SCUOLA
NORMALE
SUPERIORE

Classe di Scienze

Corso di perfezionamento in
Data Science

XXXIV ciclo

“Understanding Emotive Response to Textual Stimuli: A Multimodal Approach”

INF/01 - Informatica

Candidato
dr.ssa Benedetta Iavarone

Relatori

dott. Felice Dell'Orletta
prof.ssa Anna Monreale

Anno accademico 2023/2024

“More human than human” is our motto.

Abstract

In an era where human activities are increasingly intertwined with technology, electronic systems and devices are expected to be intelligent and capable of responding in the most human-like way possible. However, for artificial intelligence systems to integrate seamlessly into human society, they must possess emotional intelligence – the ability to identify, understand, and react to human emotions. As interactions between humans and intelligent systems often rely on language, a deeper comprehension of the link between language and human emotional response becomes crucial for improving human-computer interactions. This thesis aims to understand which aspects of language are connected to emotional response and elicitation. Specifically, it applies an extensive array of linguistically motivated features designed to capture the stylistic elements of language and investigates their relation to human emotions, both in isolation and in combination with other features. Initially, the thesis explores how this broad set of linguistic features correlates with perceived language complexity and the role of language as an emotion elicitor. This same set of features is then used to understand the influence of third-party language on the emotions of a group of subjects. Lastly, this work examines the impact of deliberately emotive language on bodily responses and the relationship between such responses and the previously mentioned linguistic features.

Acknowledgments

This endeavour would not have been possible without the support and collaboration of many people. I could not have undertaken this journey without my supervisor Felice Dell'Orletta, who provided knowledge, expertise, and guidance from way before this project even started. I am also extremely grateful to my supervisor Anna Monreale and her consistent support and encouragement.

Special thanks to Luca Cecchetti, Giacomo Handjaras, Giada Lettieri, and Emiliano Ricciardi from the Molecular Mind Laboratory of IMT School for Advanced Studies Lucca, that have provided me with the data and support necessary for the second case study.

I am also grateful for the cooperation of Alberto Greco and Nicola Vanello from Research Center "E. Piaggio", and Maria Sole Morelli from Fondazione Toscana Gabriele Monasterio, that have made possible the third case study of this thesis.

Thanks should also go to the members of the ItaliaNLP Lab of the Istituto di Linguistica Computazionale "A. Zampolli" CNR Pisa, who were always present to answer my doubts and questions.

I would be remiss in not mentioning my family and my loved ones. Their belief in me has kept my motivation high during this process. Last but not least, I would like to thank my cat for all the purring, entertainment, and emotional support during the writing of this thesis.

List of Publications

1. Monreale, A., Iavarone, B., Rossetto, E., & Beretta, A. (2022). Detecting addiction, anxiety, and depression by users psychometric profiles. In *Companion Proceedings of the ACM International World Wide Web Conference* (WWW'22).

2. Iavarone, B., & Monreale, A. (2021). From depression to suicidal discourse on Reddit. In *IEEE International Conference on Big Data* (IEEE BigData 2021).

3. Iavarone, B., Morelli, M. S., Brunato, D., Ghiasi, S., Pasquale, E., Scilingo, N. V., Dell'Orletta, F., & Greco, A. (2021). Analyzing the interaction between the reader's voice and the linguistic structure of the text: A preliminary study. In *Proceeding of the 12th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications* (MAVEBA 2021).

4. Iavarone, B., Brunato, D., & Dell'Orletta, F. (2021, June). *Sentence complexity in context.* In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (pp. 186-199).

5. Iavarone, B., & Dell'Orletta, F. (2020). Predicting movie-elicited emotions from dialogue in screenplay text: A study on "Forrest Gump". In *Proceedings of the Seventh Italian Conference on Computational Linguistics* CLiC-it 2020, 230.

6. Gorrell, G., Bakir, M. E., Roberts, I., Greenwood, M. A., Iavarone, B., & Bontcheva, K. (2019). Partisanship, Propaganda and Post-Truth Politics: Quantifying Impact in Online Debate. *The Journal of Web Science, 7.*

7. Brunato, D., De Mattei, L., Dell'Orletta, F., Iavarone, B., & Venturi, G. (2018). Is this sentence difficult? do you agree?. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2690-2699).

8. Gorrell, G., Roberts, I., Greenwood, M. A., Bakir, M. E., Iavarone, B., & Bontcheva, K. (2018). Quantifying media influence and partisan attention on Twitter during the UK EU referendum. In *Social Informatics: 10th International Conference*, SocInfo 2018, St. Petersburg, Russia, September 25-28, 2018, Proceedings, Part I 10 (pp. 274-290). Springer International Publishing.

Contents

List of Figures

List of Tables

Introduction

In both sci-fi cinematographic and literary works, it's clear that a recurring theme concerns Human-Computer Interactions, with a particular emphasis on the creation of intelligent machines capable of faithfully replicating human behavior and seamlessly integrating into the human world. It is possible to recall different fictional examples of *human-resembling machines* – either in appearance or behavior – that are relevant for this discussion. These include the sentient computer HAL, from "*2001: A Space Odyssey*"[60], which interacts and converses with the crew of the Discovery One spaceship as if it were a crew member itself; the Replicants from "*Blade Runner*", androids designed to look like humans and capable of acquiring emotions; and, in the more recent motion picture "*Her*", an artificially intelligent virtual assistant embedded in an Operating System develops an intimate bond with the human protagonist, effectively becoming a surrogate for a real romantic relationship. In the movie "*Ex-Machina*", an Artificial Intelligence with the exterior form of a beautiful woman interacts with the programmer that is performing the Turing Test[1] on her. She eventually manipulates the programmer into believing that she is in love with him, culminating in a surprising twist in the plot.

An exceptional example of how human imagination conceives the interactions between humans and machines can be found in the literary work of Isaac Asimov, one of the most influential science fiction writers of the last century. In his "*Robot series*", Asimov crafts a world where highly intelligent robots, often exhibiting humanoid features, coexist with humans. The behavior of the robot population is guided by the Three Laws of Robotics[2], a set of principles designed to ensure

[1] The Turing Test is a method of inquiry in Artificial Intelligence to determine whether or not a computer is capable of thinking like a human being. According to the test, a computer possesses artificial intelligence if it can mimic human responses under specific conditions. The test is named after Alan Turing, an English mathematician, computer scientist, logician, cryptanalyst, philosopher, and theoretical biologist. Turing is considered the father of theoretical computer science and artificial intelligence.

[2] The Three Laws of Robotics in Isaac Asimov's work: (1) a robot may not injure a human

that the robots do not rebel against their creators.

As technology continues to evolve at a brisk pace, we are witnessing the creation of hyper-realistic robots that emulate human features and movements, conversational agents that provide assistance with a human-like touch, and smart devices controllable with simple voice commands. However, despite their impressive computational prowess and capabilities, these machines still fall short in replicating the complexity of human behavior. The crucial element they lack is the ability to comprehend and respond to emotions.

For machines to behave like humans, they need to possess what is called *emotional intelligence* [108][246]. Emotional intelligence is defined as the collection of skills that equip humans with the ability to have, express, and identify affective states, along with the capacity to use these states for constructive purposes, regulate and manage them, and navigate the emotional states of others [218]. It has been argued that emotional intelligence is perhaps the most important facet of human intelligence for successful social interactions [108]. Moreover, it plays a vital role in learning [37][246] and in various other functions such as perception and rational decision-making [225]. Therefore, it is reasonable to assert that emotional intelligence is one of the fundamental components of human-human interaction and that its presence is essential for an interaction to be considered intelligently responsive [237]. Lacking these emotional aspects, human-computer interaction is still far from the representations seen in works of fiction over the years.

It is also important to highlight that not every computing system will require emotional skills, and it is unlikely that any of them will ever need the same range of emotional skills as humans. However, there are many instances where human-computer interactions could be enhanced by machines capable of understanding users and adapting to them. To achieve this, machines should also be able to recognize the user's affective state. This concept is tied to numerous psychological studies that led researchers to conclude that humans interact with machines (e.g., computers, televisions) in the same way they do with other humans [237]. Therefore, systems capable of perceiving and responding adequately to the user's affective state are expected to be perceived as more natural [208], trustworthy [52] [158], and persuasive [237]. These findings, coupled with the increasing integration of technology into daily life and activities, have sparked a growing interest in the field of *Affective Computing* and in methods to infer emotions and affective states from various kinds of data.

Affective Computing, as an academic discipline, is relatively recent, as the term made its first appearance in the late 90s, established by Rosalind Picard

being or, through inaction, allow a human being to come to harm; (2) a robot must obey the orders given it by human beings except where such orders would conflict with the First Law; (3) a robot must protect its own existence as long as such protection does not conflict with the First and Second Laws.

in her book "*Affective Computing*" [224]. Since then, Affective Computing has blossomed as an interdisciplinary field, garnering attention and contributions from various research areas, such as social science, cognitive science, psychology, engineering, computer science, linguistics, physiology, and others. The primary objective of Affective Computing is to develop systems and devices that can automatically recognize and infer human emotions, leading to appropriate interpretations and responses. Equipping a machine with these capabilities can yield extensive benefits, such as enhancing the quality of the interaction between humans and computers or creating devices that react according to the user's emotional states and reactions.

The importance of affect-aware computing stems from the fact that, in human interactions, a significant amount of information is communicated implicitly. A person's state may be expressed through body language and gestures, facial expressions and eye movements, or the manner of speaking (e.g., intonation pattern, pitch) – cues that are intuitively recognized and interpreted by humans. Without the appropriate tools and training, machines cannot infer these elements. Overlooking the user's comprehensive emotional state, by neglecting explicit and implicit affective cues, eliminates a significant portion of the meaningful information available during the interaction process, thereby diminishing the value of human-computer interactions.

Affective computing is a broad field that encompassing a wide range of techniques and approaches aimed at recognize affect from data, across various modalities and at different levels of granularity [230]. Particularly, granularity is employed to differentiate among the diverse approaches to affect recognition. According to the granularity used to detect affection, there are two main branches of study: *Sentiment Analysis* and *Emotion Recognition*.

Sentiment Analysis[3] is the study of people's opinions, appraisals, attitudes, and emotions towards entities such as topics, services, products, events, and their attributes [169]. It performs coarse-grained affect recognition, typically considering it as a binary classification task (*positive sentiment* vs. *negative sentiment*). Sometimes, granularity is increased to three classes, including a *neutral sentiment*, or to five classes, incorporating two different degrees of positivity and negativity along with the neutral class. Moreover, sentiment analysis solely focuses on identifying the *valence* of an opinion or sentiment (see Section 1.1 for further explanations on valence). In contrast, Emotion Recognition[4] carries out fine-grained affect recognition, aiming to classify data according to a broad array of emotion labels. A more in-depth discussion on the labels and how they are chosen is provided in Section 1.1.

Modality refers to the type of data used for affect recognition, which can be sourced from various channels: audio (e.g., speech), video (e.g., facial expressions,

[3] Sometimes also referred to as *Opinion Mining*, using the two names interchangeably.

[4] Also referred to as Emotion Detection.

body gestures), textual (e.g., user-generated text, speech-to-text transcriptions), plus, more recently, physiological signals (e.g., heart-rate variability, galvanic skin response) and brain-generated data (e.g., collected with fMRI or EEG). Traditional affective studies focused on single-modality approaches, examining one type of data at a time. However, recent research suggests that integrating multiple modalities can enhance the performance of affect detectors, thus leading to a shift from unimodal to multimodal approaches (see Section 1.2.1 and Section 1.2.2 for details).

Objectives and contributions

The research presented in this thesis explores emotion recognition from a multimodal perspective, delving into various aspects related to emotion with natural language as the consistent thread running throughout all the chapters. Each chapter investigates the relationship between language and emotions from a unique perspective, incorporating other modalities alongside language.

In every case study presented in this thesis, language is employed as the primary means for eliciting emotional responses in humans. However, the objective of this work is not to establish a state-of-the-art approach in Affective Computing and Natural Language Processing. Instead, the main goal is to deepen the understanding of the link between language and emotional response, to grasp why specific emotional responses occur and identify the precise linguistic phenomena that trigger these reactions. To achieve these goals, this research utilizes a wide range of linguistically motivated features. These linguistic features capture various levels of linguistic information and can be used to construct a textual *profile*. This profile can show language variation within and across texts, highlighting differences in genre, style, or register. The scope of this thesis is to demonstrate how these linguistic aspects are associated with human emotions, specifically the emotions a subject experiences when exposed to different textual stimuli. This objective, in essence, underscores the power of language in shaping emotional landscapes and the crucial role it plays in the broader field of affective computing. These aspects will be tackled in three different case studies.

The first case study explores the complexity of language as a tool to elicit emotions that ensure a high level of engagement in interactions. While the concept of language may seem far from the study of emotions, it is closely related. Given that both human-human and human-computer interactions often rely on written or spoken communication, efficiency is paramount. This efficiency means that the recipient of a written of spoken message should be able to decode and understand the message with minimal cognitive processing effort. If the recipient perceives the message as *complex*, the cognitive load for processing increases, potentially affecting the emotions they experience. High linguistic complexity can generate feelings of frustration, boredom, sadness, or anger and can reduce the recipient's engagement. Conversely, a level of linguistic complexity that adequately tailored to the recipient can facilitate and potentially increase engagement, producing feelings of calmness, happiness, or interest.

In this case study, language complexity is assessed in two scenarios. One scenario examines the complexity of sentences in isolation, while the other considers the same sentences in the context of additional sentences, in order to determine whether context influences human perception of linguistic complexity.

The second case study investigates how language from a third party, without direct social interaction, influences the emotions of a group of subjects. It also examines whether language can predict the emotions subjects were experiences while hearing these sentences. Here, an ecological audiovisual stimulus, such as a movie that reproduces real-life content, is used to elicit emotions. The language is extracted from the movie characters' speech transcription, meaning that it is not directed explicitly towards the subjects. Emotion elicitation comes from the subjects' interpretation of the movie events and their resonance with personal experiences.

The third case study examines how purposefully emotively encoded language influences subjects' bodily responses. The focus here is on how emotive images encoded in a text impact the subjects' emotional responses. Subjects were exposed to language by being asked to read texts aloud. Bodily signals (i.e., electrodermal activity) and voice features were recorded during the reading, as these are closely tied to the emotions experienced by the subjects. The autonomic nervous system, which regulates involuntary physiological processes, controls emotional regulation, speech, and many bodily function. In this case study, linguistic aspects are used to predict the changes in subjects' bodily responses, and changes registered in electrodermal activity and voice are used to predict the linguistic characteristics of the texts being read.

Structure of the thesis

The remainder of this work is structured in this way: Chapter 1 discusses the State of the Art and how the study of emotions is tackled in the literature; Chapter 2 presents more in detail the case studies, by also giving an overview of the modalities applied in the studies and the features representations used to describe each of the modalities; Chapter 3 introduces the statistical and machine learning modeling approaches applied throughout this work; Chapter 4 studies the complexity of language out of context and in context, and how complexity relates to humans' emotions; Chapter 5 studies the relationship between language and the emotive response to an ecological audiovisual stimulus; Chapter 6 studies how emotively encoded language influences bodily responses and voice features; the last chapter lists the conclusions of this study.

Part I
Background

Chapter 1

State of the Art

This chapter provides an overview of how emotions are examined in existing literature. It begins by exploring different theories of emotions, addressing how and why individuals experience emotions, and how these emotions can be encoded, represented, and categorized. The chapter then delves into the methodology used in Affective Computing to detect emotions, drawing information from a variety of sources, or *modalities*. Furthermore, it investigates the methods used to combine these various sources to attain a comprehensive understanding of emotions.

1.1 Theories of Emotions

Before diving into the current state-of-the-art research in this field, it is crucial to establish clear definitions for the terms used in the study of emotions. In the sphere of Affective Computing, there remains considerable ambiguity surrounding the definition of *affection*. Often, terms like emotion, mood, or affect are used interchangeably in this field, without making clear conceptual distinctions [82]. It is worth noting that affection is a psychological construct that goes beyond just emotions and moods [28][244], which are described as distinct phenomena with different facets and attributes in literature. Emotions are generally regarded as directed towards a specific object, while moods are characterized as broad states without a defined cause or target [102]. Colloquially, the term *affect* is often used to refer to all of the above concepts. However, for the purpose of this work, the term *emotion* will serve as an umbrella term, covering a wide range of affective phenomena pertinent to Affective Computing research, including feelings, moods, attitudes, or temperament.

It is difficult to precisely define what an emotion is, and there is a lack of consistency in definitions across literature [146]. Therefore, it is impossible to encapsulate all definitions into a single concept [306]. Emotional processes and states are highly complex and can be analyzed from numerous perspectives. However, for machines to effectively process emotions, a clear categorization and set of easily identifiable labels are needed. In response to this need, Affective Computing relies on traditional emotion theories, established over decades, to provide a proper categorization for emotions. Various models for representing emotions have been developed based on these theories and are applied in numerous affective detection applications.

Charled Darwin was the first scholar to adopt a scientific approach to studying emotions, proposing an *evolutionary theory of emotions.* In his book *The Expression of the Emotions in Man and Animals*[68], Darwin reports anecdotal and observational evidence regarding the expression of emotions in animals, infants, preliterate human groups, and Western Europeans. During his studies about evolution, Darwin noticed that some bodily and facial expressions of humans were similar to those found in other animals. He concluded that emotions were the result of some evolutionary process and that they evolved with definite functions. An essential aspect of Darwin's theories on emotions was his belief that emotional expressions initially corresponded to basic communicative actions (e.g., a face expressing disgust to reject an offensive object). These actions were vital for survival and communication among animals. Only later did these communicative actions evolve into pure emotional expressions (e.g., the disgusted face is used even when the offensive object of the previous example is no longer present).

1.1.1 The categorical model

Building up on Darwin's findings, the concept of *basic emotions* was introduced. Research has demonstrated that a limited number of facial expressions associated with emotions are universally recognized, leading researchers to propose the existence of innate emotions that have cross-culturally universal counterparts.

The theory that humans share a set of basic emotions was further explored and expanded in the early 1970s when Ekman discovered through his research the existence of universal facial expressions associated with six basic – *prototypical* – emotions [85, 89]: happiness, sadness, anger, fear, surprise, and disgust. According to Ekman's theory, certain facial muscle movements are associated with specific emotions, transcending cultural boundaries to assume the same meaning across diverse cultures. Additionally, each basic emotion possesses a unique set of characteristics that distinguish it from the others [84]. Ekman acknowledges the possibility of pan-cultural similarities for more than these six emotions. Still, he emphasizes that these six emotional terms are not arbitrary, as they are the emotional concepts researchers consistently found when studying facial expressions across cultures [87].

The huge resonance obtained by Ekman's theory has led to a majority of existing affect recognition studies being based on a *categorical model of represen-*

tation. This model employs a discrete set of categories that affective systems can utilize as labels for emotion recognition. The use of discrete categories provides several advantages in emotion recognition, largely because the discrete scheme is intuitive and embedded in everyday language [88, 91]. The terms used for the six basic emotions echo people's experiences with emotions, and numerous studies have utilized this scheme [46].

1.1.2 The dimensional model

Some researchers have challenged Ekman's perspective and the categorical model, arguing that it is insufficient for for describing the nuanced facets of the human emotions complex. From this disagreement, more theories have subsequently emerged.

Firstly, it is important to recognize that human emotions are often far more multifaceted than what is depicted in the categorical model. Ekman himself has stated that there are more than just six basic emotions. Indeed, there exist "families of emotion" [84]. Members of an emotional family share certain characteristics (such as commonalities in expression, physiological activity, or triggering events) that distinguish one family from another. Among the primary emotions that share a set of characteristics that make them *basic*, there are emotional states that do not encompass all the features typically associated with basic emotions [84]. Emotions like curiosity, frustration, engagement, boredom, fatigue, and anxiety, fall under the category of non-basic emotions.

Thus, it becomes clear that a categorical model, with its few discrete categories, carries considerable limitations and cannot fully capture the multifaceted world of emotions. An alternative to the categorical model for describing human affect is the *dimensional model* [113][245][293], where an affective state is represented as coordinates in a multi-dimensional space. Although this model could incorporate numerous latent dimensions, such as evaluation, activation, control, and power, the most common approach uses no more than two or three dimensions due to theoretical and practical considerations.

An early example of the dimensional model, derived from cognitive theory, is Russel's circumplex model [242][243]. This model employs the dimensions of arousal and valence to plot up to 150 affective labels. Valence, a fundamental aspect of emotional life, originates from the human mind's ability to distinguish between a pleasant, beneficial feelings and unpleasant, harmful ones [21]. Thus, moving along the valence axis, it is possible to identify a range of emotional states, from positive to negative. Arousal refers to the degree of activation of the emotional state [242][243], i.e., the amount of energy required to express a particular emotion. This can range from a state of high activation or tension to a condition of drowsiness and deactivation. A representation of Russel's model can be seen in Figure 1.1. Here, the vertical axis signifies the dimension of arousal

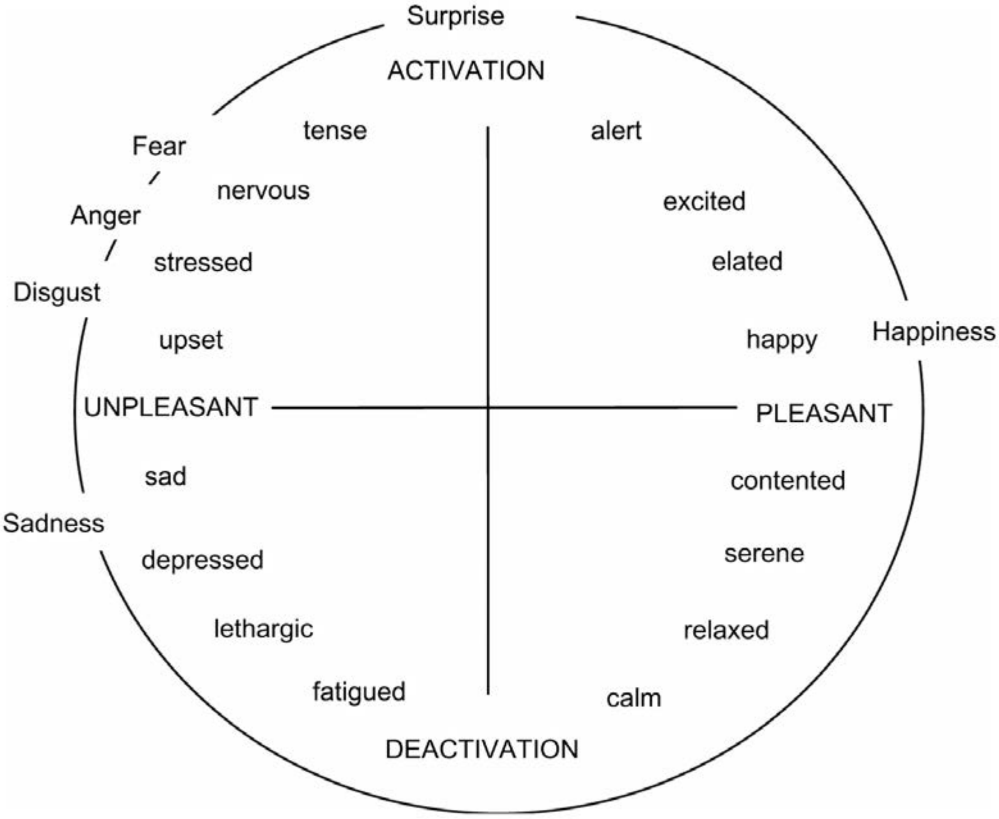


Figure 1.1: A representation of Russel's circumplex model of affect. *[Image source Pennsylvania State University website]*

(from activation to deactivation), while the horizontal axis signifies the dimension of valence (from pleasant to unpleasant). The six basic emotions are positioned along the circle, while the non-basic emotions are represented within the circle.

