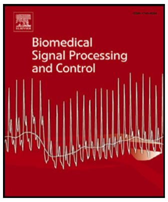
Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



The linguistic structure of an emotional text influences the sympathetic activity and the speech prosody of the reader

Benedetta Iavarone [a,b,*], Maria Sole Morelli [c], Dominique Brunato [b], Shadi Ghiasi [d], Enzo Pasquale Scilingo [d], Nicola Vanello [d], Felice Dell'Orletta [b], Alberto Greco [d]

[a] *Scuola Normale Superiore, P.za dei Cavalieri, 7, 56126, Pisa, Italy*
[b] *Istituto di Linguistica Computazionale "Antonio Zampolli", Via G. Moruzzi, 1, 56124, Pisa, Italy*
[c] *Fondazione Toscana Gabriele Monasterio, Via G. Moruzzi, 1, 56124, Pisa, Italy*
[d] *Dipartimento di Ingegneria dell'Informazione, Research Center "E. Piaggio", University of Pisa, Via G. Caruso, 16, 56122, Pisa, Italy*

ARTICLE INFO

Keywords:
Speech analysis
Linguistic profile
Electrodermal activity
Support Vector Regressor

ABSTRACT

In this study, we present an analysis of the relationship between the linguistic profile of a text and the physiological and acoustic characteristics of the reader to improve the emotion recognition systems. To this aim, we recorded the speech and electrodermal activity (EDA) signals from 33 healthy volunteers reading neutral and affective texts aloud. We used the BioVoice toolbox and cvxEDA algorithm to estimate some of the main speech and EDA features, respectively. The selected texts were analyzed to quantify their lexical, morpho-syntactic, and syntactic properties. Correlation and Support Vector Regression analyses between linguistic and speech and EDA features have shown a significant bidirectional association between the morpho-syntactic structure of the text and both sympathetic markers and voice acoustic properties. Specifically, significant relationships were observed between linguistic properties and certain EDA and speech features commonly used to evaluate human emotional state (e.g., edaSymp, mean tonic, F0). These findings suggest that lexical, morpho-syntactic, and syntactic properties may have a significant impact on an individual's emotional dynamics.

1. Introduction

Emotions play a crucial role in shaping our verbal communication and can greatly influence the effectiveness of speech and reading activities. Traditionally, assessing emotions has relied on subjective self-report measures, which may be limited by subjective biases and individual differences. However, technological advancements have enabled the integration of speech prosody and autonomic nervous system (ANS) correlates offering objective and reliable means to estimate emotional states.

ANS provides a physiological foundation for emotional regulation. Indeed, it is responsible for regulating bodily functions and plays a pivotal role in emotional responses [1]. EDA represents one of the most extensively studied ANS correlates of emotional arousal and measures change in skin's electrical conductance resulting from sweat gland activity regulated by the sympathetic branch of the ANS. EDA offers objective measures of emotional states, providing valuable insights into the physiological manifestations of emotions. The complex process that involves ANS and somatic regulation [2] is also responsible for speech production. Accordingly, variations in speech prosody can express and convey emotions [3], mood [4], stress [5,6], and personality [7] experienced by speakers or implied in the text during reading. Analyzing speech prosody [8] offers valuable insights into the emotional content [3] and intentions underlying verbal communication.

However, the linguistic structure of the pronounced text, including syntactic and semantic aspects, exerts a profound influence on speech prosody and, consequently, emotional expression. Therefore, investigating how the linguistic structure of the pronounced text influences speech prosody and ANS correlates provides a relevant pathway to understanding the interplay between language, emotions, and communication. Specifically, exploring how the linguistic structure of the pronounced text influences speech prosody, and ANS correlates, such as EDA, unravels the mechanisms through which language and emotions interact. This knowledge holds crucial implications across various practical applications. For instance, in the field of human–computer

* Corresponding author at: Scuola Normale Superiore, P.za dei Cavalieri, 7, 56126, Pisa, Italy.
E-mail addresses: benedetta.iavarone@sns.it (B. Iavarone), msmorelli@monasterio.it (M.S. Morelli), dominique.brunato@ilc.cnr.it (D. Brunato), shadi.ghiasi@centropiaggio.unipi.it (S. Ghiasi), enzo.scilingo@unipi.it (E.P. Scilingo), nicola.vanello@unipi.it (N. Vanello), felice.dellorletta@ilc.cnr.it (F. Dell'Orletta), alberto.greco@unipi.it (A. Greco).

https://doi.org/10.1016/j.bspc.2023.105776
Received 17 November 2022; Received in revised form 4 November 2023; Accepted 17 November 2023
Available online 28 November 2023

interaction, understanding how linguistic cues influence emotional responses can inform the design of emotionally intelligent systems that adapt to users' affective states. In healthcare settings, monitoring the emotional responses of patients during speech or reading tasks can aid in the diagnosis and treatment of emotional disorders. In order to capture the diverse layers of information embedded in a text, including linguistic, lexical, and stylistic aspects, increasingly sophisticated Natural Language Processing (NLP) and machine learning techniques have been developed. The advancements in these fields have led to the development of sophisticated techniques that enable the characterization of a text's linguistic profile by extracting a large number of features modeling underlying lexical, grammatical, and semantic phenomena [9]. Linguistic profiling has been successfully applied in various theoretical and application scenarios, such as automatically classifying textual genres and registers [10], as well as modeling cognitive aspects of human language. For instance, in [11], the authors have shown that linguistic features capturing lexical and (morpho-)syntactic properties of a sentence can be effectively used to predict the perception of its complexity by humans. Such evidence has been further confirmed by a subsequent study [12], which also proved the reliability of linguistic features extracted from the context in predicting humans' judgments of sentence complexity. Recently, Singh et al. [13] have proposed a deep learning hierarchy for emotion recognition, combining text analysis computed by the language model ELMo [14] with prosody, voice quality, and spectral features. However, formal modeling of the relationship between features describing linguistic profiles, ANS response, and speech prosody could unravel some specific mechanisms that influence a speaker's emotional response.

