50 words)
ndwerz	NDW (expected random 50)
ndwesz	NDW (expected sequence 50)
ttr	Type / token ratio
msttr	Mean segmental ttr (50)
cttr	Corrected ttr
rttr	Root ttr
logttr	Bilogarithmic ttr
uber	Uber coefficient

Syntactic Feature	Description
S	Number of sentences
VP	Number of verb phrases
C	Number of clauses
T	Number of T-units[3]
DC	Number of dependent clauses
CT	Number of complex T-units
CP	Number of coordinate phrases
CN	Number of complex nominals
MLS	Mean length of sentence
MLT	Mean length of T-units
MLC	Mean length of clause
C/S	Clauses per sentence
VP/T	Verb phrases per T-unit
C/T	Clauses per T-unit
DC/C	Dependent clauses per clause
DC/T	Dependent clauses per T-unit
T/S	T-units per sentence
CT/T	Complex T-units per T-unit
CP/T	Coordinate phrases per T-unit
CP/C	Coordinate phrases per clause
CN/T	Complex nominals per T-units
CN/C	Complex nominals per clause

---

[3]Here, T-unit is defined as the shortest grammatically allowable sentences into which writing can be split or minimally terminable unit. Often, but not always, a T-unit is a sentence.