A similar model to the one proposed by Russel is the 2-dimensional model proposed by Whissell, in which emotions lay in a continuous space whose dimensions are evaluation and activation. The evaluation dimension measures how individuals feel on a continuous scale, from positive to negative. The activation dimension measures the propensity for action under a particular emotional state, on a continuous scale from passive to active. Whissell uses these two dimensions to assign a score to over 4000 emotional words that constitute her *Dictionary of Affect in Language.*

Another widely used bi-dimensional model is Plutchik's wheel of emotions [229], which recalls Darwin's theories on the evolutionary role of emotions [228] – according to which emotions have evolved for a particular function (*functionalist approach*). According to Plutchik, the feeling states referred to as emotions are part of a process encompassing both cognition and behavior, containing several

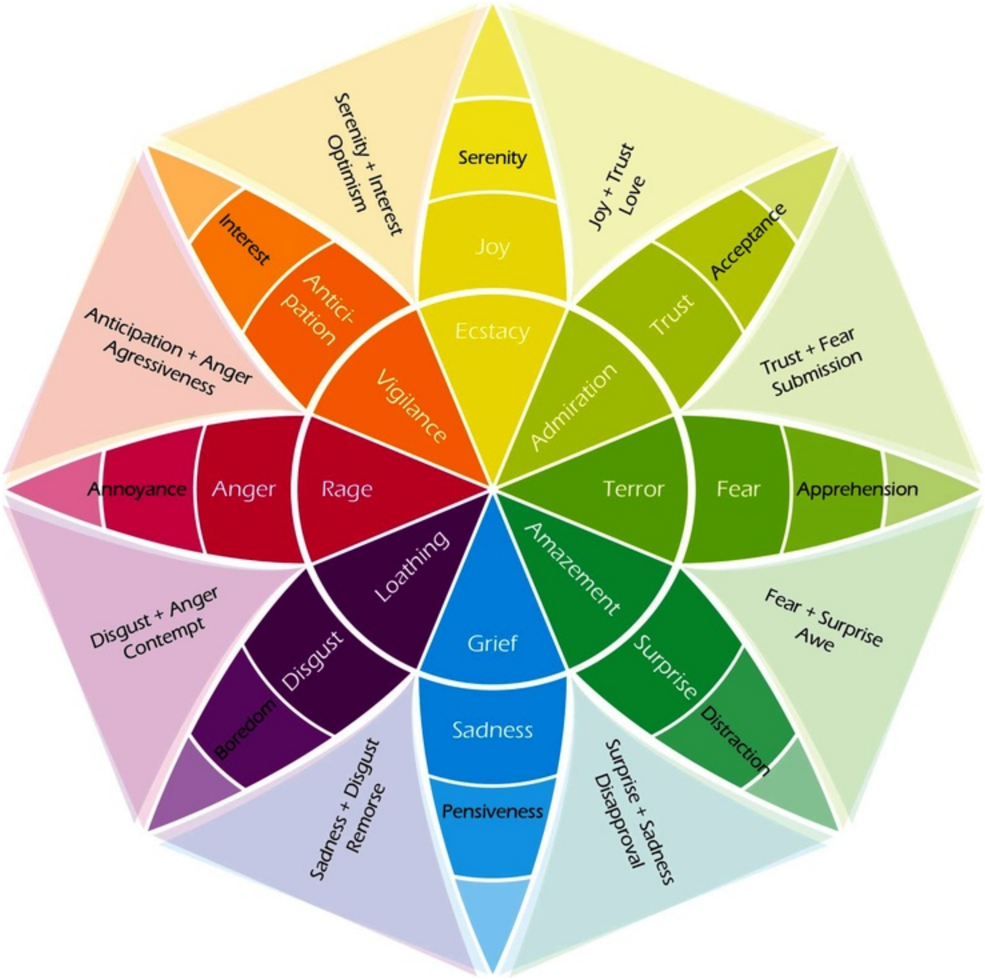


Figure 1.2: Plutchik's wheel of emotions. *[Image source positivepsychology.com]*

feedback loops. Accordingly, he designed his wheel of emotions, consisting of eight basic and eight advanced emotions, where each advanced emotions is composed of two basic ones. In this model, a vertical dimension represents intensity, and a radial dimension represents degrees of similarity among emotions.

While bi-dimensional approaches are the most commonly used due to their simplicity, models employing more than two dimensions also exist. A prevalent approach for emotion representation is the *arousal*, *valence*, *dominance* set, known by various names in literature [194]. *Dominance* is defined as a feeling of control and influence over one's environment and others, versus the feeling of being controlled or influenced by situations and others. Some theories suggest that there should also be a fourth dimension: *unpredictability* [100]. This dimension, related to surprise, uncertainty, and unexpectedness, describes an urgent reaction to a novel stimulus or an unfamiliar situation and appears necessary for adequately distinguishing emotions.

Dimensional models overcome the limitations of categorical ones, being better equipped to handle non-discrete emotions and describe emotions beyond word labels. However, dimensional approaches also have a few downsides. Although the dimensional space allows comparison of affective words according to their distance, it does not offer a method to study compound emotions or account for the interaction between different emotions. Indeed, experiencing two or more emotions simultaneously is relatively common. Additionally, most dimensional approaches operate at a word level and are incapable of capturing emotions encoded in multiple-word concepts.

1.1.3 The Hourglass of Emotions

All limitations included in categorical and dimensional models are overcome by the Hourglass of Emotions [47], inspired by Plutchik's studies. The Hourglass is also a brain-inspired and psychologically-motivated model based on the idea that the mind is made of various independent resources and that emotional states result from activating some of these resources while deactivating others. Thus, each combination of resources changes our thinking by modifying and tuning our brain's activities. This proposition is supported by different fMRI experiments that identify distinct brain activity patterns associated with different emotions (e.g., [309]).

The Hourglass of Emotions reinterprets Plutchik's model representing affective states through labels organized around four independent but concomitant dimensions. These dimensions measure: how much the user is amused by the interaction modalities (*pleasantness*), how much the user is interested in interaction contents (*attention*), how comfortable the user is with interaction dynamics (*sensitivity*), how confident the user is in the benefits of the interaction (*aptitude*). For each of these affective dimensions, the model employs multiple (polarised) activation levels, covering cases where up to four emotions can be expressed simultaneously and allowing for algebraic reasoning on these emotions. The model also enables reasoning on both single words and multiple-word expressions. It provides a formula to calculate polarity based on emotions, representing a preliminary attempt to bridge the gap between sentiment analysis and emotion recognition.

1.2 Approaches for Emotion Detection

Humans interact with the world around them through an array of senses, which include producing and interpreting sounds, observing visual cues, smelling various odors, and tasting different flavors. The term *modality*, in the context of Affective Computing and Emotion Recognition literature, refers to these different

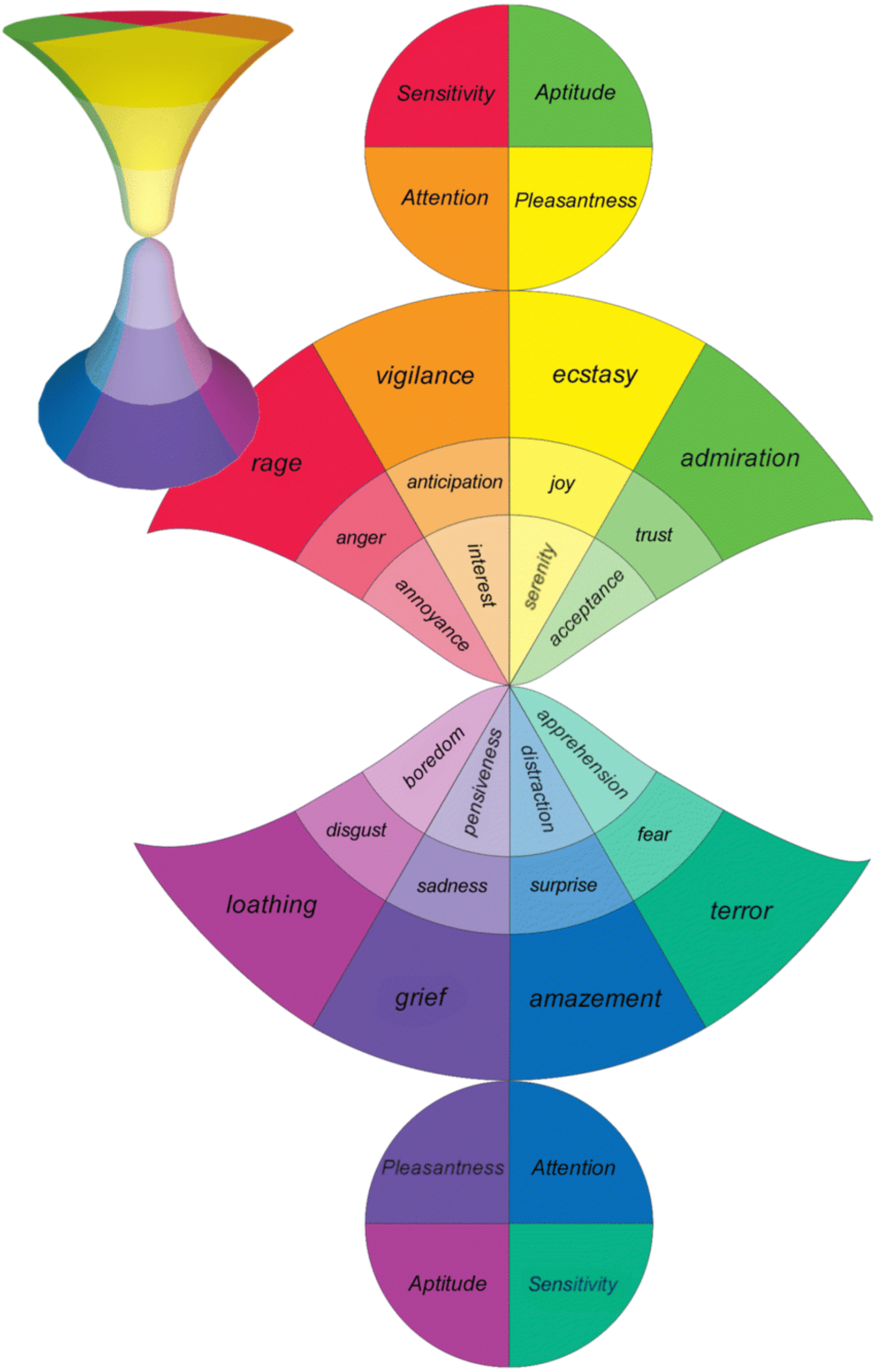


Figure 1.3: The diagram representing the hourglass of emotions. *[Image source Cambria et al. [47]]*.

channels of perception and interaction [19]. It signifies the distinct sources of data that can be anlyzed to detect and interpret emotions. Single modalities serve as fundamental components for emotion recognition. Over time, researches have investigated various modalities, generating interesting findings for affect detection. This section proposes an overview of the most commonly studied modalities and the techniques associated with them. Subsequently, the section delves into multimodal approaches, which encompass research issues that incorporate multiple modalities.

1.2.1 Unimodal approaches

Audio modality

Over the years, various trends have emerged in the study of the audio modality for emotion recognition. This field has recognized several audio features as instrumental for this purpose. More recent trends include the examination of affective reactions to everyday sounds [284], emotional responses triggered by music listening [167], and efforts to decode emotions embedded in music audio signals [170, 173, 269]. There has also been an interest in comprehending affect in naturalistic videos, such as spontaneous dialogues or audio recordings from interviews or call centers [25, 161].

Emotion recognition from speech represents one of the earliest and most enduring trends in the audio modality, primarily due to the natural role of speech in human-computer interactions. Indeed, as voice-based interactions with machines become more prevalent, speech systems are required to effectively process the underlying emotions of a conversation to reach a high level of performance, nearing human-level interaction [215]. Incorporating a component that processes emotions into such system would render them more natural and effective.

However, the implementation of speech emotion recognition is not without its challenges.

- Determining the most suitable acoustic features that can characterize and differentiate between emotions remains a subject of exploration [92]. Early research concentrated on the phonetic and acoustic properties of language. Psychological studies relating to emotion have revealed that vocal parameters, such as pitch, intensity, speaking rate, and voice quality, play a pivotal role in emotion detection and sentiment analysis [207]. Consequently, various prosodic and acoustic features have been proposed to help machines detect emotions from acoustic signals [206, 207, 302, 303].

- There are no standard speech corpora to compare the different approaches used for speech emotion recognition [149]. Current emotional speech system databases can be broadly categorized into: (*i.*) actor (simulated) emotional speech database, (*ii.*) elicited (induced) emotional speech database, (*iii.*) natural emotional speech database. In the case of simulated emotional speech corpora, the data is obtained from professional theater or radio artists, which are asked to express sentences with different emotions. Elicited speech corpora are collected by creating different contextual situations in which the conversation is constructed to elicit different emotions from the subject without their knowledge. For example, natural emotions may be collected from call center conversations and dialogues in public places. Collecting these emotions is more complex than what happens in the previous two settings, as they are usually milder, and their annotation requires the knowledge of experts. The size of the databases also heavily influences the generalizability and reliability of studies, with many datasets considered to be too small [76].

- Speech emotion recognition systems shoudl be independent of speaker and language, but they are often influenced by speaker-dependent and language-dependent information [287].

- Lastly, these systems should have the robustness to process noisy and real-life speech effectively.

Visual modality

Most studies in vision-based affect recognition primarily focus on analyzing facial expressions, where visual features are detected from images and videos. Notably, Ekman and his colleagues were pioneers in this field, conducting extensive research on human facial expressions in the 1970s [89]. They proposed the possibility of identifying six basic emotions based on cues from facial expressions (refer to Section 1.1.1). Building on this, Ekman and Firesen developed the Facial Action Coding System (FACS) [86]. This sytem encodes facial expressions based on muscle movements, where each muscular movement corresponds to an Action Unit (AU). A combination of AUs thus represents a specific facial expression.

Inspired by Ekman's work, numerous researches have used image and video processing techniques to analyze facial expressions. The majority of studies in automatic facial affect recognition are focused on detecting basic emotions [148]. There have been few attempts in detecting nonbasic affective states from facial expressions, such as fatigue [115], frustration [141], or other complex mental states [94] (e.g., agreeing, disagreeing, interested, concentrated, thinking, unsure). Moreover, earlier works on facial affect recognition largely relied on deliberate and often exaggerated facial expressions. More recently, however, there has been a growing interest in interpreting spontaneous facial expressions [24, 188].

Standard approaches to facial emotion recognition generally follow three significant steps: (*i*) the detection of the face and facial components, (*ii*) the feature extraction, and (*iii*) the expression classification [148]. The first two steps focus on the face region detection (*face acquisition*) and the extraction of either geometric features [95, 217, 292], appearance features [23, 168, 177], or a combination of both [75, 272, 273, 295, 312].

Geometric features describe the shape and location of the face and its components, such as mouth, eyes, nose, eyebrows, and chin. The facial components (or *facial feature points*) are extracted to form a feature vector that represents the geometry of the face [134]. Appearance-based features, on the other hand, detail the texture of and appearance of the skin. They focus on changes in the appearance and expression of the face, such as wrinkles and furrows. With these features, image filters are applied to either the entire face or specific regions to extract a feature vector [134]. The final steps involves applying a pre-trained classifier for facial emotion, leveraging the features extracted in the previous steps to produce recognition results.

In recent times, Deep Learning models have replaced traditional approaches for face emotion recognition. The advantage of Deep Learning methods is that they enable "end-to-end" learning directly from the input images. This eliminates the need for pre-processing stages and face-physics-based models [290]. Among the various Deep Learning models available, Convolutional Neural Networks (CNNs) are the most popular for face detection and emotion recognition. In CNN-based approaches, the input image is convoluted through a collection of filters in the convolution layers to produce a feature map. Each feature map is then combined into fully connected networks, and the facial expression is recognized as belonging to a specific class.

Face recognition poses numerous challenges, primarily due to the considerable variability in facial appearances [134]. The appearance of a face is subject to multiple factors, that affect inter-subjects and intra-subject variations. Among the factors that influence face recognition are:

- pose: the face's position within the frame of the video or image (camera point of view);
- occlusion: whether the subject's mouth is open or not, and to what extent;
- aging;
- facial expressions;
- accessories: glasses, hat, jewelry;
- technical factors: the position of the light source or the brightness of the image.

Although most research has concentrated on facial expressions, a few studies extract features from body movements, including movements of the limbs [236, 247], head movements [260], and posture [162]. Indeed, there is evidence in the communication of nonverbal behavior and psychology research that body movements convey affective expressions [147][291].

Textual modality

Research on textual modality is divided into two main tasks: sentiment analysis and emotion recognition.

Sentiment analysis targets the detection of valence (or *polarity*) in text, aiming to discern if a text is positive, negative, or neutral. Nowadays, a myriad of state-of-the-art sentiment analysis systems are available. They display impressive performance on various types of text (e.g., product reviews, social media texts, instant messages) and can accurately identify text polarity.

While sentiment analysis is a well-established task, emotion recognition and categorization continue to evolve. More attention is required on emotion detection from text [46, 50, 109, 251]. Despite the abundance of textual data, understanding fine-grained emotions remains a challenge [3]. This issue is partially attributed to the absence of other contextual and emotional cues in textual communication (e.g., facial expressions, voice modulation) [54, 109]. Emotions in text may not always be explicit, and it is necessary to interpret the meaning of the text. If emotions cannot be directly extracted, they must be inferred by interpreting the concepts expressed in the text and their interactions [153], also taking into account that some texts may convey multiple emotional expressions at the same time.

Several computational approaches have been proposed in the literature for identifying emotions in text. Traditional approaches for text emotion recognition are based on hand-crafted features engineering that attempt to map documents, sentences, and words to a set of emotions. These methods include keyword-based approaches, corpus-based approaches, and rule-based approaches. These standard approaches aim to leverage hand-crafted features to identify keywords in a sentence with explicit emotional or affective value (e.g., [264]). Following text pre-processing stages (e.g., stopwords removal, tokenization, lemmatization), words in a text are extracted via linguistic rules and matched against lexicons carrying emotional labels.

To serve this purpose, various resources have been created. For instance, *WordNet-Affect* [265], an extension of the WordNet database [201] with the representation of affective concepts and labels. Another resource, *Senti-WordNet* [96], describes the objectivity, positivity, and negativity of the terms within a synset[1].

The *NRC Word-Emotion Association Lexicon* [204] is another notorious affective lexicon, developed by the National Research Council of Canada. It comprises lists associations of words with eight emotions (anger, fear, anticipation, trust,

[1] In WordNet a *synset* is a group of terms with similar meaning (synonyms).

surprise, sadness, joy, and disgust) and two sentiments (negative, positive). This lexicon was created through manual annotation on a crowdsourcing platform and is available in 40 different languages.

EmoSenticNet [231] is a lexical resource that assigns six WordNet Affect emotions labels to *SenticNet* [48] concepts. SenticNet is a knowledge base that provides a set of semantics, sentics, and polarity associated with natural language concepts.

DepecheMood [259] is a high coverage and high precision emotion lexicon that provides emotion scores for 37 thousand terms, constructed by extracting crowdsourced affective annotation from a social news network.

Among the conventional resources for emotion annotation and recognition, there is also the *Affective Norm for English Words* (ANEWs) [38], a project that develops a set of normative emotional ratings for English words that elicit emotions. Other approaches relate to cues from emoticons and hashtags (if present in the text), e.g., [97][125], or the extraction of statistical features, such as frequent n-grams, negations, emoticons, hashtags, to generate representations for a classification model, e.g., [164].

Thanks to recent advancements in Natural Language Processing, a new era for textual emotion detection is emerging, with a focus on Machine Learning and Deep Learning methods [3]. Machine Learning approaches solve the problem of Emotion Detection by applying both supervised [9, 50, 59, 305] and unsupervised [50, 59, 179, 185, 202] learning techniques to classify texts into the different emotion categories. Frequently used Machine Learning models include Support Vector Machine [8, 20, 118, 126, 137, 171], Naive Bayes [20, 126, 266, 298], or Decision Tree [125, 171].

Although Machine Learning approaches are widely used and can deliver strong performances, the current trend is to exploit the strength of Deep Learning approaches [3]. It has been shown that a simple deep learning framework outperforms most state-of-the-art approaches in several Natural Language Processing tasks [63]. Thus, deep learning approaches are being evaluated for both sentiment analysis and emotion recognition from text. Some of the most applied Deep Learning approaches in emotion recognition include the Long Short-Term Memory (LSTM) model [128]. The LSTM has been applied in its base form [124, 142] and its variations, such as Nested LSTM [124, 142], hierarchical LSTM [132], or Bi-directional LSTM [27, 55, 178]. Another commonly used Deep Learning model is the Convolutional Neural Network [160], borrowed from studies on Computer Vision and successfully applied in the context of Natural Language and emotion recognition [4, 219]. More recently, there was a raise in the use of transformer-based architectures, such as BERT (see Section 3.3), which has been successfully applied in many emotion recognition tasks [2, 132, 153, 180].

Physiological Signals modality

Humans express emotions not only through facial expressions and their speech, but also through physiological changes in their bodies. Even when emotions are not explicitly communicated, changes in emotional state are often accompanied by alterations in physiological patterns [39]. In essence, emotional shifts are reflected in the workings of the nervous system.

The human nervous system is composed by two parts, the central nervous system (CNS) and the peripheral nervous system (PNS). The PNS consists of the autonomic nervous system (ANS) and the somatic nervous system (SNS). The ANS controls sensory and motor neurons and is the main responsible for the regulation of various bodily functions and internal organs, such as the heart, the lungs, the viscera, and the glands. The ANS is also involved in the mechanism of emotional regulation [45].When humans encounter specific situations, the CNS and ANS respond to external stimuli by initiating specific physiological changes. It is within these systems that emotional alterations occur [51]. Given that changes in the CNS and ANS are largely involuntary and cannot be consciously controlled, physiological signals offer a valuable source of authentic data on emotional activation.

In recent years, the field of emotion recognition has embraced the use of physiological signals. These signals circumvent some of the difficulties associated with other modalities. For instance, techniques used for analyzing these signals can collect vast amounts of data that can be later used for emotion recognition. Conversely, obtaining large datasets using other modalities (such as audio, visual, and textual) and techniques can be more challenging.

Common parameters monitored for applications related to affective detection include [34, 78, 254]:

- **Cardiovascular System** – Heart Rate (HR) and Heart Rate Variability (HRV), Respiratory Sinus Arrhythmia (RSA), Cardiac Output, Inter Beat Interval (IBI), Blood Pressure (BP), Electrocardiogram (ECG);
- **Electrodermal Activity** – Skin Temperature Measurements, Skin Conductance (SC), Galvanic Skin Response (GSR);
- **Respiratory System** – Breaths per minute, Respiration volume;
- **Muscular System** – Electromyography (EMG);
- **Brain Activity** – Electroencephalography (EEG), imaging (fMRI, PET).

Physiological signals are typically recorded with equipment and techniques that may be more invasive than those used for facial recognition or vocal expression. This is primarily due to the necessity of physical contact with the subject. However, recent advancements in wearable sensor technology, coupled with a growing interest in developing less invasive techniques (e.g., [12, 107]), ensure a more accessible approach to this type of research.

While working with physiological data has its benefits, it also presents significant challenges. For instance, to produce accurate predictions for emotion recognition, the collection of meaningful data is essential. This process is relatively straightforward for text, audio, or video modalities. However, for physiological signals, specialist knowledge is required for collection and interpretation [226]. Furthermore, physiological signals are the body's natural response to events. Therefore, to obtain high-quality data, these signals must be naturally elicited from subjects, adding another layer of complexity to the process.

1.2.2 Multimodal fusion

Even though unimodal approaches might perform well in many emotion recognition tasks, situations often arise where accurate emotion detection requires more than a single modality. Consider, for instance, a speech recording suddenly interrupted by a loud noise, or when the speaker falls silent for a considerable duration. Retrieving the subject's emotion from the audio signal in such a scenario would prove extremely challenging, if not impossible. However, if systems were simultaneously recording the person's facial expressions or monitoring their physiological signals, determining that person's emotional state could still be possible. This is because emotional states typically trigger multiple physiological and behavioral response systems, and integrating these different modalities can enhance emotion recognition systems.

D'Mello and Kory [74] highlight several reasons why multimodal approaches can ouperform unimodal ones. Firstly, multimodal approaches portray human affective expressions with higher fidelity. Affective expressions are complex coordination of signals that include involuntary, semi-voluntary, and voluntary responses [84, 85]. Therefore, the analysis of multiple signals and their interdependent relations likely provides a more accurate representation of the intricacies of human affective expressions. Secondly, unimodal approaches can suffer from problems associated with missing data and noisy channels.Multimodal approaches can help integrate data when they are missing in one modality and reduce inaccuracies caused by noise. However, D'Mello and Kory also caution that the interstudy variance in multimodal affect detection complicates the proper evaluation of the actual advantages of multimodality over unimodality.