In this paper, we investigated whether physiological and acoustic features, commonly used to characterize ANS activity and speech production prosody, can be significantly modulated by the linguistic structure of the pronounced text. To this aim, we asked healthy volunteers to read emotional texts designed to evoke different levels of arousal and valence. Particularly, we used neutral and highly negative arousing text. Due to their contents and linguistic characteristics, the texts could elicit a sympathetic reaction from the subjects and modulate the speech signals (e.g., fundamental frequency and formants, speech duration). We analyzed a widely used sympathetic nervous system (SNS) correlate, electrodermal activity (EDA), to quantify the sympathetic reaction. This latter reflects the activity of the sympathetic nerve on sweat glands in terms of skin conductance changes [15]. As our main objective is to understand the interaction between the linguistic profile of the texts and the speech prosody and ASN correlates, in this work we do not focus on a comparison between the features extracted during the readings of the different texts [16]. Instead, we applied correlation and regression methods to understand how the features characterizing the linguistic profiles of a text interact and influence the speech prosody and the sympathetic response elicited by the same texts.

In addition, we performed a complementary analysis aimed at assessing the strength of this relationship but from the opposite perspective, namely by testing the feasibility of exploiting speech and physiological signals to predict a set of features characterizing the linguistic structure of the pronounced text. This analysis goes in the recent direction of using cognitive signals for the grounding of NLP models in multi-modal settings to improve their performance across multiple tasks and supply more cognitive-oriented benchmarks for their evaluation. To date, the vast majority of these studies have relied on eye-tracking data, which have been proven effective in many sequence labeling and sequence-to-sequence scenarios, such as sentiment analysis and irony detection, Part–of–speech (POS) tagging, Named Entity Recognition and relation extraction [17]. On the other hand, other sources of physiological data, such as ANS correlates, still need to be investigated more.

2. Methods

A group of 33 healthy volunteers was enrolled in the study (17 females, 16 males), aged between 26.6 and 30.0. None of them suffered from heart diseases, mental disorders, or phobias. Each participant gave their written informed consent, and the study was approved by the Ethical Committee of the University of Pisa.

We selected four texts, two describing medieval tortures and two describing textual genres and writing styles. Based on the topics covered, two texts were classified as high arousal and negative valence, whereas the other two were classified as neutral. Moreover, before starting the experiment, a group of 22 subjects, other than those enrolled in this study, evaluated the texts in terms of arousal and value, confirming the arousal and valence levels supposed a-priori based on the reading topic (see Supplementary Materials).

Each participant was asked to read aloud two randomly chosen texts, one neutral and one affective [16]. All texts have similar lengths to make the duration of the reading homogeneous among subjects. After each reading, each subject was asked to score the text in terms of arousal (from 1 to 5) and valence (from −2 to 2) using the Self-Assessment Manikin (SAM) model [18]. During the reading task, the speech signal and the EDA were recorded.

2.1. Linguistic analysis

The texts were divided into sentences, using the full stop as a splitting criterion, i.e., identifying a sentence as the part of text comprised between two full stops. After the splitting, neutral texts contained 25 sentences, with an average length of 28 tokens; affective texts contained 40 sentences, with an average of 21 tokens.

Each sentence was analyzed from a linguistic point of view and represented as a vector of ~140 features, corresponding to a subset of the features described in [9]. These features model a wide range of properties extracted from the text linguistically annotated according to the Universal Dependencies (UD)[1] formalism. Specifically, these features capture, on the one hand, complex phenomena related to the syntactic structure of a text (e.g., use of subordination, the average depth of the parse tree and the length of dependency relations, the structure of the verbal predicates) and the morpho-syntactic structure (e.g., distribution of grammatical categories across the text, fine-grained aspects related to verb inflection), on the other hand, they refer to raw text properties, like the length of the text and its fundamental components (sentences and words). We chose to rely on these features as they have already been proven to be effectively involved in modeling language processing effects during natural reading, such as the user's cognitive load inferred from eye-tracking data [19], as well as from explicit judgments of perceived sentence complexity given by human annotators [11]. Accordingly, we assumed they could also offer insights into the speaker's hidden emotional dynamics.

More in detail, the considered features, which were computed at sentence level, can be grouped into the following typologies according to the linguistic phenomena they describe:

(1) Raw Text Properties. Features about the length of the sentences and of the words therein contained. For each sentence, sentence length corresponds to the number of tokens comprised in the sentence, and word length to the average number of characters per token.

(2) Lexical Richness. Features on how varied the vocabulary of a text is, determined by computing: (i) the Type/Token Ratio (TTR), i.e., the ratio between the number of lexical types and the number of tokens within the sentence, (ii) the Distribution of words and lemmas belonging to a reference frequency dictionary of the Italian language, i.e., the Basic Italian Vocabulary (BIV Tok, BIV Types), also considering the internal repertories in which it is articulated (i.e., fundamental,

[1] https://universaldependencies.org/.

high usage and high availability lexicon) [20] (iii) the lexical density, calculated as the ratio of content words (nouns, verbs, adjectives, and adverbs) over the total number of words in the sentence;

(3) Morpho-syntactic information. Features on:

(*i*) the distribution of grammatical categories for each sentence (e.g., adjectives, nouns, determiners, pronouns);

(*ii*) the inflectional morphology, i.e., the distribution, for lexical verbs and auxiliaries, of a set of inflectional features (e.g., mood, tense, and person);

(4) Verbal Predicate Structure. Features on:

(i) the distribution of verbal heads, i.e., the average number of propositions (main or subordinate) co-occurring in a sentence;

(ii) the distribution of verbal roots, i.e., the percentage of verbal roots out of the total of sentence roots;

(iii) verb arity, i.e., the average number of instantiated dependency links sharing the same verbal head;

(5) Global and local parsed tree structures. Features on:

(i) the depth of the whole parse tree, i.e., calculated as the longest path (in terms of occurring dependency links) from the root of the dependency tree to some leaf;

(ii) the average clause length calculated as the number of tokens per clause, where the number of clauses is the ratio between the number of tokens in a sentence and the number of verbal or copular heads;

(iii) the average length of dependency links, i.e., the average number of words occurring between the syntactic head and its dependent(s);

(iv) the average depth of complement chains (a list of consecutive prepositional complements modifying a nominal head);

(v) the relative order of the subject and the object in a sentence with respect to their syntactic head (i.e., the verb).