Using features from multiple modalities to classify and recognize emotions presents certain challenges. Different modalities provide heterogeneous data, meaning they vary greatly in nature and form. Therefore, creating a multimodal representation that is both understandable for the model and leverages the strengths and weaknesses of each modality is essential. The heart of the challenge lies in finding an effective fusion method to integrate these various modalities into a single representation.

The fusion method, i.e., the method applied to integrate single modalities in a multimodal representation, is the basis of any multimodal approach. It is a challenging process due to the heterogeneity of data, for which researchers adopt several common approaches:

- **Feature-level or early fusion.** Here, features from different modalities are combined into a single feature vector, which is then analyzed. This method is quite challenging due to the fact that features of different modalities often come in different formats and shapes, thus it is necessary to understand how to represent features in a univocal format.
- **Decision-level or late fusion.**This approach examines and classifies the features each modality separately. The individual results are then fused into a decision vector, which is used for the final decision.
- **Hybrid fusion.** As the name suggests, this method mixes the two previous methods to take advantage of both while minimizing their disadvantages. For instance, it can combine independent decisions of individual unimodal classifiers with the decision of a feature-level fused multimodal classifier [57].
- **Model-level fusion.** This approaches leverages the correlations and interdependencies between different modalities during the fusion process. As outlined by Atrey et al. [14] and Poria et al. [230], this fusion method can be categorized into three groups, according to the basic nature of the methods and the problem space: rule-based, classification-based and estimation-based fusion methods.

Another challenge in multimodal approaches regards which modalities to fuse. Most approaches in the literature are bimodal [74], with a preference towards audio-visual approaches (e.g., face and voice), the fusion of text and speech, or a combination of different physiological signals. Approaches that combine more than two modalities are rarer, as the complexity of fusing increases with the number of modalities to be combined.

Chapter 2

Case studies, Modalities, and Representations

This work encompasses a diverse range of case studies, each highlighting a unique way in which linguistic aspects interact with emotive response and emotion elicitation. The case studies draw from three distinct modalities (refer to Section 1.2 for a definition of modality): textual, audio, and physiological signals – in particular, the skin conductance signal from the electrodermal activity. The modality are either employed by themselves (unimodal approach, see 1.2.1) or combined together (multimodal approach, see 1.2.2). For these modalities to be used in any kind of statistical analysis and to be applied as input to machine learning models, they need to be translated into a numerical representation. This conversion involves extracting various sets of features (or different representations) from each modality.

This chapter provides an overview of the different case studies in this thesis, focusing on their main points, and offers a detailed examination of the various features extracted from each of the previously mentioned modalities. The Chapter is structured in the following way. Section 2.1 outlines the three case studies tackled in this work; Section 2.2 describes the set of linguistic features extracted to analyze the textual modality, Section 2.3 describes the features used to represent the audio modality (speech), and Section 2.4 describes the features used to represent the physiological signals modality (skin conductance response from electrodermal activity).

2.1 Case studies

This work comprises three main case studies. Each of them places natural language as the main protagonist, exploring its far-reaching implications on cognitive and emotional states, and physiological responses. In every case study, language is represented by a wide range of linguistically driven features. As necessary, these are correlated with other modalities and representations.

2.1.1 Case Study 1: The role of linguistic features on language complexity and user engagement

The first case study dives deep into the subject of natural language complexity and its impact on efficient communication. The central premise of this case study is straightforward: the complexity of a message can either increase or decrease user engagement, both in human-to-human and human-to-computer interactions. A conversation filled with overly complicated phrases or a user interface riddled with convoluted instructions are likely to trigger frustration and disinterest. Conversely, a linear and understandable language, that requires minimal cognitive effort to be processed, is more likely to elicit a positive emotional response. By understanding the linguistic aspects that cause a message to be perceived as complex, it is possible to intervene on these elements and enhance interactions, ultimately enriching the experience and making it more engaging.

However, defining complexity in natural language is a challenge due to the lack of a unified understanding in the literature. The definition of complexity is influenced by the different perspectives used to study this concept, including psycholinguistic, historical, neuroscientific, and computational angles. To navigate these intricacies, two main approaches can be applied to the study of language complexity: absolute and relative. The absolute approach, driven by theory, measures complexity based on the number of parts in a linguistic system. The relative approach, on the other hand, assesses linguistic complexity from the viewpoint of language users and their experience of processing difficulty. This case study is based on the relative approach, framing complexity as perceived complexity difficulty and focusing on the subjectivity of perception.

Language complexity is not confined to individual words and sentences. The broader context within which these words and sentences appear also plays a critical role. Models of language comprehension emphasize the significance of contextual cues, an aspect that this study takes into account. To analyze the phenomenon of linguistic complexity, this study will first focus on the perceived complexity of sentences taken in isolation, and then consider the perceived complexity of the same sentences when presented within other contextual sentences.

The primary objectives of this case study are to understand the role of linguistic phenomena in the perception of language complexity, to identify the key phenomena involved in predicting complexity, to analyze the influence of context on complexity perception, and to demonstrate the predictive superiority of models using explicit linguistic features over those using implicit features. By studying sentences in isolation and within context, this case study aspires to offer a comprehensive understanding of the nuanced subject of language complexity.

2.1.2 Case Study 2: The role of linguistic features in emotion elicitation

This second case study delves into the topic of emotion elicitation from third-party language and seeks to understand which linguistic features are associated with different emotions. In the field of psychology, eliciting emotional states within controlled environments for investigative purposes has been a long-standing tradition. Numerous techniques have been employed for this purpose, encompassing a wide variety of stimuli including, but not limited to, images, auditory cues, facial and bodily gestures, imagined scenarios, and smells. Recent decades have witnessed a paradigm shift toward understanding emotional responses through the affective processing of images.

Affective images have gained widespread popularity due to their inherent evocative potential and ease of implementation. These images are adaptable to diverse experimental designs and are easy to catalog and edit. Additionally, their static nature is particularly advantageous for studies where dynamic changes might complicate the measurement and interpretation of affective responses. However, the static nature of images also presents a downside. It's often argued that the affective experiences they yield are not as potent or realistic as those prompted by dynamic stimuli, such as movie clips.

Movies possess the standardization advantages of pictures, with the added benefit of being dynamic, thereby more closely mirroring real-life scenarios. Films provide a multimodal stimulation experience by engaging both the visual and auditory senses of viewers. Over time, films have demonstrated their versatility as emotion elicitors in studying diverse mood induction phenomena, with capabilities ranging from inducing basic emotions like fear or disgust to more nuanced feelings.

Movies also serve as a rich source of intensive longitudinal data through their dialogue, which provides insights into the emotional and mental states of the characters. Movie dialogues often contain an array of emotion-related words that resonate with the audience and trigger emotional responses through empathy.

This case study seeks to trigger empathic reactions and emotional contagion within a group of subjects using a naturalistic continuous stimulation paradigm, exemplified by the movie "Forrest Gump". This film, abundant in realistic situations and a wide array of emotions, fosters emotional resonance, i.e., the nurturing of empathic responses in viewers, influenced by the narrative decisions in the movie. Specifically, this case study examines the correlation between the language in the film's dialogues and the emotional experiences of viewers. It mainly focuses on testing the effectiveness of two different kinds of linguistic representations, an explicit linguistic profile (refer to Section 2.2.2) and an implicit vectorial representation (refer to Section 3.3), in predicting the emotions elicited in a group of subjects during the viewing of the movie.

2.1.3 Case study 3: The role of linguistic features on the activity of the Autonomic Nervous System

The primary objective of this case study is to investigate the relationship between the linguistic structure of a text and the physiological and acoustic features used to evaluate Autonomic Nervous System (ANS) activity and speech production prosody.

The ANS is responsible for the physiological aspects of emotional regulation, as it manages bodily functions and plays a key role in initiating emotional responses. Electrodermal activity, an extensively researched ANS correlate of emotional arousal, quantifies variations in the skin's electrical conductivity due to sweat gland activity controlled by the sympathetic branch of the ANS. This offers objective measurements of emotional states, providing vital insights into the physiological expressions of emotions.

In addition to the ANS, the intricate process involving somatic regulation also governs speech production. Producing human speech involves fine control of multiple muscles from respiratory, laryngeal, pharyngeal, palatal, and orofacial groups. Changes in respiratory activity induced by the ANS mirror changes in the speaker's emotional state by influencing voice spectrum characteristics. Hence, the examination of speech prosody provides crucial insights into the emotional context and intentions embedded within spoken communication.

With this in mind, this case study seeks to understand how the linguistic structure of spoken text influences speech prosody and ANS correlates. Examining this influence allows for a deeper understanding of the dynamics of language and emotions and their interconnectedness. The relationship between EDA-related features, speech prosody, and the linguistic profile of a text is analyzed using correlation and regression methods. Additionally, the study conducts a complementary analysis from a different perspective, examining the feasibility of using speech and EDA features to predict the linguistic structure of the spoken text.

2.2 Textual Modality

Text data forms the backbone of myriad applications spanning diverse domains of machine learning and artificial intelligence. Despite its immense utility, the idiosyncratic, unstructured nature of text data presents a unique set of challenges when one attempts to use this form of data as input for machine learning models. The biggest challenge is the development of effective methodologies to transform raw, unprocessed texts into a structured format that can be readily consumed by machine learning algorithms. This necessity for transformation arises from the inherent discrepancy between the design of traditional machine learning algorithms and the nature of text data.

Machine learning algorithms, in their conventional form, are crafted to handle structured, numerical data. As such, transforming unstructured, free-flowing text data into a numerical representation – a process frequently referred to as text vectorization – becomes a pivotal step in the data preparation phase for analysis. Over the years, numerous techniques have been proposed and developed to undertake this crucial task, each bearing its own set of advantages and disadvantages. This piece of work organizes the diverse array of text vectorial representations into two major groups, namely, "explicit features" and "implicit features". This categorization is done based on the techniques employed for feature extraction and the subsequent level of interpretability of the extracted features.

Explicit features have the advantage of being easily interpretable. Their primary focus lies on the grammatical function of words within a given piece of text, shedding light on various levels of linguistic phenomena. However, a glaring limitation of these features is their inability to sufficiently capture lexical and semantic information encoded in a text. Furthermore, the extraction process of these features necessitates additional pre-processing and parsing steps to be executed on the text. These steps typically involve breaking down the text into smaller components (such as sentences, words, and tokens) and subsequently analyzing them based on some syntactic formalism. The subsequent sections will delve into different types of explicit features in more detail. In particular, Section 2.2.1 will detail the traditional approaches for text vectorial representations, while Section 2.2.2 will detail the linguistic profiling approach, which has been used in all of the presented case studies to represent text.

On the other hand, implicit features are more challenging to interpret. They represent each word in a text with a high-dimensional vector, typically featuring anywhere from 50 to 300 components based on the task at hand and the dataset being utilized. These vectorial representations are learned through mathematical operations performed by a neural model. Despite the interpretation challenges that these features pose, implicit features can offer a near approximation of relationships between words (e.g., gender, verb tense) and take into account the context and surrounding words that accompany an individual word. Such features are capable of representing lexical and semantic relations between words, offering insights that explicit linguistic features may struggle to capture. The implicit features used in this work will be further described in Section 3.3.

2.2.1 Traditional text representations

One of the simplest and most straightforward methods for representing text in a numerical format that algorithms can better understand is one-hot encoding.

In this type of vectorization, each unique word in a given corpus is represented as a distinct binary vector. Formally, given a corpus C with a vocabulary of words $V = \{w_1, w_2, w_3, ..., w_n\}$, a unique word w_i can be represented with an n-dimensional vector $\mathbf{v}$, where each dimension of w_i is either 1 or 0:

$$w_i = \begin{cases} 1 & \text{if } w = w_i \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

In the vector $\mathbf{w}$ the value at the i-th position is 1, while all the other values are 0. The position that has the value 1 corresponds to the index of the word w_i in the vocabulary V. Each unique word in V has a corresponding one-hot encoded vector.

Although straightforward and previously widely used, the one-hot encoding method exhibits limitations. This approach can lead to large and sparse vector space for large vocabularies, a phenomenon also known as the “curse of dimensionality”, i.e., the exponential increase in computational complexity with each additional dimension. Since each unique word requires a new dimension, the space needed can quickly grow to be excessively large for big corpora [138]. Furthermore, it fails to capture the semantic relationships between words, as all words are equally distant from one another in the vector space [200].

Another commonly used method to represent text as input features is the Bag-of-Words approach (BoW). Originating from information retrieval science, the BoW model reduces text data to a *bag* of its constituent words, disregarding the order and context of the phrase, but maintaining their frequency [123]. In the BoW model, a text corpus is first transformed into a vocabulary of unique words. Subsequently, each document or piece of text is represented as a vector in a high-dimensional space, where the dimensionality equals the size of the vocabulary. The value in each dimension corresponds to the frequency of the word in the document [184]. The feature vector computation of the BoW model can be improved by substituting the simple word frequencies with the Term Frequency-Inverse Document Frequency (TF-IDF) weighting [174, 239]. This technique weights word frequency in a document against its frequency in the entire corpus, thereby highlighting words that are particularly significant in the document of interest.

Despite its simplicity, the BoW model is highly effective in applications such as information retrieval, spam filtering, or text classification. However, it can incur the same limitations as the one-hot encoding approach, being subject to high-dimensionality and sparsity of the vector space, leading to computational challenges. Moreover, by disregarding the order and context of words, the model loses the semantic structure of the language.

Besides the approaches that extract features from the text by looking at single words, it is also possible to rely on approaches that take sequences of words into consideration. It is the case of the n-gram model, that breaks a text corpus into chunks of n consecutive words. The choice of n depends on the scope of the model and influences the balance between the model complexity and its ability to capture context [56]. Indeed, n-gram models have the advantage of capturing more complex sequences of words (e.g., *San Francisco*), and when n-grams contain more than two words they can encode basic syntactic structures and gain more contextual information.

2.2.2 Explicit linguistic profile

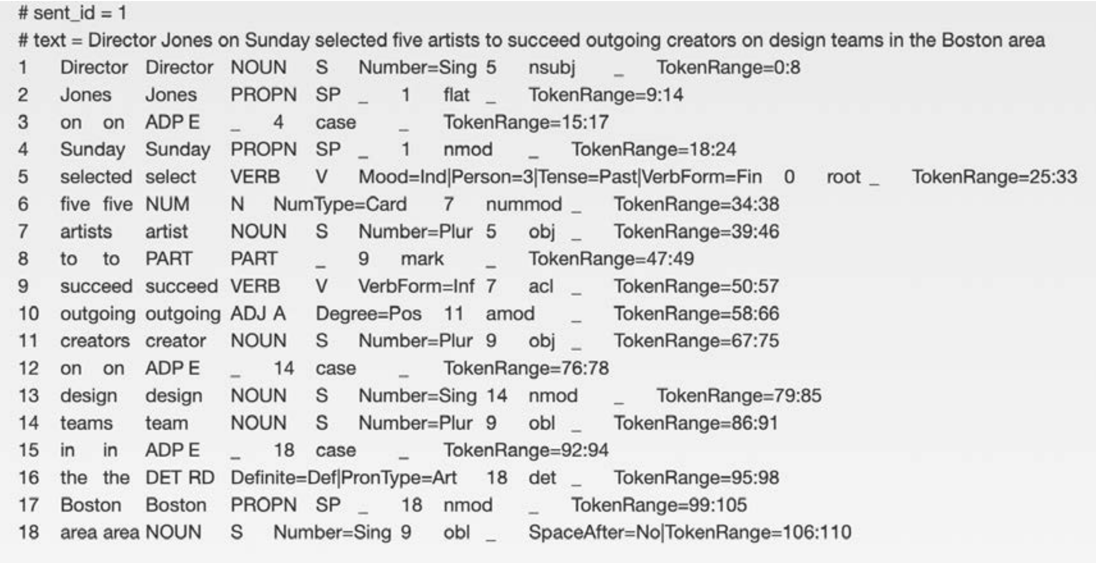
Apart from exploiting the statistical data that can be directly derived from the words and sentences of a text, many Natural Language Processing and Machine Learning applications can benefit from representations that encode information about the syntactic and semantic framework of a text. Specifically, in this work, this kind of information is encoded in a linguistic profile (henceforth referred to also as "explicit features").

The creation of a linguistic profile is based on various degrees of linguistic annotation, through which it is possible to derive a large amount of features. These features model lexical, grammatical, and semantic elements that encapsulate language variations in and across multiple texts. In this methodology, a variety of linguistic features are enumerated to identify and quantify both dissimilarities and resemblances among texts, which depict unique language types [120]. The linguistic composition of a text is analyzed to isolate significant features and a representation of a text is then created based on the statistical occurrences of these features, using either absolute or relative frequencies, or more detailed statistics.

Currently, this methodology is employed in various contexts and research areas with the common goal of recreating the linguistic profile innate to specific linguistic productions. These may originate in distinct contexts such as socio-culturally defined demographic groups or individual authors. Linguistic profiling facilitates the extraction of "meta-knowledge" from text [67]. This implies understanding which features are present and how they mix within a specific language variety as opposed to another of a similar nature. Therefore, the process of extracting meta-knowledge entails associating the feature-based representation of texts with a functional context, a class of speakers or receivers, or individual authors.

In recent years, several studies have concentrated on creating profiling features that encapsulate register, stylistic, and linguistic complexity attributes [210]. Some studies tackled features based on morphosyntactic and syntactic structures [11, 216], while others select features based on context-free grammar rules [32]. Other studies develop more sophisticated systems that allow the derivation of extensive linguistic properties spread across different levels of linguistic annotation [42].

The explicit linguistic features used in this work are extracted with Profiling-UD [42], a web-based application that performs linguistic profiling of a text, or



```
# sent_id = 1
# text = Director Jones on Sunday selected five artists to succeed outgoing creators on design teams in the Boston area
1	Director	Director	NOUN	S	Number=Sing	5	nsubj	_	TokenRange=0:8
2	Jones	Jones	PROPN	SP	_	1	flat	_	TokenRange=9:14
3	on	on	ADP	E	_	4	case	_	TokenRange=15:17
4	Sunday	Sunday	PROPN	SP	_	1	nmod	_	TokenRange=18:24
5	selected	select	VERB	V	Mood=Ind|Person=3|Tense=Past|VerbForm=Fin	0	root	_	TokenRange=25:33
6	five	five	NUM	N	NumType=Card	7	nummod	_	TokenRange=34:38
7	artists	artist	NOUN	S	Number=Plur	5	obj	_	TokenRange=39:46
8	to	to	PART	PART	_	9	mark	_	TokenRange=47:49
9	succeed	succeed	VERB	V	VerbForm=Inf	7	acl	_	TokenRange=50:57
10	outgoing	outgoing	ADJ	A	Degree=Pos	11	amod	_	TokenRange=58:66
11	creators	creator	NOUN	S	Number=Plur	9	obj	_	TokenRange=67:75
12	on	on	ADP	E	_	14	case	_	TokenRange=76:78
13	design	design	NOUN	S	Number=Sing	14	nmod	_	TokenRange=79:85
14	teams	team	NOUN	S	Number=Plur	9	obl	_	TokenRange=86:91
15	in	in	ADP	E	_	18	case	_	TokenRange=92:94
16	the	the	DET	RD	Definite=Def|PronType=Art	18	det	_	TokenRange=95:98
17	Boston	Boston	PROPN	SP	_	18	nmod	_	TokenRange=99:105
18	area	area	NOUN	S	Number=Sing	9	obl	_	SpaceAfter=No|TokenRange=106:110
```

Figure 2.1: Example of linguistic annotation in CoNLL-U format as performed by the UDPipe tool on the sentence "*Director Jones on Sunday selected five artists to succeed outgoing creators on design teams in the Boston area*".

a collection of text, for multiple languages. The tool extracts the features with a two-process that involves first a linguistic annotation stage and then a linguistic profiling stage.

Linguistic Annotation Stage

The linguistic annotation is automatically performed by UDPipe [262], a state-of-the-art trainable pipeline tool that applies tokenization, morphological analysis, part-of-speech tagging, and dependency parsing to the text. UDPipe is conceived to be language-agnostic, producing annotations that use a taxonomy that stays the same across different languages. Specifically, annotations follow the CoNLL-U Format [212, 211], and consist of the following fields:

- ID – word index, starting at 1 for each new sentence. It can be a range in the case of multi-word tokens.
- FORM – word form or punctuation symbol
- LEMMA – lemma or stem of the word form
- UPOS – Universal Part-Of-Speech (POS) tag
- XPOS – Language specific POS tag, underscore (_) if not available.
- FEATS – List of morphological features.

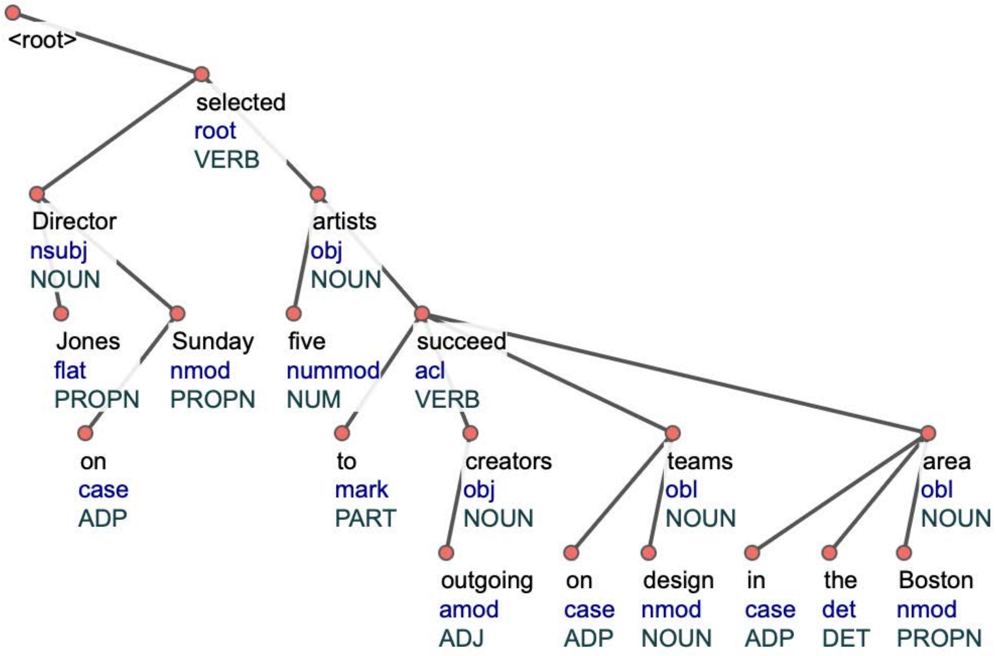


Figure 2.2: Graphical visualization of the linguistic annotation performed by the UD-Pipe tool on the sentence "*Director Jones on Sunday selected five artists to succeed outgoing creators on design teams in the Boston area*".

- HEAD – Head of the current word. Either an ID value or zero (0).

- MISC – Any other annotation that does not fit into the preceding categories.

Linguistic Profiling Stage

In the linguistic profiling stage, the different levels of annotations produced in the previous stage are automatically analyzed by a linguistic profiling component, based on a set of scripts written in Python. The component captures a vast number of linguistic phenomena, defining the rules to extract and quantify the formal properties of the texts. It extracts over 140 features, each belonging to one of the following categories: (1) raw text properties, (2) lexical variety, (3) morphosyntactic information, (4) verbal predicate structure, (5) global and local parse tree structures, (6) syntactic relations, (7) use of subordination.

Hereafter, each category is described in detail, along with the list of features that belong to it. For each feature, there is a description of how it has been extracted and quantified from the UD representation created in the annotation stage. To exemplify some of the features, the following sentence will be taken as a reference:

> "*Director Jones on Sunday selected five artists to succeed outgoing creators on design teams in the Boston area*".

The output the tool gives for this sentence is shown in Figures 2.1 and 2.2. Figure 2.1 shows the linguistic annotation in CoNLL-U format as performed by UDPipe, while Figure 2.2 shows the graphical visualization of the annotation and the structure of the analyzed sentence. The complete list of features and their acronyms and abbreviations as used in this work is reported in Appendix A.

1. Raw Text Properties

- *Sentence length.* Average length of sentences in a text, calculated as the average number of tokens per sentence.

- *Word length.* The average number of characters per word, excluding the punctuation.

2. Lexical Variety

- *Type/Token Ratio (TTR).* A standard metric to assess the lexical variety of a text. It is computed as the ratio between the number of lexical types[1] and the number of tokens in a text. The higher the value, the more different terms there are in a text. A lower value indicates the text contains many repeated words. The tool calculates this feature for the first 100 and 200 tokens of a text.