(6) Syntactic relations. The percentage distribution of the 37 syntactic relations comprised in the UD annotation scheme (e.g., subject, object, nominal modifier).

(7) Subordination phenomena. Features on:

(i) the distribution of main vs. subordinate clauses;

(ii) the distribution of subordinates following or preceding the main clause;

(iii) the number of subordinates recursively embedded in the top subordinate clause.

2.2. Speech signal processing

To analyze the speech time series and extract acoustic parameters from each sentence, we used the BioVoice toolbox [21]. The toolbox detected first only voiced parts of each segment. These segments have mean duration 0.81 ± 0.90 in the first emotive text, 0.83 ± 0.81 in the second emotive text, 0.67 ± 0.75 in the first neutral text, and 0.64 ± 0.70 in the second neutral text. Then, F0, F1, F2 and F3 were calculated. In each voiced frame, F0 was estimated with a two-step procedure. First, Simple Inverse Filter Tracking (SIFT) was applied to signal time windows of length as $3Fs/Fmin$, where Fs is the signal sampling frequency and $Fmin$ is the minimum F0 value allowed for the signal under consideration (i.e., 40 Hz) [22]. Secondly, F0 was adaptively estimated on signal frames through the Average Magnitude Difference Function (AMDF) within the range provided by the SIFT [22]. Autoregressive Power Spectral Density (AR PSD) was considered to extract formants values over time. Furthermore, in each sentence, the total time duration of reading, the overall voiced duration within the sentence time window (*signal duration*), and the average duration of the voiced segments were extracted (*mean duration*).

Of note, due to their subject-dependency, the frequency features (F0, F1, F2, and F3) were scaled according to: $F_i^{scaled} = F_i/\overline{F0}_{neu}$ where F_i represents the frequency feature of interest (in neutral or emotional test in each sentence) and $\overline{Fi}_{neu}$ the mean of the frequency of the corresponding neutral texts, computed for all time duration.

2.3. EDA signal processing

The EDA represents changes in the skin conductance of the non-dominant hand due to the activity of the sweat gland. This latter is controlled directly by the activity of the ANS and, more specifically, of the SNS. The EDA comprises two components, the tonic and the phasic signals, which contain complementary information about the SNS. In particular, the tonic represents the EDA slow varying baseline and reflects the subjects' general psychophysiological state [23]. The phasic, instead, are relatively quick stimulus-evoked changes in the EDA signal [24]. In this study, to decompose the EDA signal into the phasic and tonic components, we applied the cvxEDA algorithm [25]. After the decomposition process, several features were extracted within the time window corresponding to each sentence: the mean (*mean ph, mean ton*), standard deviation (*std ph, std ton*), and maximum value (*max ph, max ton*) of both component; the number of phasic peaks (*no pks*) and the sum of their amplitudes (*sum pks*); the power spectrum within the 0.045–0.25 Hz interval (*edaSymp*), which reflects the sympathetic activity [26]. These features were normalized according to the time window length.

2.4. Statistical analysis and modeling

Using a Wilcoxon signed rank test, the SAM valence and arousal scores were statistically compared between the a priori neutral and negative texts. As a first statistical analysis of the features, we examined the relationship between linguistic features and both EDA and speech features. This investigation aimed to identify which linguistic properties of the text are most related to physiological arousal and speech production, thus allowing us to discover the underlying interaction between linguistic structure and SNS dynamics and speech. To do so, we computed the correlation between each linguistic feature and every EDA and speech one, using Spearman's correlation coefficient as the evaluation metric.

We selected all pairwise correlations that were statistically significant (with a p-value < 0.05 after FDR correction for multiple hypothesis testing [27]) and had a correlation coefficient different from zero. For each feature, we calculated the percentage of subjects for which the pairwise correlation was significant to deeply investigate whether some patterns are more stable across participants and which phenomena they involve.

We then assessed the relationship between linguistic, speech, and EDA features from a modeling standpoint. To this end, we devised two complementary tasks. The first task aimed to test the predictive strength of the features characterizing the linguistic profile. Specifically, we employed a Support Vector Regression (SVR). The SVR was implemented with a Radial Basis Function (RBF) kernel and standard parameters. It used all linguistic features as input and predicted the EDA and acoustic features. To account for within-subject repetitions, we used leave-one-subject-out (LOSO) cross-validation, training the model on all subjects minus one and testing on the left-out subject. Of note, the baseline was calculated by running the model with only the length of sentences as an input feature. The second task aimed at assessing whether and to which extent physiological and acoustic features can be effective predictors of features underlying the internal structure of a text. To this end, we built a regression model leveraging acoustic and EDA features to predict the whole set of linguistic parameters. As in the first scenario, we employed RBF-SVR and standard parameters. The model performance was compared to a baseline SVR model that used Voiced Duration as the sole input feature.

In addition for each model, we performed a feature importance analysis. Specifically, after the two aforementioned tasks were performed, we selected for each participant all the features for whom the predicted value by the RBF-SVR correlated for at least ±0.30 with their actual value and were statistically significant (p-value < 0.05). We then applied a SVR model with a linear kernel to predict each one of these selected features. As in the tasks previously described, we used all the EDA and speech features to predict the selected linguistic features, and all the linguistic features to predict the selected EDA and speech features. We then extracted the coefficients the linear-SVR assigned to the features used as predictors and used them to build the features rankings.

3. Results

3.1. SAM statistical analysis

The Wilcoxon test confirms the hypothesis of significant differences between the a priori neutral and negative texts used in the experiment. Remarkably, both the valence and the arousal scores were significantly different between the two classes of texts: the arousal score was significantly higher after the negative reading ($p < 0.01$), whereas the valence score was significantly lower ($p < 0.01$).

3.2. Correlation analysis

Tables 1 and 2 report an overview of the most significant results of, respectively, the correlations between speech frequency features and linguistic features and between EDA features and linguistic features (the complete results, including the mean correlation values, can be found in the Supplementary Materials to this paper). In both tables, linguistic features are grouped according to the linguistic phenomenon they describe. Notice that, in the tables, each cell does not report the value of correlation but the percentage of subjects for which the corresponding linguistic features in the group were significantly correlated with acoustic features and EDA features, independently from the correlation value.

Comparing the two tables, we observe that linguistic features belonging to the same group were significant for a similar and, in some cases, the very same number of subjects. Moreover, correlations with features encoding syntactic-related phenomena were, on average, more significant for a high number of subjects than correlations with lexical and morpho-syntactic features.

Focusing on the correlations between linguistic and acoustic features (Table 1), we can observe that *mean* and especially *signal duration* are the acoustic parameters reporting significant correlations with almost all the considered linguistic features for most subjects. Significant correlations for many subjects were also found for F0 and F3, while F1 and F2 were the least correlated. Focusing on the distinction between the different typologies of linguistic phenomena, as expected, acoustic features strongly related to the length of the sentences (Mean and Signal Duration) were consistently correlated with linguistic features that encode aspects of sentence length for most subjects. High correlations were also found with syntactic features regarding the use of subordination and the structure of the parsed tree. This result is especially true for F3, with up to 70% of the subjects showing a significant correlation. Notably, most linguistic features showing significant correlations are related to aspects of linguistic complexity spanning across different domains. In particular, beyond sentence length, which is considered as a shallow proxy of linguistic complexity and text readability [28], we observe significant correlations with properties of the syntactic structure (e.g., longer dependency links and prepositional chains) and verbal morphology (e.g., a past verbal tense may be perceived as more complex than the present tense). Conversely, features targeting the use of lexicon like the *lexical density*, turned out to be significantly correlated with acoustic parameters for very few subjects.

Concerning the correlations between EDA and linguistic features summarized in Table 2, *std ph*, *sum peaks* and *no peaks* are the three phasic features that report a large number of significant correlations with linguistic features for more than 50% of the subjects. Conversely, for *max peaks* and *mean pk*, fewer correlations were found for a more restricted number of participants. For the tonic component, the feature that reports the higher number of significant correlations is *std ton*, the standard deviation of the tonic component, which is especially highly correlated with linguistic features that describe syntactic phenomena. The other features of the tonic component, *max ton*, *mean ton*, and the feature of the power spectrum, *edaSymp*, are significant for fewer subjects. The highest number of significant correlations for these features is found with the group of linguistic features related to subordination. As in the previous case, if we take a closer look into the distinct groups of linguistic features, we notice that features encoding syntactic-related phenomena, especially related to the use of subordination, are overall more correlated than morpho-syntactic and especially lexical features.

Table 1
Summary results of the correlations between Speech Features and Linguistic Features. For each pairwise correlation, each number in the rows corresponds to the *percentage of subjects* for which the correlation was statistically significant (with a p-value < 0.05) and had a correlation coefficient different from zero. The cells where no number is available indicate that there were no subjects for whom that correlation was significant.

Linguistic feature	Speech					
	F0	F1	F2	F3	Mean duration	Signal duration
Raw text properties						
Sentence length	24	9	3	58	73	100
avg clause length	33	12	3	45	55	100
Lexical variety						
Lexical density	·	·	·	12	18	100
Morpho-syntactic information						
Auxiliary form	30	9	·	42	64	100
Auxiliary mood	33	9	·	39	58	100
Auxiliary person	30	12	3	45	58	100
Auxiliary tense	30	9	3	42	58	100
Adjective (possessive)	·	·	·	9	12	88
Adverb	·	·	·	6	9	70
Conjunction (coordinative)	·	·	·	6	9	79
Conjunction (subordinative)	·	·	·	6	12	79
Preposition	·	·	·	6	9	61
Article (determinative)	·	·	·	12	18	100
Article (indeterminative)	·	·	·	18	30	100
Noun (proper)	·	·	·	6	12	85
Verb (main)	·	·	·	12	21	100
Verbal predicate structure						
Verbal arity	61	36	21	73	97	100
Verbal roots dist.	33	12	3	45	58	100
Syntactic relations distributions						
Clausal modifier of noun	42	15	9	67	88	100
Adverbial clause modifier	36	18	9	61	82	100
Conjunct	39	15	12	64	85	100
Nominal modifier	36	12	6	58	82	100
Nominal subject	33	12	3	42	55	100
Passive nominal subject	36	21	9	55	82	100
Object	33	12	3	42	64	100
Oblique nominal	33	15	6	45	73	100
Global and local parsed tree structure						
avg dependency links length	33	12	3	45	55	100
avg prepositional chains length	45	30	15	70	91	100
Post-verbal object	39	27	12	67	91	100
Pre-verbal object	42	24	12	64	85	100
Post-verbal subject	42	24	9	64	85	100
Pre-verbal subject	42	21	9	64	85	100
Use of subordination						
Principals dist.	48	27	15	70	94	100
Subordinates dist.	52	27	15	70	97	100
Post-verbal subordinate	55	30	18	70	97	100
Pre-verbal subordinate	48	30	15	70	97	100

Table 2
Summary results of the correlations between EDA Features and Linguistic Features. For each pairwise correlation, each number in the rows corresponds to the *percentage of subjects* for which the correlation was statistically significant (with a p-value < 0.05) and had a correlation coefficient different from zero. The cells where no number is available indicate that there were no subjects for whom that correlation was significant.