3. Morpho–syntactic information

- *Distribution of grammatical categories.* The percentage distribution in the text of the 17 core part-of-speech categories as defined in the Universal POS tagset[2]. The tagset is subdivided into open class words (i.e. adjective, adverb, interjection, noun, proper noun, verb), closed class words (adposition, auxiliary, coordinating conjunction, determiner, numeral, particle, pronoun, subordinating conjunction), and a class 'other' used for punctuation and symbols.

- *Lexical density.* The ratio of content words (verbs, nouns, adjectives, and adverbs) over the total number of words in a text.

- *Inflectional morphology.* For each verb and auxiliary, calculates the distribution of the following inflectional features: mood, number, person, tense, and verb(al) form.

[1] A lexical type is the base form of a word, i.e., the form in which that word is listed in a dictionary.

[2] `https://universaldependencies.org/u/pos/` (retrieved 3 July 2023).

4. Verbal Predicate Structure

- *Distribution of verbal heads.* The average number of verbal heads in a sentence. The number of verbal heads corresponds to the number of propositions appearing in a sentence, whether they are main or subordinate propositions.

- *Distribution of verbal roots.* The percentage of verbal roots out of the total roots within a sentence.

- *Verbal arity.* The number of dependency links sharing the same verbal head, excluding punctuation and copula UD dependencies, e.g. a value of arity=2 for a verb, means that the verb is the head of two dependency links. In the reference sentence, the average arity score is 2, since the main verb ‘selected’ has four dependents (Director, Sunday, artists, succeed), the first embedded verb ‘succeed‘ has two dependents (to, creators), and the gerund verb ‘outgoing’ has no dependents.

5. Global and Local Parsed Tree Structures

- *Average depth of the syntactic tree.* The mean of the maximum depths extracted for each sentence in a text. The maximum depth of a sentence is calculated as the longest path (in terms of dependency links) from the root of the dependency tree to some leaf. In the reference sentence, this feature is equal to 5. This number corresponds to the five intermediate dependency links that are crossed in the path going from the root of the sentence (‘selected’) to each of the equidistant leaf nodes, represented by the words ‘in’, ‘the’, and ‘Boston’.

- *Average clause length.* Calculated as the average number of tokens per clause. The number of clauses is the ratio between the number of tokens in a sentence and the number of verbal or copular heads.

- *Length of dependency links.* Calculated as the number of words occurring between the syntactic head and its dependencies. This information is complemented with the feature *Maximum dependency link* corresponding to the average length of the longest dependency link for each sentence in a given text. In the reference sentence, there are 17 dependency links. Hereafter all of them are reported, with the head of the link highlighted in bold. Eight links have a one-token distance: [‘**Director**’, ‘Jones’], [‘on’, ‘**Sunday**’], [‘Sunday’, ‘**selected**’], [‘five’, ‘**artists**’], [‘to’, ‘**succeed**’], [‘outgoing’, ‘**creators**’], [‘design’, ‘**teams**’], [‘Boston’, ‘**area**’]. Four links have a two-token distance: [‘**selected**’, ‘artists’], [‘**succeed**’, ‘creators’], [‘on’, **teams**], [‘the’, ‘**area**’]. Two links have a three-token distance: [‘**creators**’, ‘teams’], [‘in’, ‘**area**’]. Three links show the maximum four-token distance: [‘Director’, ‘**selected**’], [‘**selected**’, ‘succeed’], [‘**teams**’, ‘area’]. The average value, calculated as the ratio between the sum of all distances over the total number of links, is two.

- *Average depth of embedded complement chains governed by a nominal head.* The average depth of embedded complement chains, i.e. a list of consecutive complements (either prepositional complements or nominal and adjectival modifiers) sharing the same nominal head. The value of this feature corresponds to the average depth of complex nominal chains extracted from all sentences in a given text. In the reference sentence, the depth of the nominal chain headed by the noun ‘creators’ is equal to 2; as visible in the graphical representation in Figure 2.2, the chain covers two embedded prepositional modifiers (‘on design teams’ and ‘in the Boston area’), which are both governed by the noun ‘creators’.

- *Word order phenomena.* This feature applies only to the subject and the object of the sentence. It captures the relative order of subject and object with respect to the verb and captures the probability distribution of such occurrence, i.e. the subject and the object can appear in pre-verbal (before the verb) or post-verbal (after the verb) position. This feature is particularly useful for capturing word order variation across different languages and, within the same language, across varieties of language use.

6. Syntactic Relations

- *Distribution of dependency relations.* The distribution (in percentage) of the 37 universal relations in the UD dependency annotation scheme.

7. Subordination phenomena

- *Distribution of subordinate and main clauses.* The percentage distribution of main clauses against subordinate clauses, as defined in the UD scheme[3].

- *Relative order of subordinates with respect to the verbal head.*: The percentage distribution of subordinate clauses in post-verbal and pre-verbal positions.

- *Average depth of embedded subordinate clauses.* Given the subordinate clause sub–tree, a subordinate ‘chain’ is calculated as the number of subordinate clauses recursively embedded in the top subordinate clause. In

[3] https://universaldependencies.org/u/overview/complex-syntax.html#subordination

addition to the average value of the chain depth, the percentage distribution of subordinate chains by depth is also provided. The reference sentence is articulated into a main clause and a subordinate clause governed by the verbal root 'selected'. The adverbial subordinate clause headed by the verb 'succeed' occurs in a post-verbal position and does not contain embedded subordinates.

2.3 Audio modality

So far, language as a means of communication has been described only in terms of its textual representation. However, humans' primary mode of communication is through speech and the ability to communicate through speech is what differentiates humans from animals.

Speech is a complex activity that involves both neural and physical activation. When a person wants to speak, their brain constructs a sentence with the desired meaning and then maps the sequence of words into the movements required to produce the sounds associated with the words. From a physical point of view, speech relates to the activation of three components: the respiratory system, the phonatory system, and the resonatory system, as illustrated in Figure 2.3. The production of sounds begins in the respiratory system, or breathing apparatus, which acts as a compressor by activating the contraction of the lungs. The lungs are supported by the diaphragm, a *shelf* of muscles and tendons extending across the bottom of the ribcage. When air is inhaled, the diaphragmatic muscles contract, shortening and tightening, and the diaphragm moves downward in the body, allowing it to create a vacuum in the lungs that are then filled with air. During exhalation, the diaphragm relaxes and rises, decreasing the volume of the lungs and creating a positive pressure difference, pushing the air outside. The air passes then through the throat, activating the phonatory system. In the throat, the air encounters the vocal folds, that act as a sound generator: through vibrations, they chop the airstream from the lungs into a sequence of air pulses. Finally, the resonatory system is activated when the air passes the oral and nasal cavities, that act as a resonator (or a *filter*) by shaping the sound generated by the vocal folds.

Airflow by itself is not audible, as sound is the consequence of an oscillation in air pressure. For airflow to produce a sound, it must be obstructed to obtain an oscillation or turbulence. Oscillations are mainly produced in the phonatory system when air passes through the vocal folds. Vocal folds (or vocal cords or voice reeds) are folds of throat tissues that can be tensioned to reduce or augment the space between them (the glottis). Oscillations that produce sounds can also happen in other parts of the speech-production organs. For example, they can be caused by the tongue's movements or the oscillations of the uvula caused by the airflow.

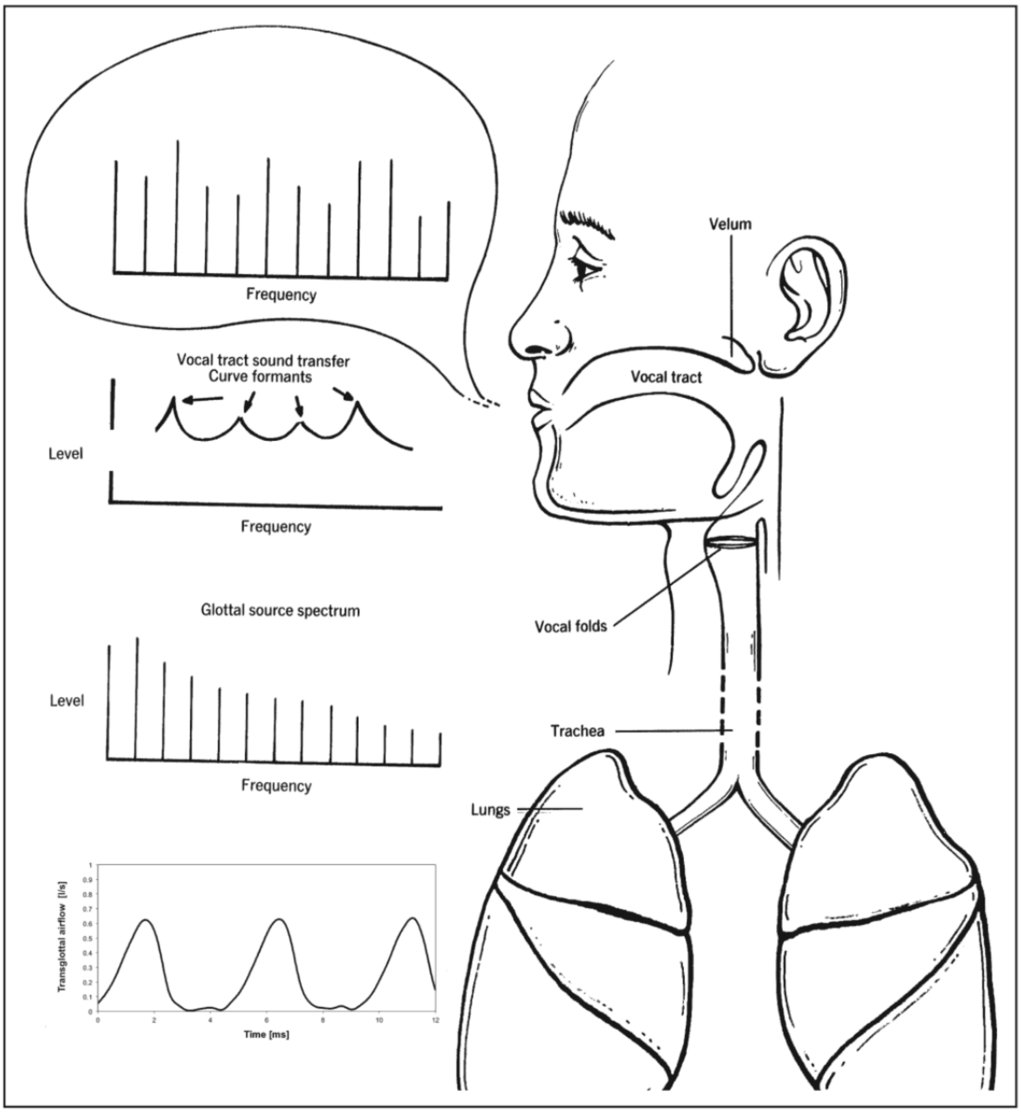


Figure 2.3: The functioning of the voice organ. *[Image source [227]].*

The vocal tract (composed of the larynx, the pharynx, and the oral cavities) has an essential effect on the timbre of the sound produced. In particular, the shape of the vocal tract determines the acoustic space's resonances and anti-resonances, boosting or attenuating the different frequencies of the sound. The resonances can be easily modified by the speaker, an act that is called *articulation*. Instead, the structures used to arrange the shape of the vocal tract are called *articulators*, e.g., the tongue or the mandible. Changes in resonance can be easily perceived by the listener, contributing to conveying information in communication. These changes in resonances can be detected by acoustic signal analysis and divided into different features.

2.3.1 Speech-based features

Features that can be extracted from speech can be divided into four categories:

1. Teager Energy Operator (TEO) Based Features
2. Voice Quality Features
3. Spectral Features
4. Prosodic Features

Different combinations and selections of these features have been applied in various studies where speech was used as a modality for emotion recognition [1, 5, 144].

Teager Energy Operator Based Features

These features are reliant on the Teager Energy operator, a tool used for detecting stress in speech, first introduced by Teager and Teager [270] and further detailed by Kaiser [139, 140]. Teager's theory suggests that speech is produced by a non-linear interplay of vortex and airflow within the human vocal system. When under stress, the speaker's muscle tension changes, leading to adjustments in the airflow during the creation of sound. The operator developed by Teager to gauge the energy of speech through this non-linear method was captured by Kaiser in the following formula

$$\Psi[X(n)] = x^2(n) - x(n+1)x(n-1) \tag{2.2}$$

where Ψ represents the Teager Energy Operator and $x(n)$ denotes the sampled speech signal.

Three new TEO-oriented features were proposed by Zhou et al. [313]: the TEO-decomposed FM (frequency modulation) variation (TEO-FM-Var), normalized TEO auto-correlation envelope area (TEO-Auto-Env), and the critical band based TEO auto-correlation envelope area (TEO-CB-Auto-Env). These features investigate the energy variation in airflow attributes in the vocal tract for stress-induced voiced speech. Zhou et al. [313] contrasted these with pitch and Mel Frequency Cepstral Coefficients (MFCC) features, using both text-dependent and independent stress classifications via the SUSAS dataset [122]. The performance of TEO-FM-VAR and TEO-AUTO-ENV was found to be subpar compared to traditional pitch and MFCC features. However, the TEO-CB-Auto-Env feature surpassed both pitch and MFCC under stress conditions. Comparable outcomes were reported by Low et al. [172] in their research on detecting clinical depression in adolescents. They used a combination of prosodic, spectral, voice quality, and TEO-based characteristics. Among all these, TEO-based features, particularly TEO-CB-Auto-Env, excelled beyond the rest, including their combinations.

Voice Quality features

Voice quality parameters describe the properties of the glottal source, that is to say, they are influenced by the physical characteristics of the vocal tract. They include parameters such as jitter, shimmer, and the harmonics to noise ratio. A strong link exists between voice quality and the emotional context of speech [65].

Jitter and shimmer. Jitter refers to the fluctuation in the fundamental frequency from one vibratory cycle to the next, while shimmer pertains to amplitude variation. Jitter gauges frequency instability, while shimmer measures amplitude instability.

Harmonics to noise ratio. The harmonics to noise ratio quantifies the relative noise level in a vowel's frequency spectrum, signifying the ratio between the periodic and aperiodic components in voiced speech signals. These fluctuations are often interpreted as voice quality changes.

Other quality metrics. Other metrics that have been utilized in studies include the Normalized Amplitude Quotient (NAQ), Quasi Open Quotient (QOQ), the variance in the amplitude of the first two harmonics in the differentiated glottal source spectrum (H1H2), Maxima Dispersion Quotient (MDQ), the spectral tilt or slope of wavelet responses (peak-slope), Parabolic Spectral Parameter (PSP), and the shape parameter of the Liljencrants-Fant model of glottal pulse dynamics (Rd) [277].

Spectral features

The sound produced by an individual is shaped and filtered by the form of their vocal tract, which shape significantly impacts the resulting sound. The vocal tract has been described as a close tube resonator [267], i.e., a closed cylindrical air column that produces resonant standing waves[4] at a fundamental frequency and at odd harmonics. The size and shape of the vocal tract and the resonance cavities will select some of the frequencies in the sound and diminish the range of other frequencies (such as what would happen for a clarinet or a flute). When a sound is fed to a resonator, the frequencies of that sound are going to be limited by the resonator cavities. The spectrum of the resulting available frequencies is called a *formant*. Spectral features are obtained by converting the time domain signal into the frequency domain signal using Fourier transform. They are extracted from speech segments ranging from 20 to 30 milliseconds, partitioned through a windowing method.

[4] Waves with characteristic patterns that arise from the combination of reflection and interference, such that the reflected waves interfere constructively with the incident waves.

Mel Frequency Cepstral Coefficients (MFCC). MFCC feature illustrates the short-term power spectrum of the speech signal. The extraction of MFCC involves dividing utterances into segments, transforming each segment into the frequency domain using short time discrete Fourier transform, calculating sub-band energies with a Mel filter bank, logging these sub-bands, and finally applying the inverse Fourier transform [152].

Linear Prediction Cepstral Coefficients (LPCC). LPCC also encapsulate vocal tract features of speakers, which show variations with distinct emotions. LPCC can be directly derived from Linear Prediction Coefficient (LPC) using a recursive method. LPC represents the coefficients of all-pole filters and equates to the smoothed envelope of the speech log spectrum [301].

Log-Frequency Power Coefficients (LFPC). LFPC replicate the logarithmic filtering attributes of the human auditory system by measuring spectral band energies with Fast Fourier Transform [213].

Gammatone Frequency Cepstral Coefficients (GFCC). GFCC is obtained similarly to MFCC extraction, but using a Gammatone filter-bank on the power spectrum instead of a Mel filter bank.

Formants (resonance of the vocal tract). Formants refer to the resonance frequencies of the vocal tract and are calculated as amplitude peaks in the sound's frequency spectrum. The human vocal system has two main resonance cavities: the mouth and the larynx. The positions assumed by the larynx and the mouth during speech influence sound production, so each sound results in different formants. The formants are referred to as F_1, F_2, F_3, and so on, according to the frequency at which a specific sound resonates. By convention, the lower the number associated with the formant, the lower the frequency at which it resonates. Roughly, an average man shows a different formant every 1000Hz, a range that shifts to 1100Hz for the average woman. However, the true range depends on the dimensions (length) of the vocal tract.

Prosodic features

Prosodic features refer to elements in speech that are perceptible to the human ear, including intonation and rhythm. An example of this kind of features is the rising intonation at the end of a sentence to signal a question. For instance, in the sentence "You are coming tomorrow?", the upward inflection on the term "tomorrow" implies it's a question. Such elements, known as paralinguistic features, pertain to larger speech units such as syllables, words, phrases, and sentences.

As these features are drawn from more extensive speech segments, they are classified as long-term features. According to research, prosodic elements are key in revealing the unique emotional aspects in the field of speech emotion recognition [310]. The three most commonly applied prosodic features are those based on fundamental frequency, energy, and duration.

The fundamental frequency. Also denoted by $F0$ or F_0, it refers to the approximate frequency of the (quasi-)periodic structure of the voiced speech signal. In other words, the fundamental frequency is the average number of oscillations per second expressed in *Hertz*. The oscillation of the speech signal originates in the vocal folds. However, since it originates from an organic structure, it cannot be precisely periodic: it contains significant fluctuations (such as variations in period length, *jitter*, or variations in amplitude, *shimmer*). Moreover, F_0 is usually non-stationary within a sentence but changes repeatedly. Indeed, the fundamental frequency can be used for expressive purposes in speech, for example, to add emphasis or model a question. The typical values for the fundamental frequency lay between 80Hz and 450Hz, with males having lower frequencies than females and children. The fundamental frequency of an individual speaker varies according to the length of the individual's vocal cords, which is also correlated to body size. The fundamental frequency is also closely related to *pitch*, which is how our ears and brain perceive and interpret the signal coming from the F_0.

Energy. The energy contained in the speech signal, often termed as volume or intensity, provides a depiction of the amplitude fluctuations in speech signals over time. It is theorized by researchers that heightened emotional states such as anger, joy, or surprise result in an escalation of this energy, while feelings of disgust and sorrow lead to a decline in energy [166].

Duration. Duration refers to the time required to construct speech elements like vowels, words, and other similar constructs. Key features related to duration include the rate of speech, the length of silent intervals, the speed of voiced and unvoiced sections, and the duration of the longest spoken voice.

2.3.2 Analysis of human voice

The analysis of human voice is a challenging task, due to the considerable variations influenced by factors such as gender, age, health conditions, and emotional states.

The recent advancement and increased dependability of computer systems have facilitated the creation of numerous voice analysis tools and software. One of the initial and most commonly utilized tools was the Multi Dimensional Voice Program (MDVP) [275]. The software was developed and commercialized by

Kay Elemetrics Corporation (now Pentax Medical) in the 1990s. The software, requiring the use of a Computerized Speech Lab (CSL), provides a comprehensive set of over 35 quantitative voice parameters, including Frequency Perturbation, Amplitude Perturbation, and Noise measures.

MDVP is widely used for the analysis of dysphonia, a condition that causes changes in voice quality, including hoarseness, breathiness, or roughness. It's capable of producing numerical data that can help detect subtle changes in voice quality, therefore being a valuable tool for clinicians and speech pathologists. One of the unique aspects of MDVP is its ability to provide multi-dimensional voice assessments. This means it can measure and evaluate multiple aspects of voice quality simultaneously, providing a more complete picture of a person's vocal health. Despite its usefulness, it does have some limitations. The cost of the license for the software can be high, which may be a barrier for some users. Also, the earliest versions of the program were criticized for their lack of usability, particularly for non-expert users. Over time, newer versions and alternative software have improved on this aspect.

Other prominent tools, such as DrSpeech [77] and Wevosys [297], have found value in complementing traditional speech therapy approaches and assessing voice range profiles.

Dr. Speech is a comprehensive software suite used for a wide range of speech and voice assessment and analysis tasks. It is designed for use by speech-language pathologists, voice coaches, scientists, and clinicians. The software features multiple modules, each targeted at a different aspect of voice and speech analysis. This can include speech acoustics, perception, and physiology. It enables users to record, play back, and analyze voice samples, offering tools to examine aspects such as pitch, volume, voicing, and nasality. One of the strengths of Dr. Speech is its flexibility. It offers a wide range of tools and analytical functions, which can be used for many different applications. For example, it can be used to help evaluate and treat voice disorders, to provide biofeedback for speech therapy patients, or to analyze vocal performances.

Wevosys, or Voice System GmbH, is a company that develops and provides innovative solutions for speech analysis, voice assessment, and therapy. They offer LingWAVES, a comprehensive suite of tools for the evaluation, analysis, and therapy of voice, speech, and language disorders. One of the key features of LingWAVES is its broad range of applications. It can be used in several areas such as speech-language pathology, voice coaching, and scientific research. The system is designed to assess various parameters of voice and speech, such as fundamental frequency, intensity, and spectral characteristics. It also includes modules for voice range profile measurement, perturbation analysis, and Electroglottography analysis, among others. One important advantage of LingWAVES is its user-friendly interface. The software presents complex voice and speech data in an easy-to-understand visual format. This aids clinicians in explaining voice phenomena to their patients and helps patients understand and visualize their progress throughout therapy. LingWAVES also includes tools that provide visual biofeedback to patients during therapy. This can be particularly beneficial in vocal training and rehabilitation, as it enables patients to visualize their voice production in real-time and adjust their technique accordingly.

The market has also seen the emergence of a variety of open-source tools, including VoceVista [289], Wavesurfer [294], Speech Analyzer [257], WASP [300], and VAT [288]. These tools specialize in fundamental frequency, spectrographic and spectral analysis, with some being especially apt as feedback tools for singing practice and for recording vocal progression throughout voice training or therapy.

VoceVista is an educational software tool designed primarily for singers, voice teachers, and vocal researchers. It provides visual feedback and analyses to help improve singing technique and understand the voice better. The tool allows to view and analyze the overtone structure (i.e., the formant) of vocal sounds, which is particularly useful for understanding and teaching vocal technique. The tool includes a spectrogram (a visual representation of the spectrum of frequencies in a sound) and a scope to show pitch, giving a real-time or recorded visual representation of the voice. It allows for pitch tracking, showing the fundamental frequency, and the intensity of the overtones (harmonics) relative to each other.

WaveSurfer is a tool designed for sound visualization and manipulation. It was developed primarily for the linguistic research community, but its flexibility and modular design allow it to be useful in various other areas such as music, speech therapy, and education. The software provides a range of functionalities, including waveform visualization, spectrogram analysis, pitch tracking, and formant analysis. It allows users to analyze and manipulate audio files in various formats, making it a versatile tool for different audio analysis tasks. One of the primary advantages of WaveSurfer is its extensibility. It was developed with a plug-in architecture, allowing developers to extend its functionality with new custom-made plug-ins.

Speech Analyzer is a computer program developed by SIL International, a nonprofit organization committed to serving language communities worldwide. The software is designed for acoustic analysis of speech sounds and is often used by linguists, phoneticians, and other language professionals. Speech Analyzer offers several key features that enable the detailed study and analysis of speech. It allows users to perform waveform, spectrogram, and pitch analyses, providing a multifaceted view of speech data. With the waveform analysis, users can observe the shape and intensity of the sound waves. The spectrogram analysis, on the other hand, gives a visual representation of the spectrum of frequencies of a signal as it varies with time. This can be useful for studying the spectral characteristics of different speech sounds. Pitch analysis, meanwhile, can be used to study the melody or intonation of speech. Moreover, the software allows for the annotation of recordings, making it easier to study specific segments of speech. Users can also slow down or speed up the playback without altering the pitch, which can be beneficial for careful analysis of fast speech sounds.