Linguistic feature	Electrodermal Activity (EDA)								
	eda symp	Phasic component					Tonic component		
		max pks	no pks	sum pks	mean ph	std ph	max ton	mean ton	std ton
Raw text properties									
Sentence length	3	12	39	52	3	64	·	·	52
avg clause length	3	6	21	27	3	39	3	3	39
Lexical variety									
Lexical density	·	·	·	·	·	3	·	·	3
Morpho-syntactic information									
Auxiliary form	·	6	21	3	3	42	3	3	36
Auxiliary mood	·	6	18	24	3	36	3	3	33
Auxiliary person	3	6	21	3 ·	3	42	3	3	33
Auxiliary tense	·	6	18	21	3	36	3	3	36
Article (determinative)	·	·	·	·	·	6	·	·	·
Article (indeterminative)	·	·	·	·	·	6	·	·	9
Verb (main)	·	·	·	·	·	6	·	·	6
Verbal predicate structure									
Verbal arity	18	48	7	64	27	82	21	18	88
Verbal roots dist.	3	6	21	3	3	42	3	3	36
Syntactic relations distributions									
Clausal modifier of noun	12	21	42	45	12	58	9	9	70
Adverbial clause modifier	9	18	36	39	9	52	6	3	58
Conjunct	12	18	45	45	12	55	9	9	61
Nominal modifier	6	18	36	36	6	48	3	3	58
Nominal subject	3	6	18	3 ·	3	39	3	3	39
Passive nominal subject	9	18	42	36	9	52	6	3	55
Object	6	6	21	33	3	39	3	3	42
Oblique nominal	6	12	33	33	6	48	3	3	48
Global and local parsed tree structure									
avg dependency links length	3	6	18	3	3	39	3	3	39
avg prepositional chains length	15	3 ·	52	55	15	64	9	6	79
Post-verbal object	15	24	52	52	15	61	9	6	79
Pre-verbal object	9	24	52	45	9	55	6	6	67
Post-verbal subject	9	24	48	45	12	55	6	6	70
Pre-verbal subject	9	24	45	45	12	58	6	6	67
Use of subordination									
Principals dist.	15	39	64	58	18	79	12	12	88
Subordinates dist.	15	39	64	58	21	79	12	15	88
Post-verbal subordinate	15	45	64	58	24	82	15	15	88
Pre-verbal subordinate	12	42	64	58	21	82	12	15	88

3.3. SVR prediction of EDA and speech features using linguistic features as independent variables

In this section, we report the regression analysis results to investigate the feasibility of predicting the acoustic and the EDA features using our set of multi-level linguistic properties. To evaluate the goodness of the implemented SVR models, we correlated each model's predictions with the actual values of the features under examination, calculating the mean Spearman's correlation and variance over all subjects. For both types of signals, we present the results as percentages that indicate the number of subjects for which the predictions were significantly correlated, independently from the correlation value.

Table 3
Regression results for the prediction of speech features using linguistic features as independent variables. Highlighted in bold are the features that obtain a mean correlation value across subjects > 0.50.

	% significant subjects	Mean correlation	Correlation variance	Baseline
F0	15%	0.4032	0.0027	0.3622
F1	61%	**0.5419**	0.0181	−0.0272
F2	97%	**0.5424**	0.0089	0.0524
F3	27%	0.4593	0.0061	0.3264
Mean duration	91%	**0.5836**	0.0123	0.4399
Signal duration	100%	**0.9559**	0.0008	0.9447

Table 3 presents the results of predicting acoustic features using linguistic features. It includes the percentage of subjects that exhibit a significant correlation between the predicted variable and the target variable. Additionally, it provides the mean and variance of the correlation coefficient, as well as the correlation achieved by the baseline model. Of note, the predicting model always performed better than the baseline. The robustness of the model is also confirmed by the low correlation variance across subjects, indicating that the acoustic values predicted are consistent among the different subjects. The prediction of *mean duration* and *signal duration* was significant for almost every subject, as expected from the correlation results. Indeed, these features are directly linked to the length of sentences, a feature that the model could see in input. However, the model that uses all linguistic features slightly outperformed the baseline, especially for the prediction of *mean duration*, suggesting that also signals are influenced by text properties that go beyond sentence length. The predictions of F1 and F2 were significant for a large number of subjects (>60%). Contrary to what was seen previously in the correlation analysis, where F0 and F3 obtained significant results for a high number of subjects, when predicting them

Table 4
Regression results for the prediction of EDA features. Highlighted in bold are the features that obtain a mean correlation value across subjects > 0.50.

	% significant subjects	Mean correlation	Correlation variance	Baseline
edasymp	64%	**0.5033**	0.0082	0.0561
max_pks	33%	0.4836	0.0118	0.2790
no_pks	76%	**0.5394**	0.0103	0.4453
sum_pks	67%	**0.5357**	0.0184	0.3532
mean_ph	42%	0.2607	0.2291	0.0524
std_ph	82%	**0.5785**	0.0207	0.4947
max_ton	48%	0.1956	0.1982	0.0342
mean_ton	58%	0.1664	0.2455	0.0429
std_ton	73%	**0.5558**	0.0202	0.5066

with the SVR their predictions are significant for a low number of subjects (<30%).

Table 4 shows the results for the prediction of EDA features. Also in this case, our predictions are always higher than the baseline, which – as previously mentioned – corresponds to the scores of a SVR model using only the sentence's length in input. This is especially the case of *edaSymp* and *sum_pks*, which are among the best-predicted ones if compared to the baseline results. Nevertheless, the variance of the predictions is higher for some features (e.g., mean_ph, mean_ton) compared to the relatively low variance obtained by the model predicting speech features. We can also see that, except for *mean ph*, features referring to the phasic component are generally predicted with higher accuracy. On the contrary, both the mean and the maximum value of the tonic component showed to be less predictable by our set of linguistic features. This result reflects the correlation analysis results (see Table 2), where we observed that the pairwise correlations between these features and the whole set of linguistic features were significant for a lower percentage of subjects (from 3 to 30%, on average).

By inspecting the results of the feature importance analysis for the prediction of speech features, we found a clear impact coming from features connected to the length of the sentence. This includes not only the sentence length itself, but also other related aspects such as the number of verbal heads, as longer sentences tend to exhibit more clauses combined through coordination or subordination. We further observed that the distribution of subjects and their position (pre- or post-verbal) within the sentence turned out to be highly predictive. As regards the feature importance analysis for the prediction of the EDA, the tonic component does not exhibit a clear pattern of influence. Instead, the prediction of the EDA phasic component is found to be more related to the length of the sentence and features associated with it, such as the number of prepositional chains, and subordination phenomena. Additionally, punctuation seems to have some level of influence, although it is assumed to be related to the sentence length, as longer sentence tend to have more punctuation.

3.4. *SVR prediction of linguistic features using EDA and speech features as independent variables*

As the last analysis, we discuss the results of the prediction of linguistic features using speech features and EDA features as input. Also in this scenario, the goodness of the model is evaluated by correlating the model's predictions with the true values of the predicted features by calculating the mean Spearman's correlation and its variance for all subjects. These results are shown in Table 5, which reports only the features for which the number of significant subjects was ≥15. Complete results are shown in the Supplementary Materials of this paper.

We observed that the predictions of our model are always better than the baseline for all features. The very low variance of the correlation coefficients across the different subjects also shows that the model is quite robust in its predictions. The highest correlations are found for sentence length, as well as for features still related to length but modeling more complex properties of the global and local structure of the parse tree. These features include: the depth of the whole parse tree, the average length of dependency links, and the presence, and internal structure, of complex nominal complements headed by a preposition (i.e., *avg prepositional chains length, prepositional chain number*). These features are also the ones for which the correlations are significant for a high percentage of subjects (≥90%). Overall, considering the distinction of linguistic features into the different groups of phenomena, the best results are found in the predictions of features covering the use of subordination, for which the mean correlation is above 0.60, and the prediction is significant for almost all subjects. Conversely, EDA and voice features contribute to a small extent to the prediction of morpho-syntactic properties. Indeed, focusing on the distribution of grammatical categories, although the correlations are around 0.4 or above, these correlations are significant only for a few subjects. As shown in Table 5, the only two exceptions are represented by the presence of subordinating conjunctions and of auxiliaries in the present tense, which are significantly correlated for a high number of subjects (i.e., 20 out of 33 and 16 subjects out of 33, respectively).

The feature importance analysis reveals patterns of influence on the different type of linguistic features. Features regarding the lexical density have as most relevant predictors F1, F3, signal duration, and the EDA tonic component. In the case of morpho-syntactic features, the most influential predictors are F2 and signal duration, with no clear pattern observed from the EDA features. For what concerns syntactic relations features and the local and global parsed tree structure features, F2 and signal duration are identified as the most relevant predictors, but also the EDA phasic component exhibits a great impact. Lastly, in the context of subordination phenomena, signal duration emerges as the most relevant predictor, while no distinct pattern is observed from the EDA features.

4. Discussion and conclusion

In this study, we combined the analysis of the linguistic profile of neutral and emotional texts with the reader's EDA and speech signal analysis. We assumed that both EDA and speech signal reflected the emotional elicitation induced by the task and assessed by the SAM. Correlation and regression methods were used to understand how the linguistic structure of the texts interacts with both signals.

As regards the correlation analysis, we found a statistically significant relationship with some of the linguistic properties of the text. In particular, significance was found between linguistic features related to aspects of syntactic complexity, including the use of subordination and the verbal predicate structure, and the speech features that describe some prosodic aspects of speech often related to the human emotional state (e.g., F0, F3 variation over time). The results also show how speech features like the signal duration can be indicators of linguistic complexity, because of its strict relationship with the sentence length. Indeed, sentence length serves as an indicator of complexity because longer sentences typically involve more intricate dependencies and syntactic structures, such as multiple subordinate clauses. This increases the cognitive effort required to process and comprehend the sentence and its underlying structure. As a result, since sentence length and signal duration share a direct correlation, signal duration can also be regarded as a measure of complexity. Likewise, the EDA features describing the variability of both phasic and tonic components (std ph, std ton), as well as the number of phasic responses, were strongly correlated with most of the linguistic properties of the texts. These features often reflect arousing states such as fear and anxiety [29].

The strong significant relationship between linguistic characteristics and acoustic and EDA features was also confirmed by the prediction performance of the linguistic-driven SVR models. Indeed, the combination of linguistic features showed a significant and relevant prediction ability of the ANS-related features both when they described some

Table 5

Regression results for the prediction of Linguistic Features using in input speech features and EDA features. Highlighted in bold are the features that obtain a mean correlation value > 0.50.