WASP (Waveform and Spectral analysis Program) is a simple tool used for recording, displaying, and analyzing speech sounds. Developed by Mark Huckvale at University College London, it's commonly utilized in speech and language research. WASP offers multiple features including waveform display, spectrogram display, pitch contour analysis, and intensity analysis.

VAT, which stands for Voice Analysis Toolkit, is a software designed for the analysis of speech and voice. Developed in the MATLAB programming environment, VAT provides a set of algorithms and graphical user interfaces for the analysis of speech signals.

Another widely used tool among top-tier researchers is Praat [35], developed by Paul Boersma and David Weenink of the University of Amsterdam. It provides a variety of features to analyze speech or other acoustic signals, with its key functionalities including Spectrogram, Pich, Formant, and Intensity analysis, annotation, and speech syntesis. While Praat is a powerful tool with an extensive range of functionalities, its interface might not be immediately intuitive to beginners.

The BioVoice tool

Recently, Morelli et al. [205] have introduced BioVoice, a software tool for voice analysis based on new algorithms that make it particularly suitable also for high-pitched and quasi-stationary voices, such as signing voice, children vocalizations, and newborn and infant cry.

In this work, the BioVoice toolbox will be applied in the case study of Chapter 6 to estimate the parameters of the voice of a group of subjects reading some texts out loud. For this purpose, the tool estimates the following parameters:

- the fundamental frequency F_0;
- the formants F_1, F_2, F_3;
- signal duration, voiced duration, and mean duration of the recorded speech signal.

To extract the parameters, the tool first loads the wave file and divides it into non-overlapping windows. The windows' size fluctuates based on the inverse relation to the varying vocal frequency F_0. The length of these frames is linked to the gender of the subject (whose voice is recorded and analyzed), because F_0 is influenced by the dimensions, thickness, and tension of the vocal cords. The chosen frame length is within the boundaries $3F_s/F_{min} \leq M \leq F_s/F_{max}$, where F_s is the sampling frequency of the signal, while F_{min} and F_{max} denote the lowest and highest allowed F_0 values for the evaluated signal (established at 50-250 Hz for males, and 100-350 Hz for females).

Estimation of $\mathbf{F_0}$. The fundamental frequency F_0 is estimated in a two-step procedure. Initially, Simple Inverse Filter Tracking (SIFT)[5] is employed on the signal time windows of a set length $M = 3F_s/F_{min}$. This step results in a rough F_0 tracking along with its range of variation $[F_{low}, F_{high}]$, where F_{low} is the lowest F_0 value and F_{high} is the highest F_0 value [58, 256, 274]. Following that, F_0 is computed within $[F_{low}, F_{high}]$ using a blended approach of Short-Time Autocorrelation Function (STACF) and Average Magnitude Difference Function (AMDF).

STACF is a measure of the correlation between a signal and a time-shifted version of itself, calculated over short, overlapping segments of the signal. It provides information about the temporal structure and repeating patterns present in a signal. It is computed on the fixed duration frames in which the input signal is divided. Within each frame, the autocorrelation function is calculated by multiplying the signal with a time-shifted version of itself, and then summing the products over a specified time lag range. STACF is calculated applying this formula:

$$R(k) = \int [s(m) * x(m-k)]dt \tag{2.3}$$

where $R(k)$ is the autocorrelation at time lag k, $s(m)$ is the input signal at sample index m, and $*$ denotes the convolution operation.

For each signal frame, STACF is computed, and the tool searches for the maximum autocorrelation R_{max} and the corresponding cycle length T_{max} in the range $[1/F_{high}, 1/F_{low}]$, with F_{high} and F_{low} estimated in the previous step. If T_{max} is outside the range, the cycle length is recalculated for the same frame utilizing the AMDF. AMDF is a measure used to estimate the periodicity of a signal, computed by comparing the magnitude differences between a signal and its delayed versions at different time lags. The basic idea is to calculate the average absolute difference between the signal and its delayed versions across a range of

[5] Simple Inverse Filter Tracking is a technique used in signal processing and control systems to estimate the state of a dynamic system based on input-output measurements. The goal of inverse filter tracking is to find an inverse filter that can reproduce the original input signal from the observed output signal. In a simple inverse filter tracking scheme, the inverse filter is designed to model the inverse dynamics of the system. By applying the inverse filter to the observed output signal, the estimated input signal can be obtained. The basic idea behind inverse filter tracking is to exploit the known dynamics of the system and its input-output relationship to estimate the input signal. However, it is important to note that inverse filtering can be challenging in practice due to noise, modeling errors, and limitations in the observability of the system. Inverse filter tracking has applications in various fields, including audio signal processing, image restoration, and control systems. It can be used, for example, to remove noise or distortions from a recorded audio signal by estimating the original clean signal. It's worth mentioning that the term "Simple Inverse Filter Tracking" does not refer to a specific, well-defined technique but rather describes a general approach to inverse filtering. The specific implementation and methodology may vary depending on the application and the characteristics of the system being tracked.

time lags, as shown in the following formula:

$$AMDF(k) = \sum_{m=k}^{N} |s(m) - s(m-k)| \tag{2.4}$$

where $AMDF(k)$ is the AMDF at time lag k, $s(m)$ represents the signal at sample index m, N is the length of the signal, and the summation is performed over the entire signal. In the case of a periodic signal with a period T, the function is expected to have a significant minimum at $k = T$. Consequently, the pitch period is assessed as the one corresponding to the minimum value of the AMDF.

In a second step, the tool removes all unvoiced frames (i.e., frames that do not show a periodical structure, that is to say, frames that have a low periodicity) and generates a new audio file made up of only voiced frames. This step is based on a revised version of the pitch continuity function in the AMPEX algorithm [182].

Estimation of the Formants ($\mathbf{F_1}$, $\mathbf{F_2}$, $\mathbf{F_3}$). The formants are estimated applying Autoregressive Power Spectral Density (AR-PSD) [44]. This method is used to estimate the power spectral density (PSD) of a signal based on an autoregressive (AR) model.

The power spectral density is a representation of the frequency content of a signal and provides information about the distribution of signal power across different frequencies. The AR-PSD estimation approach models the signal as an autoregressive process, where each sample is assumed to be a linear combination of its past samples.

In the autoregressive model, the current sample of the signal is expressed as a weighted sum of the previous samples, with the weights determined by the model's parameters. The AR model is typically represented as follows:

$$x(n) = a(1) * x(n-1) + a(2) * x(n-2) + ... + a(p) * x(n-p) + e(n) \tag{2.5}$$

where $x(n)$ represents the current sample, $a(1)$ to $a(p)$ are the model parameters, $x(n-1)$ to $x(n-p)$ are the previous samples, p is the model order, and $e(n)$ is the error term or noise.

To estimate the AR-PSD, the AR model is fitted to the pre-processed voice signal. The parameters of the AR model are determined using methods such as the Yule-Walker equations [307] or the Burg [44] method.

Once the AR model parameters are obtained, the PSD is computed using the Fourier transform of the AR model's power spectral density function. The PSD represents the distribution of signal power across different frequencies. The formants correspond to the peaks in the estimated PSD. The peaks indicate the resonant frequencies of the vocal tract.

Estimation of Signal, Voiced, and Mean Duration. These parameters pertain to the time domain of the voice recordings and are automatically extracted by the tool at the beginning of the analysis. They simply reflect the length of the whole recorded signal (comprising the voiced and unvoiced sections), the length of the sole voiced sections, and the mean length of the recorded signals.

2.4 Physiological signals modality

As seen in Chapter 1, multiple physiological sources can be measured when assessing changes caused by emotions. Among them, a common, simple, and non-invasive one to measure is Electrodermal Activity (EDA). EDA broadly refers to any variation of the electrical properties of the skin, covering a myriad of potential fluctuations influenced by factors such as emotional state, cognitive load, and even environmental conditions. One of the most frequently used measures of EDA is the *skin's electrical conductance* in response to sweat secretion. By applying a low, constant, undetectable voltage it is possible to non-invasively measure the variance in the skin's conductance.

Electrodermal signals reflect the activity of the eccrine sweat glands, which are stimulated by the sympathetic branch of the Autonomic Nervous System, primarily through the sudomotor nerves [101]. Sweat production, triggered by the sudomotor nerves, leads to alterations in the conductivity measured on the skin's surface. This change is a consequence of both the sweat secretion and the changes in the ionic permeability of the sweat gland membranes [36, 79, 80].

Sweat gland activity is initiated by the postganglionic sudomotor fibers [143, 238]. Each sweat gland is innervated by multiple distinct sudomotor fibers, and in turn, each sudomotor fiber innervates a skin region of approximately 1.28 cm^2. This network of fibers and glands forms a sophisticated system for distributing sweat across the skin's surface. It was calculated that the sudomotor fibers fire at an average rate of 0.62 Hz, i.e., they fire 0.62 times per second. The simultaneous firing of multiple fibers is recognized as a nerve burst in the integrated nerve record. A sudomotor nerve burst corresponds to a visible Skin Conductance Response (SCR). The density of spikes (as indicated by the nerve burst's amplitude in the consolidated nerve record) is directly related to the number of the activated sweat glands and the amplitude of the SCR. Therefore, the SCR amplitude can be considered as a measure of the activity of the Sympathetic Nervous System, which is involved also in emotional response.

While sudomotor activity, i.e., sweat secretion, plays a major role in thermoregulation, the sweat glands found on the palms and the soles may have evolved also to improve gripping ability and increase sensory discrimination. For this reason, these glands potentially respond more to psychologically significant stimuli compared to thermal ones [36, 80].

2.4.1 Decomposition of the Skin Conductance signal

The variations in skin conductance are captured by electrodes that are easy and non-invasive to apply to subjects. Usually, data is acquired with sampling rates between 1-10 Hz and is measured in units of micro-Siemens (μS). The skin conductance signal is characterized by two main components, a *tonic* component (i.e., the skin conductance level, SCL), and a *phasic* component (i.e., the skin conductance responses, SCRs), that differ in temporal scales and relationships with the stimuli that initiated them. The tonic component fluctuates slowly (seconds to minutes) and represents the overall psycho-physiological state of a subject. Tonic events encompass slow shifts in the baseline skin conductance level and unpredictable changes in the skin conductance. The phasic component varies faster (fluctuating within seconds), and its changes are the short-time response evoked by an external stimulus. The typical shape of the phasic component, i.e., of the SCR, exhibits a fairly quick increase from the conductance level, succeeded by a slower exponential decay back to the baseline.

When the temporal gap between two consecutive stimuli, also known as the interstimulus interval (ISI), is shorter than the recovery period of the first response, the skin conductance responses generated by the two stimuli will overlap. This situation arises in many experimental protocols, especially in cognitive neuroscience. In this field, the typical ISI values (1-2 seconds) are usually shorter than the recommended ISI of 10-20 seconds necessary to avoid this overlap [41, 69]. This overlapping issue is one of the main limitations for the mathematical algorithms applied to decompose skin conductance into its phasic and tonic component.

In the past, numerous mathematical approaches have been developed to separate the phasic signal into individual Skin Conductance Responses connected to each stimulus, even for experiments with short inter-stimulus intervals. These methods also model how the activity of the Autonomic Nervous System – particularly the activity of the sudomotor nerve – leads to Skin Conductance Responses. This process enables the estimation of Autonomic Nervous System activity with potentially higher resolution, compared to using the raw Skin Conductance Response signal. However, many of these early methods, primarily conceived to tackle the overlapping issue, required visual revision and would introduce subjective elements in the analysis [22, 165].

Further improvement in this analysis was achieved by depicting the peripheral system as a linear time-invariant system. From this assumption, derived several classes of models [16]. Alexander et al. [6] introduced the first linear time-invariant model, proposing a decomposition methods by means of deconvolution. Their model allows the estimation of sudomotor nerve activity by conceiving Skin Conductance as the result of the convolution between discrete sudomotor nerve activity bursts and a biexponential impulse response function.

The deconvolution approach proposed in [6] relies on the assumption that the impulse response function is known apriori and is stable, i.e., time-invariant. However, multiple studies have demonstrated that the Skin Conductance Response shape shows significant variability both inter-individual and intra-individual [40, 81, 135]. To address this aspect, Benedek and Kaernbach [29, 30] proposed two different approaches in which the linear time-invariant assumption was adjusted to accomodate for the variability of the Skin Conductance Response shape. These models are the nonnegative deconvolution [30] and the continuous deconvolution analysis [29]. The two models split the sudomotor nerve activity into two portions, one that represents variations in the electrodermal activity derived from different origins, the other that represents the phasic activity. Both the models use a biexponential impulse response function (called Bateman function) and assume a pharmacokinetic model for the dynamic law of sweat diffusion. Even though these models [6, 29, 30] do not formally model observation noise, they all assume its presence, estimating a noisy sudomotor nerve activity and then deriving a filtered phasic component via a low-pass filter, with a subsequent heuristic and predifined peak-detection process.

More recently, Bach [15] presented the SCRalyze toolbox, that includes several models based on a linear time-invariant system. These models, as the one presented in [6], use a heuristic impulse response function optimized on large datasets. The algorithms in the toolbox try to estimate either the model input (the sudomotor nerve activity) or the parameters that best explain the observed Skin Conductance data, based on optimization methods. Additionally, they include a noise term to account for possible deviations from the time-invariance assumption.

Drawing from several elements of these prior methodological approaches, such as the impulse response function, Greco et al. [110] proposed a method for decomposing Skin Conductance signals into smooth tonic and sparse phasic components by solving a convex optimization problem. This solution incorporates physiological knowledge about electrodermal activity through an appropriate selection of constraints and regularizers. Specifically, the nonnegativity of the sudomotor nerve activity is enforced by the model using a nonnegative constraint on the corresponding optimization variable, as opposed to the soft penalty used in [30].

The cvxEDA algorithm

More recently, Greco et al. [111] have introduced the cvxEDA algorithm for the decomposition of the Skin Conductance signal into the phasic and tonic components. This algorithm estimates Autonomic Nervous System activity from Electrodermal Activity through a convex optimization approach.

The model represents the observed skin conductance as the sum of three elements:

1. a tonic (baseline) component;

2. a phasic (event-related) component, i.e., the output of the convolution of an Impulse Response Function with a sparse non-negative sudomotor nerve activity phasic driver;

3. an additive noise term.

The phasic component, related to event-driven responses, is modeled as a sum of decaying exponential functions that are triggered by impulse signals (i.e., the first derivative of the phasic component is a sparse positive signal). The advantage of this formulation is that it leads to a convex optimization problem that can be efficiently solved. The Impulse Response Function, associated with the phasic component, is modeled as an infinite impulse response functions through an autoregressive moving average model.

The tonic component, or the slower baseline of the EDA signal, is modeled as a smooth signal that captures the non-specific variations of the EDA, such as diurnal changes and slow trends in skin conductance levels.

The cvxEDA algorithm solves for these components using a technique called convex optimization. Specifically, it formulates the following convex optimization problem

$$\begin{aligned} &minimize\ \frac{1}{2}||Mq + Bl + Cd - y||_2^2 + \alpha||Aq||_1 + \frac{\gamma}{2}||l||_2^2 \\ &subj.to\ Aq \geq 0 \end{aligned} \tag{2.6}$$

that minimizes the sum of the squares of the residuals between the modeled signal and the observed EDA data while enforcing constraints that ensure the model's physiological plausibility. These constraints include that the first derivative of the phasic component is a sparse positive signal and the tonic component is smooth.

In this work, the cvxEDA algorithm is applied to Skin Conductance time series and the following electrodermal activity features are extracted:

The power spectrum. The dynamics of the normalized sympathetic component of the Electrodermal Activity, abbreviated as *edaSymp* in this work. This feature is meant to assess the dynamics of the autonomic nervous system in a noninvasive and quantitative way, by computing the power spectral density of heart rate variability. The power spectrum is computed within the 0.045-0.25Hz interval, as it has been demonstrated that at this frequency this feature reflects the sympatetic activity [232].

The skin conductance response. Describing the phasic changes in electrical conductivity of the skin, thus capturing the quick stimulus-evoked changes in the EDA signal [69]. In this work the following features and abbreviations are used:

- *max_pks*: maximum amplitude of the phasic component;

- *no_pks*: number of the phasic peaks;
- *sum_pks*: sum of the amplitudes of the phasic peaks;
- *mean_ph*: mean value of the phasic component;
- *std_ph*: standard deviation of the phasic component.

The skin conductance level. Describing the tonic level of the electrical activity of the skin, thus capturing the EDA slowly varying baseline and the subjects general psychophysiological state [112].

- *mean_ton*: mean value of the tonic component;
- *std_ton*: standard deviation value of the tonic component;
- *max_ton*: maximum value of the tonic component.

Chapter 3

Approach and Methods

This chapter takes a comprehensive journey into the algorithms, models, and metrics that have been employed to untangle the complex relationship between textual constructs and the vast array of human emotions. The systematic approach employed in this work seeks to provide an in-depth analysis and precise modeling of the diverse datasets, each playing a crucial role in exploring and defining the nexus between language and emotional responses.

The central goal of this thesis is to understand and illustrate the nuanced phenomena that link language and its features to the wide spectrum of emotive responses. The aim is not only to understand this interrelation but also to communicate it clearly and concisely. To this end, all the chosen techniques for data analysis are grounded in transparency and rigor, producing results that are straightforward and easily explicable.

Section 3.1 describes different methods and metrics for statistical analysis, including Correlation Coefficients (Section 3.1.1), the concept of Agreement among annotators of a manually labeled dataset (Section 3.1.2), and the Wilcoxon Rank-Sum Test (Section 3.1.3). Section 3.2 describes the Support Vector algorithm, that in this work has been used both for classification and regression purposes, along with the Recursive Feature Elimination algorithm (Section 3.2.3) applied in the cases were a feature selection method was needed. Finally, Section 3.3 introduces the approach for the implicit representation of the linguistic features. Despite being the sole approach applied in this work that does not yield explicitly interpretable results, this method is a state-of-the-art model for numerous Natural Language Processing tasks. Therefore, it was chosen for comparison with explicit linguistic features representations (see Section 2.2.2). This juxtaposition was instrumental in determining which of these two distinct methodologies for feature representations better encapsulates the relationship between language and emotional phenomena.

3.1 Methods and metrics for statistical analysis

3.1.1 Correlation coefficients

A correlation coefficient (ρ) is a statistical metric that measures the strength of the linear relationship between two variables x and y, namely, the degree to which the movement of the two variables is associated.

Typical correlation coefficient formulas return a value that ranges from -1 to $+1$. The absolute value of the correlation coefficient denotes the strength of the relationship between the two variables. A value greater than 0 indicates a positive correlation: for every positive increase in one variable, there is a positive increase in the other variable. A value lower than 0 indicates a negative correlation: for every positive increase in one variable, there is a negative decrease in the other. A correlation of 0 means the two variables are unrelated.

Spearman rank correlation coefficient

Spearman Rank Correlation Coefficient (or Spearman's Rho, denoted with the Greek letter rho ρ) is a non-parametric test that measures the strength of the rank correlation between two variables. Spearman's correlation is the non-parametric version of another popular correlation metric, Pearson Product-Moment Correlation. Pearson's correlation is applied to data based on some assumptions: (i) the data are in an interval or ratio level, (ii) the data are linearly related, and (iii) the data are bivariantly distributed. When the data do not meet these assumptions, it is necessary to apply Spearman's Correlation Coefficient. However, while Pearson's metric is applied to measure the strength and direction of the linear correlation between variables, Spearman's correlation determines the strength and direction of the *monotonic* relationship between the variables. In a monotonic relationship (function), either one of the following happens:

1. the function is monotonically increasing, i.e., when the variable x increases, so do the variable y;

2. the function is monotonically decreasing, i.e., when the variable x increases, the variable y decreases.

The monotonic relationship is less restrictive when compared to the linear relationship used in Pearson's coefficient. Indeed, a linear relationship is always monotonic, but a monotonic relationship is not always linear.

For a sample size n, the n raw scores X_i, Y_i are converted to ranks $R(X_i), R(Y_i)$ ad the Spearman's Rank Correlation Coefficient is calculated as follows:

$$r_s = \rho R(X), R(Y) = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}, \tag{3.1}$$

where ρ indicates the Pearson correlation coefficient, but in this case applied to rank variables, $cov(R(X), R(Y))$ is the covariance of the rank variables, and $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviations of the two variables.

In the case in which all n ranks are distinct integers, Spearman's correlation can be calculated with the following:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.2}$$

where $d_i = R(X_i) - R(Y_i)$ is the difference between the two ranks at each observation (among all n observations).

3.1.2 Measurement of agreement

In machine learning the quality of the training data can significantly impact the performance of the models built. This is especially true when it comes to tasks that involve manual annotations. When a dataset is manually labeled, it is good practice to have multiple subjects performing the annotations and then evaluate how coherent they are. Having only one person labeling the dataset would result in a biased (subjective) annotation. To evaluate the quality of manually annotated dataset, it is possible to resort to metrics like inter-annotator agreement.

Inter-annotator agreement, also referred to as inter-rater reliability or inter-coder agreement, is a statistical measure used to determine the level of consistency or consensus among multiple annotators who independently label or categorize a set of data. High agreement indicates that the labels or annotations assigned to the dataset are consistent across the different annotators, suggesting that they are likely to be reliable and accurate. On the other hand, low agreement may suggest problems with he data, such as ambiguous examples or unclear labeling instructions, which could lead to unreliable and noisy data.

The analysis of inter-annotator agreement can also provide insights into the complexity and ambiguities of tasks. By analyzing the instances where annotators disagree, researchers can gain a better understanding of the task's challenges and potentially improve the instructions or training provided to the annotators. Moreover, the level of inter-annotator agreement can provide an upper bound on the expected performance of the machine learning models. If humans have a hard time agreeing on the correct label, it's likely to be a challenging tasks for a machine learning model as well.

There are different metrics to assess inter-annotator agreement, depending also on the number of annotators involved. Commonly used ones are Cohen's Kappa [61], Fleiss' Kappa [99], and Krippendorff's Alpha [150]. In this work, Fleiss' Kappa was applied.

Fleiss' Kappa

Fleiss' Kappa[99] is a statistical measure of inter-rater agreement. Named after biostatistician Joseph L. Fleiss, it determines the level of agreement between two or more raters (judges, observers, annotators) when assigning *categorical ratings* to a number of items or classifying items. It is a generalization of Scott's pi (π) evaluation metric for two annotators[250] extended to multiple annotators. Whereas Scott's pi works for only two annotators, Fleiss' Kappa works for any number of raters. In addition, in Fleiss' Kappa, it is not necessary for all raters to annotate all given items.

Fleiss' Kappa can be calculated as:

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{3.3}$$

The factor $1 - \bar{P}_e$ gives the degree of agreement that is attainable by chance, while $\bar{P} - \bar{P}_e$ gives the degree of the agreement actually achieved above chance. If the annotators agree entirely, then $k = 1$. If there's no agreement between the annotators, other than the one attainable by chance, then $k \leq 0$. In practice, the interpretation of Fleiss' Kappa results often follows the guidelines suggested by Landis and Koch [156]:

- Below 0.00: Poor agreement
- 0.00 - 0.20: Slight agreement
- 0.21 - 0.40: Fair agreement
- 0.41 - 0.60: Moderate agreement
- 0.61 - 0.80: Substantial agreement
- 0.81 - 1.00: Almost perfect agreement.

3.1.3 Wilcoxon Rank-Sum Test

Wilcoxon Rank-Sum Test (also known by the names *Mann-Whitney U test*, *Mann-Whitney-Wilcoxon*, or *Wilcoxon-Mann-Whitney test*) is a non-parametric statistical test of the null hypothesis. The test was proposed in 1945 by Frank Wilcoxon [299] and by Henry Mann and Donald Whitney in 1947 [183], but actually dates back to 1914, first introduced by Gustav Deuchler [151, 209].

The Wilcoxon Rank-Sum Test determines whether two independent samples come from the same population. Being a non-parametric test, this test makes no assumption about the underlying distribution of the data and is based on ranks rather than the original observations. For this reason, the test is well-suited for data that is not normally distributed. However, the test makes other assumptions about the data and the experimental design:

- *Independence.* The data in each group must be independent of each other, meaning that the values in one group should not be related to the values in the other group [64].

- *Random Sampling.* The data should be collected through random sampling, which means that every member of the population has an equal chance of being selected for inclusion in the sample [253].