	Number (and %) of significant subjects	Mean correlation	Correlation variance	Baseline
Raw text properties				
Sentence length	33 (100)	**0.8447**	0.0018	0.4563
Lexical variety				
Types fundamental lexicon	15 (45)	**0.5103**	0.0087	0.1336
Type/token ratio lemma	33 (100)	**0.6482**	0.0084	0.3439
Morpho-syntactic information				
Subordinating conjunctions	20 (61)	0.4416	0.0031	0.0563
Auxiliaries present tense	16 (48)	**0.5355**	0.0075	0.1362
Syntactic relations				
Adverbial clause modifier	28 (85)	**0.5355**	0.0072	0.2023
Marker	28 (85)	**0.5631**	0.0099	0.2836
Nominal modifier	20 (61)	0.4226	0.0034	0.1947
Nominal subject	15 (45)	**0.5112**	0.0103	0.2812
Object	15 (45)	0.4475	0.0049	0.0567
Global and local parsed tree structure				
Parsed tree depth	33 (100)	**0.7603**	0.0032	0.3852
Clause length	19 (58)	0.4995	0.0039	0.2985
avg dependency links length	33 (100)	**0.6771**	0.0085	0.3486
avg prepositional chains length	32 (97)	**0.5248**	0.0052	0.2120
Prepositional chains number	33 (100)	**0.6316**	0.0081	0.2990
Post-verbal object	28 (85)	0.4715	0.0064	0.1816
Prepositions distribution	17 (52)	0.4564	0.0083	0.1760
Subordination phenomena				
Principal propositions dist.	32 (97)	**0.6581**	0.0228	0.2647
Subordinate propositions dist.	33 (100)	**0.7234**	0.0077	0.2984
Post-verbal subordinates	31 (94)	**0.5542**	0.0087	0.2350
Subordinate chains length	33 (100)	**0.6594**	0.0063	0.3098

characteristics of the voice spectrum (i.e., fundamental frequency and formants) that could be altered by the respiratory activity and when they describe the physiological arousal manifested by the sweat gland activity.

Concerning the EDA features, in addition to those already shown by the correlation analysis, the SVR has shown a remarkable prediction performance of *edaSymp* values. Such a feature is indeed a reliable marker of the activity of the sympathetic system and a proven stress marker, supporting the hypothesis of a relationship between features typically recognized as proxies of linguistic complexity, especially at the syntactic level, and a stress reaction of the subject [26]. However, this result could suggest a double possible interpretation. On the one hand, the linguistic structure of the pronounced sentence may be a confounding factor that masks the actual contribution of voice prosody and EDA in estimating the emotional state when a subject is speaking. Indeed, the prosody and EDA dynamics variations could be due to the speech-related mechanical changes induced in the respiratory activity, which is known to influence both acoustic and EDA characteristics. On the other hand, the linguistic structure itself could directly influence the subject's emotional state, which would be correctly identified by the speech and EDA features. This last hypothesis has already been supported by some studies that have combined the features derived from speech processing with some linguistic features to feed classifiers for the recognition of the emotional state [13,30]. However, in these studies, the encoding of the text takes into account the lexical and contextual aspects of language but does not consider other important features examined in our studies, such as those encoding morpho-syntactic or syntactic information. Indeed, these features could substantially impact an individual's emotional state because they are related to a variety of psycholinguistic phenomena and could affect the cognitive load and processing difficulty of the language user. In this regard, the results we obtained are in line with previous studies (see, in particular, [11,12]) in which the same set of linguistic features on which we relied was shown to be highly correlated with conscious judgments of perceived sentence complexity given by native speakers.

Additional evidence that emerged from our study is that acoustic and physiological signals are reliable predictors of a large array of linguistic features, which contribute to encoding the morpho-syntactic and syntactic structure of the text. This finding emphasizes the strong interplay between the speaker's underlying emotional state and the linguistic structure of the text. Moreover, it holds valuable practical implications in the realm of computational modeling of language phenomena. In line with current research trends, cognitive signals have been shown to enhance NLP models through multi-task learning, offering informative explanations for distinguishing between human and machine language processing [17,31]. Multimodal fusion of these signals with textual data can thus lead to the development of more comprehensive and context-aware NLP models that can better understand and respond to users in various real-life settings and applications concerned with affective phenomena: from multimodal sentiment analysis systems to cognitively-inspired readability assessment tools and AI technologies in the educational scenario able to adapt to personalized learning paths. Additionally, recognizing which text features elicit specific emotional responses in readers could assist in regulating the automated processing and generation of language [32]. The incorporation of controlled emotions in automatically generated texts can be beneficial in a variety of scenarios: for instance, for improving the quality of conversational systems as this technology allows the system to respond to human users in a more empathetic manner, thereby fostering more meaningful conversations between the user and the AI-agent. Moreover, the ability to establish an emotional connection with the reader holds particular value for conversational therapy bots that need to generate appropriate emotional responses based on the user's mental state.

To expand our theoretical comprehension of human engagement dynamics and facilitate the advancement of emotionally-aware AI technologies, our future studies will thus involve a thorough exploration of the chosen linguistic features to assess their influence on predicting emotional states. Additionally, we find it intriguing to extend our investigation to include other physiological parameters, such as the electrocardiogram recorded during the reading task. By examining the correlation between these physiological signals and speech and linguistic parameters in affective reading, we aim to gain further insights into the interplay between physiological responses and emotional processing.

CRediT authorship contribution statement

Benedetta Iavarone: Methodology, Software, Validation, Formal analysis, Writing – original draft. **Maria Sole Morelli:** Data curation, Methodology, Software, Validation, Formal analysis. **Dominique Brunato:** Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Shadi Ghiasi:** Methodology, Software, Validation, Formal analysis. **Enzo Pasquale Scilingo:** Writing – review & editing. **Nicola Vanello:** Conceptualization, Writing – review & editing. **Felice Dell'Orletta:** Conceptualization, Project administration, Supervision. **Alberto Greco:** Conceptualization, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bspc.2023.105776.

References

[1] A.L. Callara, L. Sebastiani, N. Vanello, E.P. Scilingo, A. Greco, Parasympathetic-sympathetic causal interactions assessed by time-varying multivariate autoregressive modeling of electrodermal activity and heart-rate-variability, IEEE Trans. Biomed. Eng. (2021).