- *Ordinal Data.* The data should be ordinal, meaning that it can be ranked. The Wilcoxon Rank-Sum Test is designed to compare the medians of two independent samples, and ranking the data allows for this comparison [308].

- *Homogeneity of Variance.* The variances of the two groups should be approximately equal. If the variances are significantly different, the Wilcoxon rank-sum test may not be appropriate [64].

- *Equal Shape of Distributions.* The two groups should have similar shapes, meaning that the distribution of values in one group should not be markedly different from the distribution of values in the other group [253].

It is important to note that the Wilcoxon rank-sum test is robust to violations of the homogeneity of variance assumption and can be used even when the variances are not equal. However, violating the independence, random sampling, or ordinal data assumptions can lead to inaccurate results and should be avoided. If these assumptions are not met, alternative tests, such as the Kruskal-Wallis test [191] or the Jonckheere-Terpstra test [7], may need to be considered.

The Wilcoxon Rank-Sum test compares the medians of two independent samples. It works by converting the original numerical data into ranks and then using these ranks to calculate a test statistic, U, which is used to test the null hypothesis that the two groups have equal medians.

The Wilcoxon rank-sum test involves the following steps:

1. Combine the two independent samples into a single dataset and rank the values from lowest to highest, ignoring ties.

2. Calculate the sum of the ranks for the observations in one of the groups. This is denoted as R_1.

3. Calculate the test statistic U using the following formula:

$$U = n_1 * n_2 + (n_1 * (n_1 + 1))/2 - R_1 \tag{3.4}$$

   where n_1 is the sample size of the first group, n_2 is the sample size of the second group, and R_1 is the sum of the ranks for the observations in the first group.

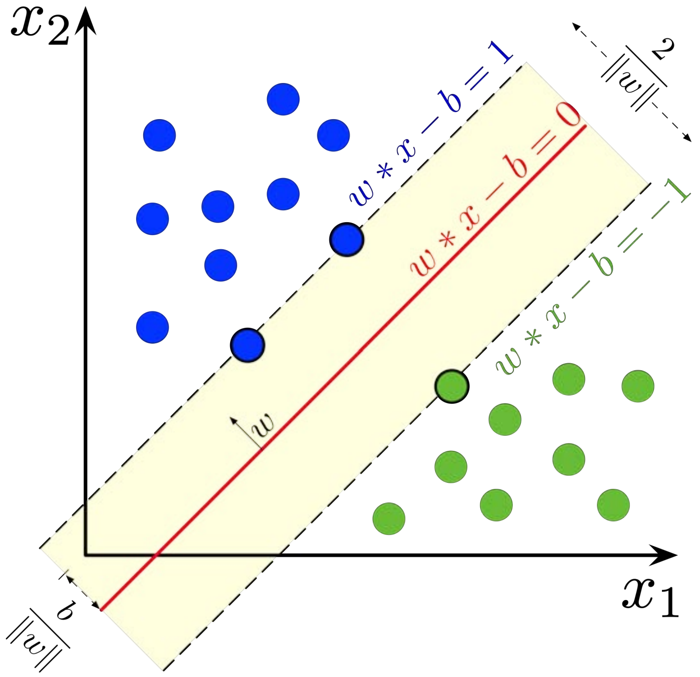


Figure 3.1: Maximum-margin hyperplane and margins for a Support Vector Machine trained with samples from two classes. *[Image source en.wikipedia.org]*

4. Compare the calculated value of U to a critical value from a standard normal distribution or look up the corresponding p-value from a Wilcoxon rank-sum test table.

If the calculated value of U is greater than the critical value or the p-value is less than the chosen level of significance, reject the null hypothesis and conclude that the two groups have different medians. Otherwise, fail to reject the null hypothesis and conclude that there is no evidence to suggest that the two groups have different medians.

3.2 Support Vector algorithm

The foundations of the Support Vector algorithm are set in the framework of statistical learning theory or Vapnik–Chervonenkis theory. The Support Vector algorithm is indeed a generalization of the *Generalization Portrait* algorithm first theorized in Russia during the sixties [278, 281] and then further developed in the following decades by Vapnik and Chervonenkis [279, 280, 282]. Vapnik-Chervonekis theory is a theory for non-parametric (distribution-free) dependency estimation with finite high-dimensional data. This theory is based on the empirical risk minimization principle, an approach used in neural network training that tries to minimize the average training error (empirical risk) with respect to model parameters (weights of the neural network) in order to estimate the unknown dependency. The Vapnik-Chervonekis theory derives necessary and sufficient distribution-free conditions for fast convergence and consistency of the empirical risk minimization principle.

Support Vector Machines (SVMs) implement a learning algorithm that leverages the structural risk minimization principle to recognize patterns in complex datasets and generalize effectively on unseen data. The algorithm attempts simultaneously to minimize the empirical risk and the Vapnik-Chervonenkis dimension.

In the field of Machine Learning, the typical application of Support Vector Machines is in data classification (see Section 3.2.1). By looking at previously labeled data, the SVM is capable of leveraging this information to separate and label unseen data. SVMs can also be applied to regression problems by introducing an additional loss function (see Section 3.2.2).

3.2.1 Support Vector Classification

There are four main concepts at the base of the SVM algorithm for classification:

1. the separating hyperplane;
2. the maximum-margin hyperplane;
3. the soft margin;
4. the kernel function.

The separating hyperplane is a decision boundary that serves as the decision boundary between classes. This is optimally chosen as the maximum-margin hyperplane to ensure maximum distance from the nearest datapoints, while a soft margin approach allows some classification errors for better generalization. Kernel functions enable these linear classifiers to operate effectively in high-dimensional and non-linear spaces. In what follows, these main concepts will be analyzed more in depth.

Consider a training dataset of n points such that

$$D = \Big\{(x_1, y_1), ..., (x_n, y_n)\Big\}, \qquad x \in \mathbb{R}^n, \;\; y \in \{-1, 1\}\,, \tag{3.5}$$

where the x represents the data points of the dataset and the y represents the two classes assigned to the data points (in this example, the two possible classes are -1 and 1). In the Support Vector Classification, the training points belonging to D need to be optimally separated by a hyperplane,

$$\langle w, x \rangle + b = 0. \tag{3.6}$$

The points are optimally separated by the hyperplane if these conditions verify: (i) the points need to be separated without error, i.e., all points belonging to one class should be separated from the ones belonging to the other class and there should not be misclassified points, (ii) the distance between the closest point to the hyperplane should be the maximal possible. Figure 3.1 shows a graphical representation of the maximum-margin hyperplane and margins for an SVM trained to separate two classes.

When data is linearly separable, it is possible to select two parallel hyperplanes that separate the two classes -1 and 1, so that the distance between the two hyperplanes is as large as possible. The region between them is called *margin*, while the hyperplane that lies between them is the *maximum-margin hyperplane*, i.e., the hyperplane that is at the maximum distance possible from the data points. The optimal hyperplane should satisfy this constraint:

$$y_i \left[\langle w, x_i \rangle + b\right] \geq 1, \quad i = 1, ..., n. \tag{3.7}$$

The constraint simply states that each data point must lie on the correct side of the margin and they should not fall into the margin itself.

The distance $d(w, b; x)$ of a point x from the hyperplane (w, b) is

$$d(w, b; x) = \frac{|\langle w, x_i \rangle + b|}{||w||}. \tag{3.8}$$

The optimal hyperplane is obtained by maximizing the margin p, respecting the constraint of 3.7. The margin is given by

$$p(w, b) = \underset{x_i : y_i = -1}{min} d(w, b; x_i) + \underset{x_i : y_i = 1}{min} d(w, b; x_i) = \frac{2}{||w||}. \tag{3.9}$$

Given that the maximum-margin hyperplane is dependent on the points x_i that lie next to it, these points are called *support vectors.*

So far, the SVM Classification problem was examined assuming that data are only linearly separable. However, in many cases, data cannot be separated linearly in a clean way. To handle these cases, the SVM algorithm has been modified by adding a *soft margin* (also called *degree of tolerance*). In this way, some data points can push their way through the margin (i.e., they are allowed to appear on the *wrong side* of the margin) without affecting the final result. The number of points that are allowed to violate the hyperplane is determined by a user-specified parameter, in order to limit the amount of misclassified points present.

The soft margin is indeed helpful for separating non-linear data, but in some settings, it may not be enough. When this case occurs, the SVM is helped by *kernel* functions. Kernel functions can be used to solve complex non-linear classification problems without having to resort to complex calculations. Simply, kernels can be used to construct a mapping of the data into a higher dimensional space. Different kinds of kernel functions can be applied. In this work, the following have been used:

1. **Linear Kernel**, the basis kernel used in all cases in which there are linearly-separable data, with equation $K(x_i, x_j) = x_i \cdot x_j + c$;

2. **Gaussian Radial Basis Function**, widely used when there is no prior knowledge about the distribution of the data. It projects the data into a Gaussian distribution following this equation: $K(x_i, x_j) = exp(-\gamma \, ||x_i - x_j||)^2$, where $\gamma > 0$.

3.2.2 Support Vector Regression

Support Vector Machines can be also applied to solve regression problems and predict discrete values. Support Vector Regression (SVR) uses the same principle behind Support Vector Classification, but instead of finding a hyperplane to separate the data, it looks for the best hyperplane that contains (*fits*) the maximum number of points.

SVRs work by introducing an alternative loss function, accordingly modified to include a distance measure. Different loss functions are possible, such as (*i*) quadratic, (*ii*) Laplace, (*iii*) Huber, and (*iv*) ϵ-Insensitive. In this work, the latter is applied.

Usually, in Linear Regression models, the goal is to minimize the Sum of Squared Errors (SSE). However, with SVR models the objective is not to minimize the SSE but rather to minimize the coefficients – more specifically, the l2-norm of the coefficient vector. Therefore, the objective is to minimize

$$MIN \frac{1}{2} \, ||w||^2 \quad , \tag{3.10}$$

following this constraint

$$|y_i - w_i x_i| \leq \varepsilon \; . \tag{3.11}$$

Notice that the error (ε) is handled not in the minimizing function but in the constraint, where it is set to be less or equal to a specified margin. The value of ε (or maximum error) can be tuned to gain the desired accuracy in the SVR model. Ideally, all data points should lie between $w_i x_i - \varepsilon$ and $w_i x_i + \varepsilon$. There are indeed cases in which the data do not respect the constraint and some data points fall outside the established value of ε. In these cases, the problem can be solved by adding the so-called *slack variables*. Slack variables simply add a deviation ξ from the margin to account for values that fall outside of ε. Even though these deviations can exist, it is still necessary to minimize them. To do this, the deviations get added to the objective function

$$MIN\frac{1}{2}\left|\left|w\right|\right|^{2}+C\sum_{i=1}^{n}\left|\xi_{i}\right| \tag{3.12}$$

and to the constraints

$$\left|y_{i}-w_{i}x_{i}\right|\leq+\left|\xi_{i}\right| \ . \tag{3.13}$$

The additional hyperparameter C can be tuned to modify the tolerance of the algorithm. With a higher value of C, the tolerance for points outside of the error ε increases. If C approaches 0 the equation collapses into the simplified one (3.10). In most cases, it is good practice to test different values of C to find the ones that maximize the performance of the algorithm.

3.2.3 Recursive Feature Elimination

Recursive Feature Elimination is a feature selection algorithm used in Machine Learning to identify the most important features in a dataset and reduce the dimensionality of the feature space. In the machine learning and pattern recognition literature, the algorithm is cited as first appearing in the work of Guyon et al. [117], who used it to select a subset of genes from broad patterns of gene expression data. In their work, they demonstrated that the genes selected by this technique yield better classification performance and are biologically relevant to cancer. Indeed, the method eliminates gene redundancy automatically and provides better and more compact gene subsets.

Typically, Recursive Feature Elimination is utilized in conjunction with other algorithms to identify a subset of features that are most relevant to the outcome of a predictive model (e.g., a Support Vector Machine). The algorithm begins by training the predictive model on the entire set of features of the dataset and determining the feature importance. The feature with the least impact on the model's performance is removed, and the process is repeated iteratively until a desired number of features is reached.

The core mathematical concept behind Recursive Feature Elimination is the ranking of features based on their importance to the model. Feature ranking can be accomplished using different methods, such as feature weights, coefficients, or feature importances, which are determined according to the characteristics of the predictive model implemented. Once the features are ranked, the feature selection process can be formalized as an optimization problem.

Given a set of features $F = \{f_1, f_2, ..., f_n\}$ and a feature ranking function $R(F)$, the goal of the Recursive Feature Elimination is to find a subset of features $S \subseteq F$, such that $R(S)$ is maximized. In this optimization problem, the feature ranking function $R(F)$ is calculated for each feature subset S. The feature subset that results in the highest ranking is selected as the final set of features.

Recursive Feature Elimination is especially advantageous for high-dimensional datasets, where the number of features exceeds the number of samples. Traditional methods such as correlation analysis or mutual information can be challenging to use in such cases to identify the most relevant features. One of the primary benefits of Recursive Feature Elimination is that it is an iterative method, which permits the gradual elimination of features, rather than removing all irrelevant features at once. This allows for a more thorough exploration of the feature space and can lead to more accurate predictions.

By reducing the dimensionality of the dataset and removing redundant features, recursive feature elimination can enhance the predictive accuracy of a model, by also mitigating overfitting. Moreover, having fewer features means less computational complexity, which leads to faster training time. Finally, recursive feature elimination can make the model easier to understand and interpret.

3.3 Implicit representation of linguistic features

The implicit features, i.e., the high dimensional vectors that represent text, employed in this work were created leveraging BERT (*Bidirectional Encoder from Transformers*) [73], a Neural Language Model based on the architecture of Transformers. BERT is publicly available as an open-source framework and is nowadays one of the state-of-the-art models for most Natural Language Processing applications. This model and features were selected to be counterposed to the explicit linguistic features (detailed in Section 2.2.2) selected to profile the linguistic aspects of the texts used in this work. The goal of this comparison was to establish which of the two methodologies can better describe the relationship between language and emotional phenomena. Hereafter, this section describes the concepts on which BERT is based and how the model is used in this work.

A *Language Model* is a probabilistic model that can predict the probability of a sequence of words, i.e., the model can predict the next word in a sequence given the words that precede it. A probabilistic language model would define the probability of a sentence $s = [w_1, w_2, ..., w_n]$ as:

$$P(s) = \prod_{i=1}^{n} P(w_i | w_1, w_2, ..., w_{i-1}) \tag{3.14}$$

Language Models are primarily of two types:

- Statistical Language Models: these models use traditional statistical techniques like n-grams, Hidden Markov Models, and certain linguistic rules to learn the probability distribution of words

- Neural Language Models: also called *continuous space language models*, these are new players in the Natural Language Processing field and have surpassed the statistical language models in their effectiveness. They use different kinds of Neural Networks to model language creating a continuous representation or embedding words to make their prediction [199].

3.3.1 Neural Language Models

Traditional models such as n-grams predict the next word in an n-gram sequence by following the Markov assumption that the probability of a given word is solely dependent on the previous $n-1$ words. However, there are inherent limitations to language models that rely on n-grams. Initially, while a range of smoothing methods has been suggested to mitigate data sparsity issues, n-gram language models persistently demonstrate suboptimal performance when confronted with rare and unseen words. Furthermore, when language models are trained on progressively larger texts, the count of distinct words and potential sequences in the model vocabulary escalates exponentially with the number and length of texts. An extensive vocabulary results in highly sparse data because the number of possible word sequences augments exponentially.

An approach to assess the curse of dimensionality in modeling natural language has been given by the approach developed by Bengio et al. [31]. The authors propose a Neural Probabilistic Language Model that assigns a unique vector to each word and uses a neural network structure to predict the subsequent word, as illustrated in Figure 3.2. The Figure shows the architecture of the model, defined as:

$$f(i, w_{t-1}, ..., w_{t-n+1}) = g(i, C(w_{t-1})), ..., C(w_{t-n+1})) \tag{3.15}$$

where g is the neural network and $C(i)$ is the feature vector of the $i-th$ word. When trained on a specific text corpus, the model gains the ability to model the joint probability of sentences. Concurrently, it yields word embeddings, otherwise referred to as low-dimensional word vectors, as a part of its learned parameters. Unlike preceding methodologies, the word embeddings generated by the Neural Language Model have the advantage of decreasing the dimensionality of categorical variables and facilitating a more meaningful representation of categories within the transformed space.

This pioneering work on Neural Language Models has catalyzed a wave of methodologies centered on embedding words into distributed representations using a neural network. Noteworthy implementations of this include, for example, *Word2Vec* [200], and *GloVe* [221]. Despite their distinct characteristics, these models are bound by their exceptional efficiency and their widespread adoption in numerous Natural Language Processing tasks in recent years. In particular, during the past years, the *Word2Vec* algorithm has been the lead algorithm for many Natural Language Processing tasks. The algorithm is implemented as

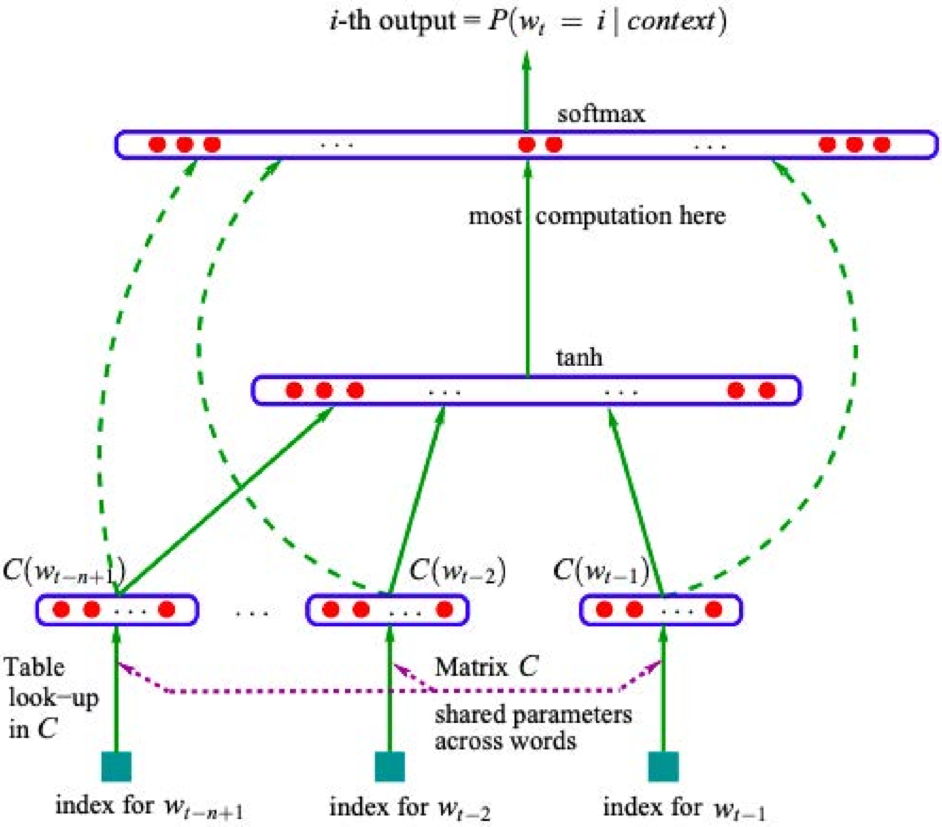


Figure 3.2: The architecture of the Neural Probabilistic Language Model. *[Image source Bengio et al. [31]].*

a software package[1] that include two primary models: the Continuous Bag of Words (CBOW) and the Skip-gram model, whose architectures are represented in Figure 3.3.

The CBOW model predicts the current word based on the context. It treats the context as a single observation, taking the average of all word vectors, without considering the order of words. For example, for the sentence "the cat sat on the couch", with "sat" as the target word, the CBOW model uses "the", "cat", "on", "the", "couch", to predict "sat". Formally, the model predicts the word w_i given a window of context:

$$P(w_i | w_{j(|j-i| \leq l, j \neq i)}) = softmax\left(M\left(\sum_{|j-i| \leq l, j \neq i} w_f\right)\right) \tag{3.16}$$

[1] https://code.google.com/archive/p/Word2Vec/

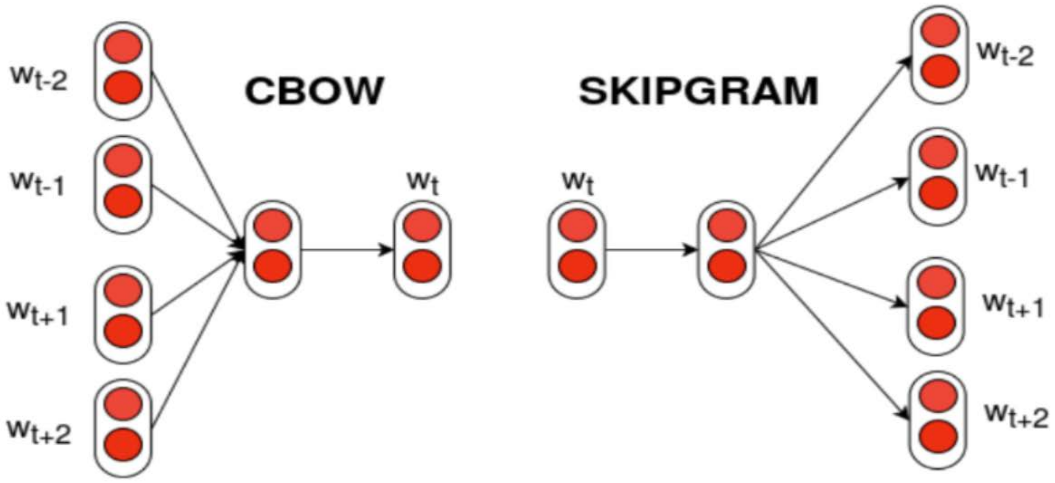


Figure 3.3: The architecture of the CBOW (left) and Skip-gram (right) models in the Word2Vec algorithm. *[Image source machinelearninginterview.com].*

where $P(w_i|w_{j(|j-i|\leq l,j\neq i)})$ is the probability of the word w_i given its context, l is the size of the training context, M is the weight matrix in $\mathbb{R}^{|V|\times m}$, V is the vocabulary, and m is the dimension of the word vector. Subsequently, the CBOW model is optimized by minimizing the sum of negative log probabilities as in the following loss function:

$$L = -\sum_i \log P(w_i|w_{j(|j-i|\leq l,j\neq i)}) \tag{3.17}$$

The Skip-gram model predicts the surrounding words given a current word. In contrast to the CBOW model, Skip-gram treats each context-target pair as a new observation and predicts context words ("the", "cat", "on", "the", "couch") from the target word "sat". Formally, the model predicts the context given the word w_i:

$$P(w_j|w_i) = softmax(M_{w_i})(|j-i| \leq l, j \neq i) \tag{3.18}$$

where $P(w_j|w_i)$ is the probability of the context word w_j given the word w_i, and M is the weight matrix. The model is optimized with the following loss function:

$$L = -\sum_i \sum_{(|j-i|\leq l,j\neq i)} P(w_j|w_i) \tag{3.19}$$

While models like Word2Vec and GloVe have been instrumental in the progression of natural language processing, they also have a few drawbacks.

As a first issue, these models cannot handle out-of-vocabulary words. If they encounter a word during inference that was not included in their training corpus, the models are unable to generate an appropriate embedding for it. Furthermore, given that they treat words as the smallest training unit, these models cannot account for morphological nuances, which becomes problematic for languages with a rich morphology where a single root word can have many different forms. At the same time, even though these models can capture the semantic meaning of individual words, they do not encode the sequence of words in a sentence, which is an essential aspect of linguistic comprehension.

Moreover, storage poses a challenge for these models, as each unique word requires a separate vector representation. For large vocabularies, this requirement may result in significant storage overhead. Similarly, the computational expense of training these models on extensive corpora can be considerable, posing a potential obstacle for resource-limited environments.

However, the primary challenge associated with these models is that they generate static word embeddings. In essence, each word is given the same vector representation regardless of its context. For instance, in the case of a word like "bank", these models would not differentiate between its usage in the phrase "river bank" as opposed to "bank account". Consequently, the distinct meanings depending on the contextual usage are not encapsulated by the vectors generated by these models.