[2] J.R. Deller, J.H.L. Hansen, J.G. Proakis, Discrete-Time Processing of Speech Signals, 2000, p. 908.

[3] S.G. Koolagudi, K.S. Rao, Emotion recognition from speech: a review, Int. J. Speech Technol. 15 (2) (2012) 99–117.

[4] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T.F. Quatieri, A review of depression and suicide risk assessment using speech analysis, Speech Commun. 71 (2015) 10–49.

[5] C.L. Giddens, K.W. Barron, J. Byrd-Craven, K.F. Clark, A.S. Winter, Vocal indices of stress: a review, J. Voice 27 (3) (2013) 390–e21.

[6] R. Fernandez, R.W. Picard, Modeling drivers' speech under stress, Speech Commun. 40 (1–2) (2003) 145–159.

[7] A. Guidi, C. Gentili, E.P. Scilingo, N. Vanello, Analysis of speech features and personality traits, Biomed. Signal Process. Control 51 (2019) 1–7.

[8] Z. Zhang, Mechanics of human voice production and control., J. Acoust. Soc. Am. 140 (4) (2016) 2614, http://dx.doi.org/10.1121/1.4964509.

[9] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-ud: a tool for linguistic profiling of texts, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 7145–7151.

[10] S.E. Argamon, Computational register analysis and synthesis, arXiv preprint arXiv:1901.02543 (2019).

[11] D. Brunato, L. De Mattei, F. Dell'Orletta, B. Iavarone, G. Venturi, Is this sentence difficult? Do you agree? in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2690–2699.

[12] B. Iavarone, D. Brunato, F. Dell'Orletta, Sentence complexity in context, in: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, 2021, pp. 186–199.

[13] P. Singh, R. Srivastava, K. Rana, V. Kumar, A multimodal hierarchical approach to speech emotion recognition from audio and text, Knowl.-Based Syst. 229 (2021) 107316.

[14] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, 2018, arXiv preprint arXiv:1802.05365.

[15] A. Greco, G. Valenza, E.P. Scilingo, Advances in Electrodermal Activity Processing with Applications for Mental Health, Springer, 2016.

[16] S. Ghiasi, G. Valenza, M.S. Morelli, M. Bianchi, E.P. Scilingo, A. Greco, The role of haptic stimuli on affective reading: a pilot study, in: 2019 41st Annual EMBC, IEEE, 2019, pp. 4938–4941.

[17] N. Hollenstein, M. Barrett, L. Beinborn, Towards best practices for leveraging human language processing signals for natural language processing, in: Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources, European Language Resources Association, Marseille, France, 2020, pp. 15–27, URL https://aclanthology.org/2020.lincr-1.3.

[18] M.M. Bradley, P.J. Lang, Measuring emotion: the self-assessment manikin and the semantic differential, J. Behav. Ther. Exp. Psychiatry 25 (1) (1994) 49–59.

[19] G. Sarti, D. Brunato, F. Dell'Orletta, That looks hard: Characterizing linguistic complexity in humans and language models, in: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, Association for Computational Linguistics, Online, 2021, pp. 48–60, http://dx.doi.org/10.18653/v1/2021.cmcl-1.5, URL https://aclanthology.org/2021.cmcl-1.5.

[20] T. De Mauro, Grande Dizionario Italiano Dell'Uso (GRADIT), Torino, UTET, 2000.

[21] M.S. Morelli, S. Orlandi, C. Manfredi, BioVoice: A multipurpose tool for voice analysis, Biomed. Signal Process. Control 64 (2021) 102302.

[22] C. Manfredi, L. Bocchi, G. Cantarella, A multipurpose user-friendly tool for voice analysis: Application to pathological adult voices, Biomed. Signal Process. Control 4 (3) (2009) 212–220.

[23] A. Greco, G. Valenza, J. Lázaro, J.M. Garzón-Rey, J. Aguiló, C. De-la Camara, R. Bailón, E.P. Scilingo, Acute stress state classification based on electrodermal activity modeling, IEEE Trans. Affect. Comput. (2021).

[24] M.E. Dawson, A.M. Schell, D.L. Filion, The electrodermal system., 2017.

[25] A. Greco, G. Valenza, A. Lanata, E.P. Scilingo, L. Citi, Cvxeda: A convex optimization approach to electrodermal activity processing, IEEE Trans. Biomed. Eng. 63 (4) (2015) 797–804.

[26] H.F. Posada-Quintero, J.P. Florian, A.D. Orjuela-Cañón, T. Aljama-Corrales, S. Charleston-Villalobos, K.H. Chon, Power spectral density analysis of electrodermal activity for sympathetic function assessment, Ann. Biomed. Eng. 44 (10) (2016) 3124–3135.

[27] J.D. Storey, A direct approach to false discovery rates, J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (3) (2002) 479–498.

[28] K. Collins-Thompson, Computational assessment of text readability: A survey of current and future research, ITL - Int. J. Appl. Linguist. 165 (1) (2014) 97–135.

[29] A. Baldini, S. Frumento, D. Menicucci, A. Gemignani, E.P. Scilingo, A. Greco, Subjective fear in virtual reality: a linear mixed-effects analysis of skin conductance, IEEE Trans. Affect. Comput. (2022).

[30] B.T. Atmaja, M. Akagi, Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM, Speech Commun. 126 (2021) 9–21.

[31] O. Eberle, S. Brandl, J. Pilot, A. Sø gaard, Do transformer models show similar attention patterns to task-specific human gaze? in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4295–4309, http://dx.doi.org/10.18653/v1/2022.acl-long.296, URL https://aclanthology.org/2022.acl-long.296.

[32] I. Singh, A. Barkati, T. Goswamy, A. Modi, Adapting a language model for controlled affective text generation, in: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 2787–2801.