To address this issue and keep the context-specific representation of a document, different context-based models have been introduced. In 2016, Melamud et al. [195] proposed Context2Vec, a model to generate context-dependent word representation. Context2Vec is a modified version of Word2Vec's CBOW model, but the most significant change is the replacement of the traditional average word representation within a fixed window with a more robust and advanced Bi-directional Long Short-Term Memory neural network [128]. The method uses a large text corpus to train a neural model that embeds in the same low-dimensional space both the context derived from a sentence and its target words. This space is then fine-tuned to accurately reflect the interrelations between the target words and their sentential context, as shown in Figure 3.4

From the foundations laid by Context2Vec, McCann et al. [189] introduced CoVe (Contextualized Word Representation Vectors). Rather than employing the methodologies used in Word2Vec (skip-gram or CBOW) or GloVe (Matrix factorization), they leveraged machine translation to construct CoVe. Their fundamental strategy involved pre-training a two-layer BiLSTM for attention sequence-to-sequence translation, initializing with GloVe word vectors. Subsequently, the output of the sequence encoder is combined with GloVe vectors and used in a task-specific downstream model through transfer learning.

In 2018, Peters et al. [222] introduced ELMo, a deep bi-directional LSTM architecture capable of representing each word in relation to the full context in which it appears. Specifically, ELMo converts words into low-dimensional vectors by inputting the word and its surrounding text into a two-layer bi-directional language model (BiLM). For a sequence of n words $(w_1, w_2, ..., w_n)$ ELMo calculates

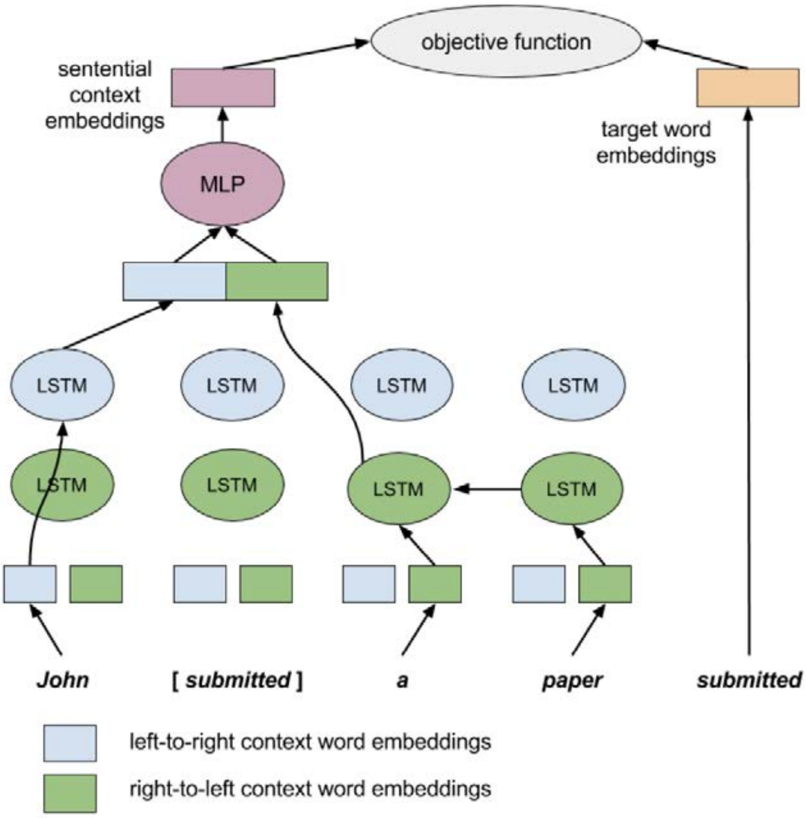


Figure 3.4: The architecture of Context2Vec. *[Image source Melamud et al. [195]].*

a forward language model (as in equation 2.2) and a backward language model and computes the final vector as a concatenation of hidden representations from the two models. The backward model is similar to the forward one, with the sole difference that the input word sequence is reversed $(w_n, w_(n-1), ..., w_1)$ and each word is predicted according to the future context:

$$P(s) = \prod_{i=1}^{n} P(wi|w_{i+1}, w_{i+2}, ..., w_n) \tag{3.20}$$

3.3.2 The transformer architecture

The BERT framework is based on Transformers architecture [286], a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection – a process called *attention*. These models are designed to handle sequential data and compensate for one of the disadvantages of recurrent model approaches, namely the need to process data in an orderly manner. In the Transformer architecture, the recurrent operations are substituted with the attention mechanism to learn contextual relationships between words in a text.

In its base form, the Transformer architecture includes two separate mechanisms, an encoder and a decoder, both containing the same stacked sequence of layers that transform the input embeddings into the output for the predictions for the task. The encoder first takes an input sequence $(x_1, ..., x_n)$ from a text and maps it to a sequence of embeddings $z = (z_1, ..., z_n)$. The embeddings in z are used by the decoder to autoregressively produce the output sequence $(y_1, ..., y_m)$. Each of the layers in the encoder and the decoder has the same architecture, comprising two sublayers: a multi-head self-attention mechanism and a fully connected feed-forward network, with residual connections around them and followed by a normalization layer. Additionally, the decoder has a third layer that receives the output from the previous stack and modifies the self-attention sublayer to look only at the preceding words. Indeed, while the encoder attends all words in the input sequence, regardless of their position, the decoder needs to prevent attending at future context. For the nature of the language model task itself, for the decoder to predict a word at a position i, the prediction should be dependent only on the outputs of the words that come before it in the sequence. The architecture of the Transformer model and its encoder-decoder mechanism is shown in Figure 3.5.

The attention mechanism of the transformer allows the model to focus on different words in the input sequence when generating each word in the output sequence. It allows the model to deal in a flexible manner with the variable length of sentences: thanks to the attention mechanism, the model can draw connections between any parts of the sequence, and long-range dependencies have the same likelihood to be taken into account as short-range ones.

The attention mechanism operates by assigning a weight to each word in a sequence, which signifies the degree of ‘attention’ each word should receive. In essence, the attention function can be seen as a mapping between a query and a set of key-value pairs to an output, where the query, the keys, the values, and the output are all vectors.

The transformer architecture implements a scaled dot-product attention. In this procedure, for each word in the input three vectors are created: the query vector $\mathbf{q}$, the key vector $\mathbf{k}$, and the value vector $\mathbf{v}$. Subsequently, the key vector $\mathbf{k}$ is matched with the query to provide the weighting score, i.e., the score is obtained from the dot product between these two vectors. The intuition is that if $\mathbf{k}$ and $\mathbf{q}$ are similar, the dot-product will be large and the model will pay *more attention* to that word. The scores are then normalized, generally with a *softmax* function, to be interpreted as probabilities. This ensures that all the attention scores for an output word sum up to 1. To obtain the final output for each word, the value vectors are weighted by the normalized attention scores and summed.

The computations performed by the scaled dot-product attention can be applied to entire sets of queries simultaneously. To do so, sets of queries, keys, and

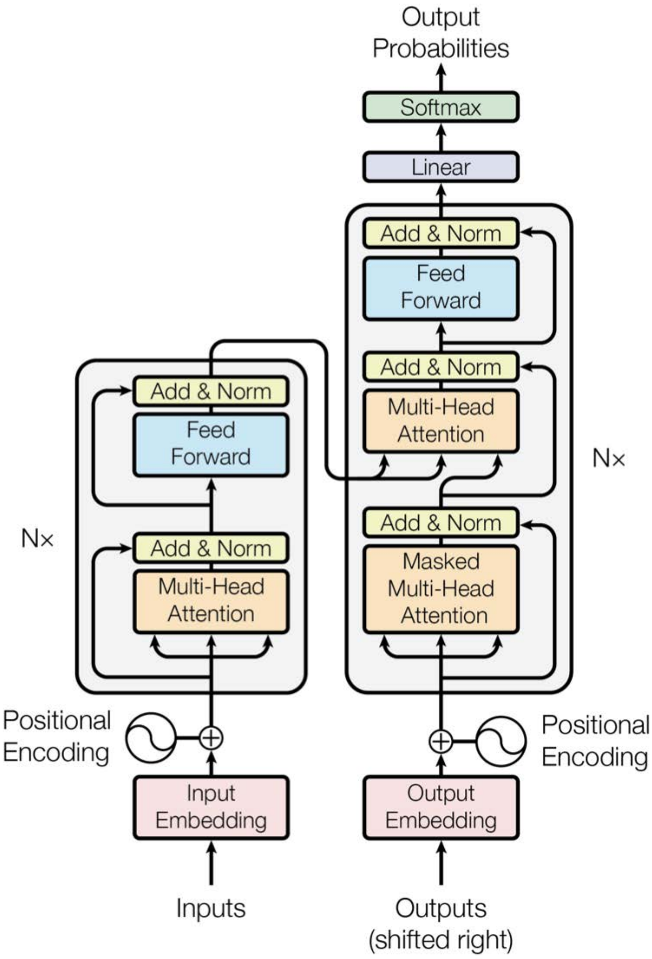


Figure 3.5: The architecture of the Transformer model *[Image source Vaswani et al. [286]]*.

values, are respectively packed in the matrices Q, K, and V, which are supplied as inputs to the attention function as shown in Figure 3.6 (left). Formally, given the query matrix Q, the key matrix K, and the value matrix V as inputs, the output of the attention is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.21}$$

where d_k is the dimension of the query matrix.

In the Transformer model, instead of performing the attention process once per word, it can be performed multiple times in parallel, with each instance referred to as “head” – hence, the term multi-head attention. The multi-head attention mechanisms linearly projects the queries, keys, and values h times, using

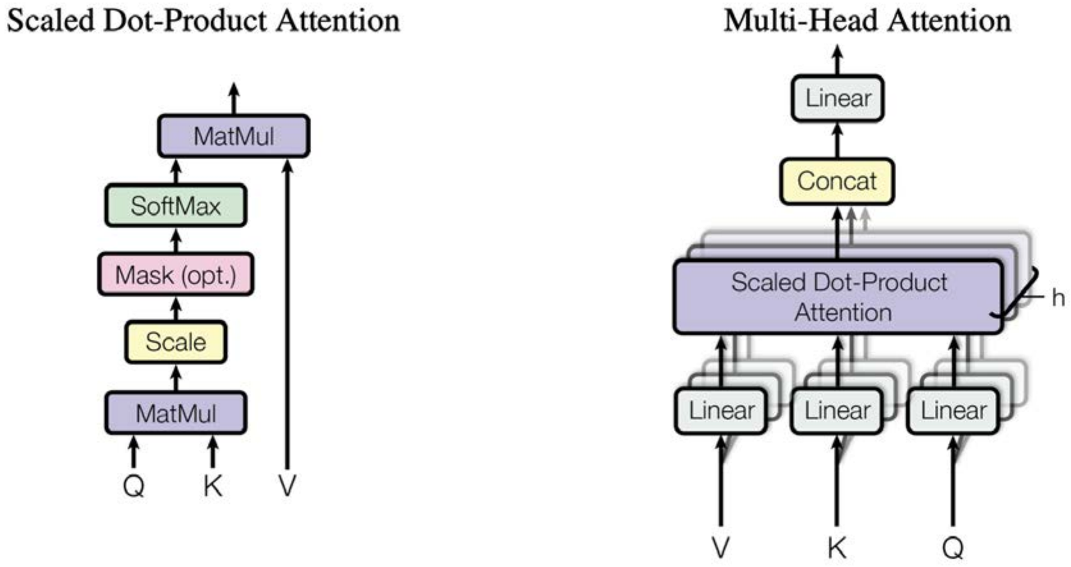


Figure 3.6: The scaled dot-product attention (left) and the multi-head attention (right) mechanisms *[Image source Vaswani et al. [286]]*.

a different learned projection each time. The single attention mechanism is then applied to the head of the h projections in parallel to produce h outputs, which, in turn, are concatenated and projected again to produce a final result. This process is also shown in Figure 3.6 (right). Each head has its own set of learned parameters and can therefore learn to focus on different features in the input data: the attention function can extract information from different representation subspaces, which would be otherwise impossible with a single attention head. The multi-head attention function can be formalized as:

$$multiead(Q, K, V) = [head_1, head_2, ..., head_h]W^O \tag{3.22}$$

where each $head_i$, with $i = 1, ..., h$, implements a single attention function characterized by its own learned projection matrices such that

$$head_i = attention(QW_i^Q, KW_i^K, VW_i^V)W^O \tag{3.23}$$

where W_i^Q, W_i^K, and W_i^V are the projection matrices used to generate different subspace representations of the query, key, and value matrices, and W^O is a projection matrix for the multi-head output.

The BERT model

One of the most famous Language Models based on the Transformer architecture is BERT (*Bidirectional Encoder from Transformers*) [73]. BERT is publicly available as an open-source framework and is nowadays one of the state-of-the-art models for many Natural Language Processing applications.

BERT has a series of advantages that made it the top model in recent years. The first advantage is its bidirectional nature. Being a bidirectional model means that BERT learns information from both the left and the right side of a token's context during the training phase. This is a significant improvement over previous models like GPT [235] and the original Transformer model, which are unidirectional and can only understand the context of a word based on words that came before it (left-to-right) or after it (right-to-left).

The second advantage of BERT is that it is pre-trained on a large amount of data. Currently, many pre-trained BERT models are available for different languages and purposes. However, the original pre-trained model came in two sizes: BERT-base, trained on a corpus of around 800 million words, and BERT-large, trained on English Wikipedia, a corpus of around 2,500 words. The pre-trained BERT models can then be fine-tuned on other domains and tasks. This way, it is possible to leverage the knowledge already included in the pre-trained model and transfer it to new data. This possibility eliminates the need to collect a large amount of data to create a language model or perform a specific Natural Language Processing task.

Given that BERT's goal is to generate a language representation model, it only needs the encoding part of the Transformer architecture. BERT uses the encoder to collect the semantic and syntactic information into the embedding representation. The input for BERT's encoder is a sequence of tokens that are first converted into vectors (embeddings) and then processed by the neural network. When processing the sequence of tokens, BERT adds to the sequence some special tokens that contain extra metadata:

- the [CLS] token, a special classification token that is added at the beginning of each sentence. The final hidden state that corresponds to this token is used as an aggregate representation of the whole sequence for classification tasks.

- the [SEP] token inserted between two different sentences to help the encoder distinguish them.

Overall, the final embedding representation constructed by BERT is the sum of three components:

1. the token embedding, learned for each specific token from the WordPiece embeddings vocabulary [304];

2. the segment embedding: given that BERT can take as input a single sentence or two sentences A and B stacked together, this learned embedding indicates for each token whether it belongs to sentence A or sentence B;

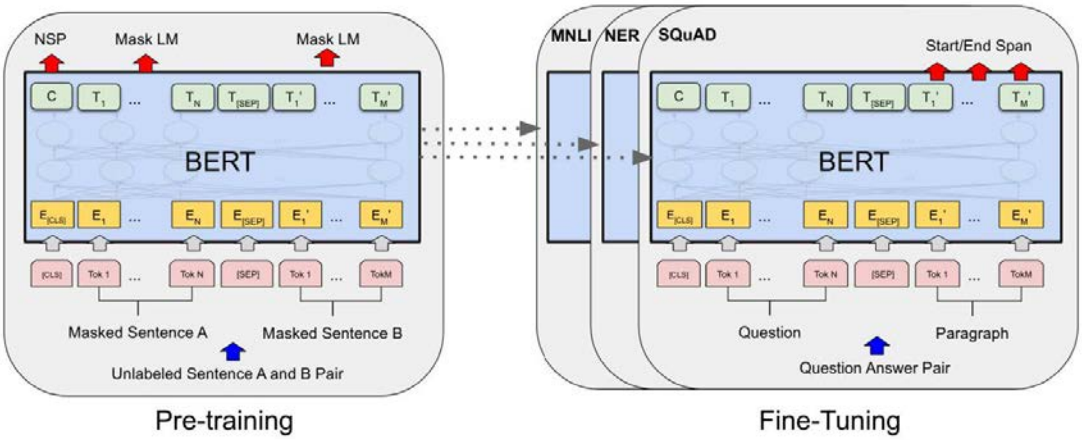


Figure 3.7: Pre-training and fine-tuning architecture for BERT. *[Image source Devlin et al. [73]].*

3. the position embedding, a learned embedding that encodes the position of the tokens within the sentence.

The BERT model is pre-trained with two unsupervised tasks. The first is the Masked Language Model, in which some tokens in the sentences are randomly masked, and BERT is instructed to predict the masked tokens. The second is the Next Sentence Prediction, in which BERT is given in input two sentences A and B, and BERT is instructed to predict sentence B given the preceding sentence A. After the pre-training stage, the model can be then fine-tuned. In the fine-tuning stage, the model is first initialized with the pre-trained parameters, which are then fine-tuned using labeled data from downstream tasks. A representation of the overall structure of the pre-training and fine-tuning procedures for the BERT model can be seen in Figure 3.7.

In Chapter 5 BERT was used to produce implicit feature vectors to be used as input for a classification task. In particular, the BERT-base model (*bert-base-uncased*) will be leveraged.

Part II
Case Studies

Chapter 4

Influence of Language Complexity on User Engagement

"Much of a language's complexity is not necessarily for effective communication."

Guy Deutscher

This first case study focuses on the aspect of engagement in communication and the influence the explicit linguistic profile has on language complexity and thus on engagement. Engagement is shaped by a range of emotions and factors, and in human-human interaction, it stems from effective communication. It's crucial to comprehend that the effectiveness of communication isn't merely determined by the sender, but also significantly influenced by how the receiver processes the information. The success of communication depends on multiple elements, including the simplicity of the message conveyed. In other words, the complexity of language can have an impact on communication and the emotions associated with engagement. This chapter provides insights into the relationship between language complexity, communication, and engagement.

This case study examines how humans perceive syntactic complexity in language by exploring two scenarios: the complexity of sentences in isolation and the complexity of sentences within a specific context. In the first scenario, complexity is studied by utilizing explicit linguistic features (Section 2.2.2), while in the second scenario, both explicit and implicit linguistic features (Section 3.3) are considered. Overall, the results indicate that explicit features are the most effective in predicting the level of complexity of a sentence.

The chapter is organized as follows: Section 4.1 describes the motivations behind this study; Section 4.2 describes the collection and creation of the data for this study; Section 4.3 analyzes the complexity of sentences presented in isolation (without providing any context); Section 4.4 analyzes the complexity of sentences when presented with two additional contextual sentences.

4.1 Background and motivations

The complexity of natural language is a topic of significant interest in linguistic research. One of the main reasons for this interest resides in the fact that complexity impacts efficient communication.

According to linguistic theories, communication is considered efficient when a message is transmitted between the speaker and the listener quickly, and with minimal elaboration effort [127]. In simpler terms, communication efficiency relies on the structural and grammatical simplicity of a message, allowing for rapid processing with minimal cognitive effort. Thus, a *complex message* in natural language can hinder effective and satisfying communication, affecting engagement.

The concept of engagement is multi-faceted and has been defined in various ways. Some theories consider engagement as something that captures and retains someone's attention [53], or as something related to playfulness and sensory integration [159]. Others view engagement as a dimension of usability, influenced by a user's initial impression of an application and their enjoyment while using it [234].

Regardless of the specific definition, engagement is widely acknowledged as a complex process that encompasses various attributes, including attention, aesthetics, interest, challenge, control, motivation, novelty, and feedback. These attributes play a role in each stage of the engagement process, including point of engagement, engagement, disengagement, and re-engagement [214].

Engagement is crucial in both human-human and human-computer interactions. A dull conversation with another person is likely to result in a lack of engagement, causing the listener's mind to wander. Similarly, a user-unfriendly machine makes it difficult for users to enjoy and find the interaction pleasant. The same holds true when a human tries to listen to or read a message that they perceive as linguistically complex. If decoding and interpreting the message requires a cognitive effort that the receiver perceives as excessive, it can lead to feelings of frustration and disinterest, jeopardizing the success of the communication and resulting in its termination due to a lack of engagement.

By understanding the linguistic aspects that humans perceive as complex, it is possible to gain a better understanding of the factors that cause humans to engage or disengage in a conversation or with any digital device they use in their daily life. This knowledge can be used to improve the design of human-computer interactions and enhance the overall user experience.

Despite its significance, the concept of complexity in natural language remains somewhat ambiguous, as there is a lack of a clear and consistent definition in the literature. One recurring theme in research is that the definition of complexity varies depending on the perspective used to study it. Indeed, complexity has been studied from various angles, including psycholinguistic, historical, neuroscientific, and computational, leading to multiple measures of complexity from each perspective. This diversity of definitions and measures highlights the complexity of the concept itself and the need for a more unified understanding.

According to a widely accepted distinction, complexity in natural language can be studied using two approaches: an *absolute* approach and a *relative* approach [198].

The absolute approach is theory-driven and defines complexity based on the number of parts in the linguistic system, i.e., the more parts a system has, the more complex it is. For example, a language with 34 phonemes is considered more complex than a language with only 18 phonemes. This measure of complexity is grounded in the idea that “an area of grammar is more complex than the same area in another grammar to the extent that it encompasses more overt distinctions and rules than another grammar” [193]. In this view, phonological and morphological phenomena play a key role in determining a language’s complexity.

The relative approach is user-oriented, defining complexity based on the cost and difficulty experienced by language users (speakers, listeners, or learners). In other words, it measures how difficult a phenomenon is to process or learn. Sentence complexity can be analyzed in terms of cognitive load, which can be inferred through offline processing measures, such as complexity judgments and error rates on comprehension tests, or through online processing measures, such as total gaze time, fixation duration, and pupil dilatation from eye-tracking data.

This case study adopts a relative approach, examining complexity from the perspective of human perception. The aim is not to provide a formal definition of complexity, but rather to understand the subjective nature of human perception of complexity in language and understand which linguistic phenomena impact on complexity perception.

There have been various metrics proposed to operationalize the factors that influence sentence processing performance. These metrics consider properties of individual words and sentences, as well as experience-based expectations. Word-level predictors that are correlated with higher cognitive load (i.e., processing difficulty) include word frequency, root frequency effect, and orthographic neighborhood frequency. At the syntactic level, a well-established complexity metric takes into account dependency length [104, 105]. Another explanation of processing difficulty is in terms of surprisal [119].

In this chapter, complexity is conceptualized in terms of perceived processing difficulty, or how difficult or easy a sentence is perceived to be by an individual. This approach recognizes the subjective nature of human perception of complexity and the impact of a person’s background and knowledge on their perception of complexity.

It is important to note that many previous studies on complexity have primarily focused on studying the complexity of sentences in isolation, linking syntactic and lexical properties with observed difficulty. However, models of language comprehension have emphasized the crucial role of contextual cues in creating a coherent representation of a text [145, 192]. Previous research has demonstrated the impact of context (i.e., the sentences preceding and following another sentence) on language comprehension and difficulty. For example, studies have shown that context can affect the readability of a sentence [249] and that ill-formed sentences are perceived as more acceptable when presented in context [33]. This highlights the importance of considering context when analyzing the complexity of language.

This study takes a comprehensive approach to analyzing and discussing the complexity of sentences both in isolation and within context. Through this double analysis, the study aims to address several research questions related to complexity:

1. To examine the role played by linguistic phenomena in human perception of language complexity.

2. To identify the key linguistic phenomena involved in predicting human agreement on complexity.

3. To understand which phenomena are associated with sentences that are perceived as complex by humans, and to use these phenomena to predict human complexity.

4. To describe the role played by context in the perception of complexity.

5. To demonstrate that models that rely on explicit linguistic features are better predictors of complexity judgments compared to state-of-the-art models that use implicit features.

4.2 Data

The data used in this case study to examine complexity consist of two monolingual specialized corpora, each containing sentences in a single language and belonging to a specific textual genre. Both corpora consist of sentences extracted from newspapers, with one corpus containing Italian sentences only and the other containing English sentences only.

By examining the complexity of sentences in two languages with different morpho-syntactic and syntactic properties, this study aims to understand if typologically different languages share common parameters of linguistic complexity. This comparative analysis was only performed when analyzing the complexity of sentences in isolation. When the context was introduced, only the English corpus was considered.

4.2.1 Dataset creation

For both languages, the sentences were extracted from manually revised treebanks to ensure accuracy and prevent errors that may be introduced by automatic annotations.

The Italian sentence corpus was created from the newspaper section of the Italian Universal Dependency Treebank (UDT) [187]. This treebank is annotated according to the Universal Dependencies (UD) scheme [211], which is designed to be interlingual, allowing for the annotation of different languages without the need for language-specific schemes. The use of the UD scheme enables easier inter-language comparisons and facilitates the creation of multilingual analysis tools.

The English sentence corpus was created from the automatically converted Wall Street Journal section of the Penn Treebank [190]. It is important to note that this treebank is not annotated with the same UD scheme as the Italian one. However, the use of a different annotation scheme for the English corpus does not pose a problem as the UD scheme is an evolution of the Stanford scheme [70] used to annotate the Penn Treebank. The choice of a different annotation scheme for the English corpus was made due to the genre of the sentences analyzed in this study. The UD treebank for English is built on texts extracted from web media (such as blogs, e-mails, and online product reviews), while the texts in the UDT mainly belong to the journalistic genre. By extracting texts from the Wall Street Journal of the Penn Treebank, this study is able to compare linguistic phenomena related to sentence complexity, minimizing possible cross-linguistic differences caused by inconsistent sentence structure representation principles. Different literary genres can encode various linguistic structures that would not make them comparable within the scope of this study.

After acquiring the data, a data selection phase was conducted. To reduce the impact of lexicon on sentence complexity, a strategy was implemented to remove sentences containing low-frequency lemmas from the two treebanks. This was done by using a lemma frequency list extracted from a large reference corpus, excluding numerals and proper nouns. For Italian, the reference corpus was PAISÁ [176], one of the most extensive corpora of contemporary Italian texts extracted from the internet. For English, the reference corpus was a large set of sentences extracted from the Wall Street Journal [212].

After the data selection phase, the sentences were grouped based on their length in terms of number of tokens. Six bins were created, each containing sentences with 10, 15, 20, 25, 30, and 35 tokens (for Italian with a range of +/- 1 token each). By controlling the length of the sentences, it was possible to determine whether the linguistic features known to correlate with sentence length still had an impact on complexity. As previously mentioned, it is widely accepted in the literature that sentence length is often associated with linguistic complexity, as longer sentences are typically perceived as more complex due to their longer syntactic dependencies or the presence of many relative clauses. Only the first 200 highest-ranked sentences from each bin were extracted, except for the last bin of Italian, which only contained 123 sentences. The final dataset used for the experiments consisted of 1,189 English sentences and 1,123 Italian sentences.

Dataset extension with context

The dataset was augmented in order to examine the impact of contextual factors on the perception of sentence complexity. The expansion was exclusively applied to the English sentences.

In this study, context is conceptualized as *"the preceding or succeeding sentence, or both, to a given sentence"*. The definition of a sentence in this scope is the text segment that lies between two full stops. For each sentence in the dataset, three separate *windows of context* were generated based on the relative position of the main sentence in relation to the contextual sentence:

- *begin window*: the sentence appears first and is followed by two contextual sentences;
- *center window*: the sentence is in the middle and is preceded by a contextual sentence and followed by another contextual sentence;
- *end window*: the sentence appears as the last one and is preceded by two contextual sentences.

The resulting dataset comprises 2,913 windows of context: 1,002 for the *begin* window, 986 for the *center* window, and 943 for the *end* window.

4.2.2 Dataset annotation

Human complexity judgments were collected with the aid of crowdsourcing. Two separate crowdsourcing tasks were established, one to assess complexity without context and the other to evaluate complexity in the presence of context. These tasks were conducted at different times, utilizing distinct participants and slightly varying conditions, as described in further detail below.

The task to assess complexity without context was implemented on the CrowdFlower platform[1]. For each language, 20 native speakers were recruited through

[1] As of the writing of this Chapter, the CrowdFlower platform has been discontinued and its services have been migrated to `www.appen.com` (last visit 09/07/2023).

the platform. Participants were instructed to read each sentence and to rate its difficulty level on a 7-point Likert scale, where 1 indicated *very easy* and 7 indicated *very difficult*. The sentences were presented in a randomized order and were displayed on separate pages, each containing five sentences. To ensure high-quality annotations, only workers with a "high quality" level assigned by the platform[2] were selected. In order to maintain the quality of the annotations, each participant was required to spend a minimum of ten seconds on each page.

The task to evaluate complexity in context was conducted on the Prolific platform[3]. For each contextual window, the sentence to be evaluated was presented in bold font, while the contextual sentences were displayed in plain font. The windows were presented in a randomized order and were displayed on separate pages, each containing ten windows. To manage the large number of context windows to be evaluated, the dataset was divided into smaller sections, with a maximum of 200 windows per section, resulting in 15 evaluation tasks. For each task, ten native English speakers were recruited. Participants were asked to read the entire window of context and to rate the complexity of the sentence presented in bold font on the same 7-point Likert scale used in the previous task. Given the subjective nature of complexity perception, the ratings were aggregated to account for individual biases among participants. There may be instances where one annotator consistently gave low scores, while another consistently gave high scores. The ratings were then re-scaled between 0 and 1 and normalized based on the range of ratings provided by each annotator.

4.3 Analysis of single sentence complexity

This Section focuses on the analysis of single sentence complexity in the data. The analysis aims to address the following objectives: (*i*) identifying linguistic phenomena that contribute to agreement among annotators in their complexity judgments, (*ii*) examining the linguistic aspects that are more strongly associated with a higher or lower perception of complexity, and (*iii*) determining the linguistic aspects that are better predictors of human complexity perception.

4.3.1 Study of the agreement

The first research question addressed in this study pertains to the evaluation of the linguistic phenomena that contribute to agreement among annotators. Specifically, the objective is to determine if certain patterns in sentence structure are correlated with complexity levels, thus leading different annotators to assign the same complexity judgment to a sentence. To achieve this, a new metric, referred

[2] This level was determined based on the worker's performance in previous tasks.

[3] `www.prolific.com` (last visit 09/07/2023)

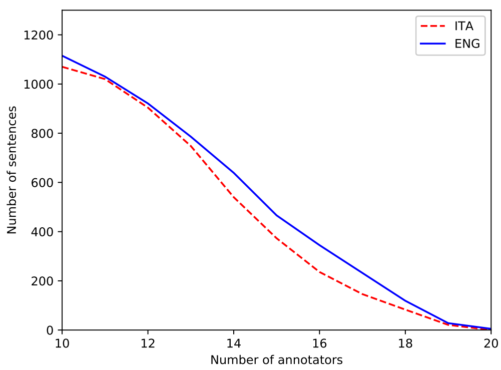


Figure 4.1: Number of sentences for each degree of agreement.

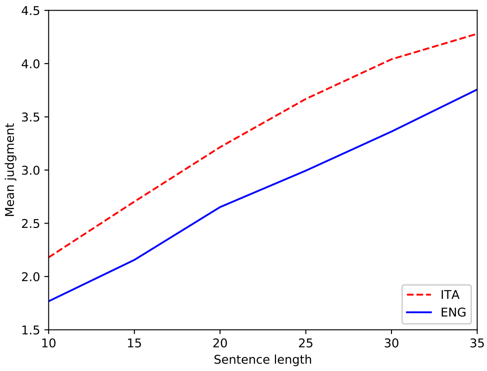


Figure 4.2: Mean complexity judgment at different sentence lengths.

to as *degrees of agreement*, is introduced. This metric measures the number of annotators who assign a complexity judgment within the same range, calculated as a standard deviation from the mean judgment for each sentence. Applying this measure, the rated sentences are split into ten sets, each corresponding to a different degree of agreement.

Figure 4.1 displays the number of sentences for each degree of agreement, ranging from a minimum degree of agreement of 10 to a maximum of 20, along with the total number of annotators involved in the study. For each language, a small number of sentences are excluded when considering a minimum agreement of 10 annotators. This indicates that there are a substantial number of sentences (approximately 1100) on which at least 10 annotators agree in assigning a complexity judgment within the same range. As the number of annotators in agreement increases, the number of sentences decreases, yet a significant number of sentences (approximately 600) still remain at a degree of agreement of 14. When the degree of agreement reaches 20, the number of remaining sentences is zero, indicating that there are no sentences to which all 20 annotators have assigned a complexity judgment within the same range.

Subsequently, the sentences are analyzed in terms of the linguistic features outlined in Section 2.2.2 to determine the linguistic phenomena that contribute to agreement among annotators. The features were extracted from sentences where annotators agreed (referred to as *agreed sentences*) and from sentences where annotators did not agree (referred to as *not-agreed sentences*). The Wilcoxon Rank-sum test (detailed in Section 3.1.3) was used to determine if there is a statistically significant difference between the two groups of sentences. This process was repeated for each agreement threshold.

The next step in addressing the first research question was to conduct a feature selection process to identify the features that maximize the accuracy of a classifier in distinguishing between agreed and not-agreed sentences. The Support Vector

Machine Classifier (SVC) algorithm (Section 3.2.1) was used as the estimator, and the Recursive Feature Elimination algorithm[4], outlined in Section 3.2.3, was used to create a ranking of feature relevance. At each iteration of the algorithm, a single feature was dropped, and the performance of the classifier was evaluated using a 3-fold cross-validation method. This process was repeated ten times for each degree of agreement, and the top-ranked features were then selected. The accuracy of the Support Vector Classifier on unseen data was evaluated against a baseline, which corresponds to the performance of a most likely classification method, where each sentence is always classified into the most likely class.

Table 4.1 shows the features that vary in a statistically significant way (marked with ✓) according to the Wilkonson Rank-sum test and the ones selected in the classification task with the Support Vector Machine (marked with $\star$), for both languages and at different levels of agreement. The features are grouped into sections based on the linguistic phenomena they describe. As evident, there is an opposite trend between the statistically significant features and those selected by the classifier as the degree of agreement increases. For what concerns the Wilcoxon test, very few features have significantly different values at lower degrees of agreement. Namely, very few features are involved in discriminating the *agreed* vs. *not-agreed* sentences, especially when the agreement is below 14.

For both languages raw text features (*n_tokens*, *char_tok*) vary significantly for most levels of agreement. It is noteworthy that while these two features are significant, they are not given equal consideration by the classifier, which instead relies on more complex syntactic features, such as features related to subordination (*subord_depth*) and nominal modification (*prep_chain_l*). Syntactic features begin to vary significantly as the agreement increases; this happens, for instance, for features related to the parsed tree structure, such as the depth of the whole parsed tree (*max_depth*) and the chains of complements (*dep_mark*, or features related to the use of subordination (section *subordination phenomena* in the table).

By comparing the two languages, some differences emerge. At the lowest agreement (degree 10), different features in all groups vary significantly for English, while for Italian, the *agreed* and *not-agreed* sentences do not vary for any features. At higher degrees of agreement, the *agreed* sentences in Italian are characterized by the variation of two language-specific features: the position of the object with respect to the verbal head (*order_obj*) and some morphological features about verbs (*verbs_num_pers*, *verb_tense*), which also contribute to the classification only for Italian.

Table 4.2 shows the accuracy of the SVC for each degree of agreement and the baseline (computed with the most likely classification method). The accuracy of

[4] Implemented in the Scikit-learn library [220].

	Agreement															
	10		**11**		**12**		**13**		**14**		**15**		**16**		**17**	
Feature	IT	EN	IT	EN	IT	EN	IT	EN	IT	EN	IT	EN	IT	EN	IT	EN
raw text properties																
char_tok	⋆	⋆	⋆	⋆	⋆	-	-	✓	✓	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆
n_tokens	-	✓⋆	✓⋆	✓⋆	✓⋆	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
lexical variety																
lex_density	-	⋆	-	⋆	-	-	-	✓⋆	-	✓⋆	-	✓	-	✓⋆	-	✓⋆
ttr_form	-	✓⋆	⋆	⋆	✓⋆	✓	✓	✓	✓	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓	✓⋆
ttr_lemma	⋆	✓⋆	⋆	✓⋆	⋆	✓	✓	✓⋆	✓	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆
morpho syntactic information																
cpos_ADJ	⋆	⋆	⋆	⋆	⋆	-	-	-	✓⋆	-	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓
cpos_ADP	⋆	⋆	⋆	⋆	⋆	-	-	⋆	-	-	-	-	✓	✓	✓	✓
cpos_ADV	⋆	-	⋆	-	⋆	-	-	-	-	-	⋆	-	⋆	-	⋆	-
cpos_AUX	⋆	-	⋆	-	⋆	-	✓	-	-	-	-	-	✓⋆	-	✓	-
cpos_CONJ	⋆	⋆	⋆	⋆	⋆	-	-	⋆	-	✓	✓	✓⋆	✓⋆	✓	✓	✓
cpos_PRON	⋆	-	⋆	-	⋆	-	-	-	✓	-	✓⋆	-	✓	-	✓	-
cpos_DET	-	⋆	-	⋆	-	-	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
cpos_NUM	-	⋆	-	✓⋆	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
cpos_PROPN	⋆	-	⋆	-	⋆	-	-	-	✓	-	⋆	-	✓⋆	-	-	
cpos_PUNCT	⋆	-	⋆	-	⋆	-	✓	-	-	-	✓⋆	-	✓	-	✓⋆	-
cpos_SCONJ	⋆	-	⋆	-	⋆	-	-	-	-	-	✓⋆	-	✓	-	✓	-
cpos_VERB	-	⋆	-	⋆	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓
verbs_num_pers	⋆	-	⋆	-	⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-
verbs_tense	⋆	-	⋆	-	✓⋆	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-
verbal predicate structure																
verb_arity	⋆	⋆	⋆	⋆	✓⋆	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
verb_head_arity	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	✓⋆	⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆
verb_head	⋆	⋆	⋆	⋆	✓⋆	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
global and local parsed tree structure																
links_len	-	✓⋆	⋆	⋆	✓⋆	✓	✓	✓	✓⋆	✓	✓	✓⋆	✓	✓	✓	✓
max_depth	-	⋆	⋆	⋆	✓⋆	-	✓	✓	✓	✓	✓	✓	✓	✓⋆	✓	✓
max_links_l	-	✓⋆	⋆	✓⋆	✓⋆	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
n_prep_chains	⋆	✓⋆	✓⋆	⋆	✓⋆	✓	✓	✓	✓	✓	✓	✓	✓	✓⋆	✓	✓
order_obj	-	-	⋆	-	⋆	-	-	-	-	-	✓	-	✓	-	✓	-
order_subj	-	-	⋆	-	⋆	-	-	-	⋆	-	-	-	✓	-	✓	-
prep_chain_l	-	⋆	⋆	⋆	⋆	-	✓	-	✓	✓	✓⋆	✓	✓⋆	✓	✓	✓
prep_depth	-	✓⋆	⋆	⋆	✓	✓	✓	✓	✓	✓⋆	✓	✓⋆	✓	✓⋆	✓	✓⋆
token_clause	-	⋆	⋆	⋆	⋆	-	-	-	-	-	-	-	✓	✓	✓	✓
syntactic relations																
dep_acl	-	-	⋆	-	⋆	-	✓	-	✓	-	✓	-	✓	-	✓	-
dep_acl:relcl	-	-	⋆	-	⋆	-	-	-	⋆	-	✓	-	✓⋆	-	✓	-
dep_adpobj	-	⋆	-	⋆	-	-	-	⋆	-	-	-	-	-	-	-	✓
dep_advcl	⋆	-	⋆	-	⋆	-	-	-	✓	-	✓	-	✓⋆	-	✓	-
dep_amod	⋆	✓⋆	⋆	⋆	⋆	✓	-	✓⋆	✓	✓	✓	✓⋆	✓⋆	✓⋆	✓	✓⋆
dep_appos	-	⋆	-	⋆	-	-	-	-	-	⋆	-	-	-	-	-	-
dep_attr	-	⋆	-	⋆	-	-	-	-	-	-	-	✓⋆	-	✓⋆	-	✓
dep_aux	-	-	⋆	-	⋆	-	✓	-	✓	-	-	-	✓⋆	-	✓	-
dep_case	⋆	-	⋆	-	⋆	-	-	-	⋆	-	-	-	✓	-	✓	-
dep_cc	⋆	⋆	⋆	⋆	⋆	-	-	-	-	✓⋆	✓	✓⋆	✓⋆	✓⋆	✓	✓
dep_ccomp	-	⋆	-	⋆	-	-	-	-	-	✓	-	✓	-	✓	-	✓
dep_compmod	-	⋆	-	⋆	-	-	-	-	-	✓⋆	-	⋆	-	✓⋆	-	✓⋆
dep_conj	⋆	⋆	⋆	⋆	⋆	-	-	✓⋆	-	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓	✓⋆
dep_det	-	⋆	-	⋆	-	-	-	⋆	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
dep_dobj	⋆	-	⋆	-	⋆	-	-	-	-	-	✓	-	✓⋆	-	✓	-
dep_mark	⋆	⋆	⋆	⋆	⋆	-	✓	⋆	✓	⋆	✓⋆	⋆	✓⋆	✓	✓	✓
dep_nmod	⋆	⋆	⋆	⋆	✓⋆	-	✓	-	✓	-	-	✓⋆	✓⋆	✓	✓	✓
dep_nsubj	-	✓⋆	-	✓⋆	-	✓	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
dep_num	-	⋆	-	⋆	-	✓	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆
dep_partmod	-	⋆	-	⋆	-	-	-	-	-	-	-	✓	-	✓	-	✓
dep_poss	-	⋆	-	⋆	-	-	-	-	-	✓	-	✓	-	✓	-	✓
dep_punct	⋆	-	⋆	-	⋆	-	✓	-	-	-	✓⋆	-	✓	-	✓	-
dep_rcmod	-	⋆	-	⋆	-	-	-	⋆	-	-	-	✓⋆	-	✓⋆	-	✓
dep_xcomp	⋆	-	⋆	-	⋆	-	-	-	-	-	-	-	✓	-	✓	-
subordination phenomena																
n_principal_clauses	-	⋆	⋆	⋆	⋆	-	✓	✓	✓	✓⋆	✓	✓	✓	✓	✓	✓
n_subord_chain	⋆	⋆	⋆	⋆	⋆	✓	✓	-	✓⋆	✓⋆	✓⋆	✓	✓⋆	✓	✓	✓
n_subord_clauses	⋆	-	⋆	-	⋆	-	✓	-	✓⋆	-	✓⋆	-	✓⋆	-	✓⋆	-
order_subord	⋆	⋆	⋆	⋆	⋆	-	✓	✓	✓	✓⋆	✓	✓	✓	✓	✓	✓
subord_depth	⋆	⋆	⋆	⋆	⋆	-	✓⋆	⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆	✓⋆

Table 4.1: Linguistic features that vary statistically and features selected by the SVM classifier in at least 50% of the 10 runs (⋆) for Italian and English at different degrees of agreement.

the classifier is computed as the average classification score of the 10 best results of the feature selection process. At low degrees of agreement (< 14) the classifier achieves lower accuracy compared to the baseline, showing that the selected features do not contribute to discriminating *agreed* vs. *non-agreed* sentences. These features begin having a more significant impact on the classification of sentences as the degree increases (from degree 14 to degree 17). This result indicates that at these degrees of agreement, the values of the features characterizing the *agreed* sentences are considerably different from those of the *non-agreed* sentences. Moreover, even though a very high number of features are considered statistically significant by the Wilcoxon test for these sentences, the classifier needs fewer features to assign the correct class (as shown in Table 4.1).

	Baseline Accuracy (%) - SVC Accuracy (%)							
	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Italian	95.4-95.4	91-90.8	80.6-80.5	66.7-66	51.9-59.1	66.8-68.8	79-80.7	87-87.1
English	94-94	86.8-86.8	83.6-77.4	66.3-66.1	53.9-60	60.7-71.8	70.9-79.3	80.4-84.6

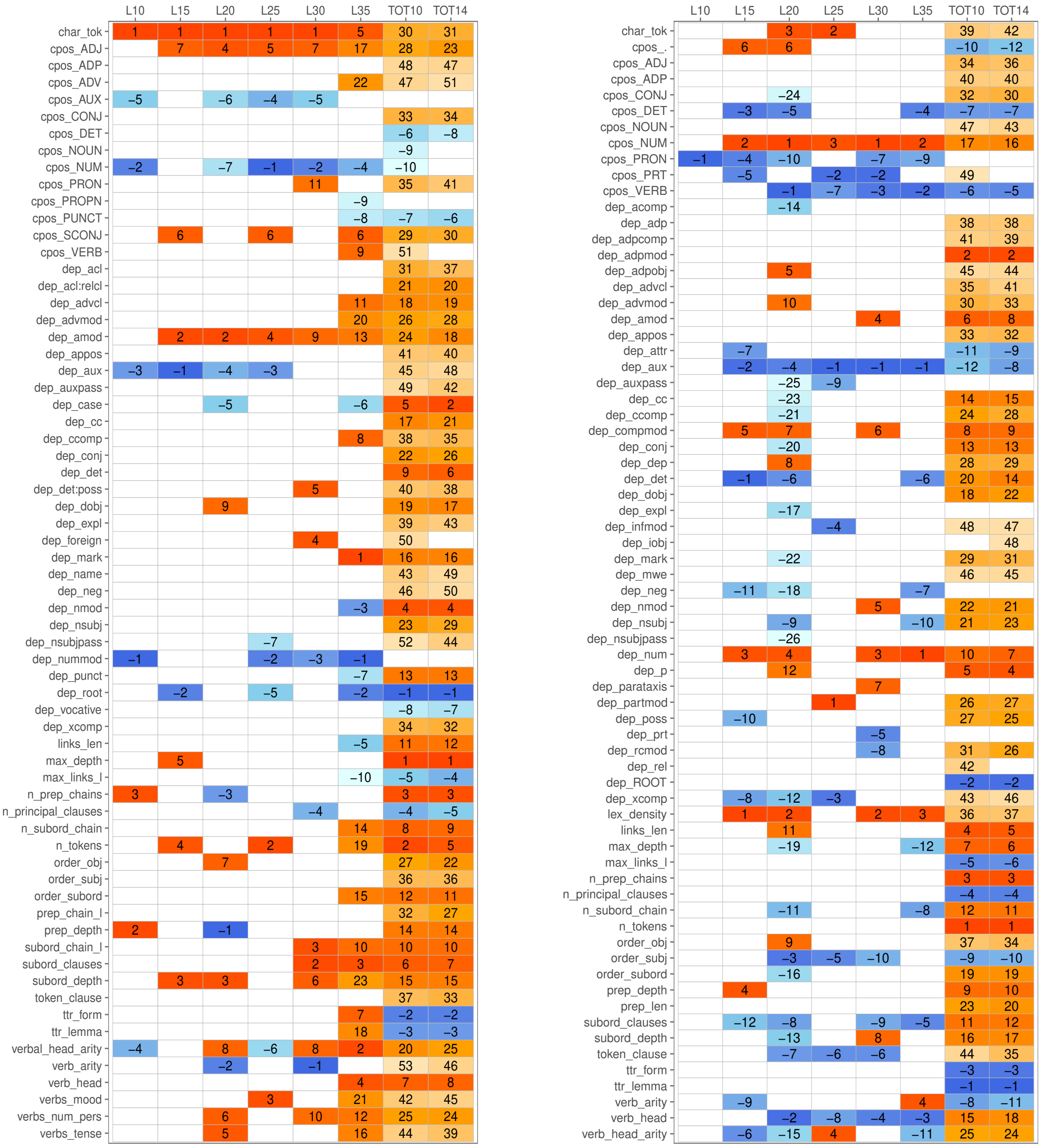
Table 4.2: Baseline and Support Vector Classifier accuracy at different degrees of human agreement.

4.3.2 Correlation between linguistic features and complexity

The second research question addressed in this study aims to model human perception of complexity by examining the correlation between linguistic features extracted from sentences and the complexity judgments assigned to each sentence. To achieve this, the average complexity judgment for each of the six bins of sentences of the same length (10, 15, 20, 25, 30, 35 tokens) was calculated. As illustrated in Figure 4.2 on page 84, both languages rated longer sentences as more complex, although all sentences were consistently rated as more complex in Italian.

Next, the Spearman's Rank Correlation Coefficient (detailed in Section 3.1.1) was calculated between the values of each feature and the average complexity judgments to obtain a ranking of features. The correlation was computed at two distinct degrees of agreement, 10 and 14, as these two thresholds were selected for analysis. At degree 10, the *agreed* sentences correspond to nearly all of the rated sentences, and at degree 14, the Support Vector Machine classifier starts to outperform the baseline (as shown in Table 4.2). Additionally, at degree 14, there is still a substantial set of *agreed* sentences remaining, allowing for a reliable statistical examination of the features (as depicted in Figure 4.1). For threshold 10, the ranking of features with respect to the six bins of sentences of the same length (L10, L15, L20, L25, L30, L35) was also calculated.

Figure 4.3 displays the ranking of features with $p < 0.05$. A positive value



(a) Italian

(b) English

Figure 4.3: Features correlating with human judgments at different sentence lengths and with respect to the sentences at agreement 10 (*TOT 10*) and 14 (*TOT 14*).

indicates that as the feature value increases, the sentence is perceived as more complex (i.e., the feature ranked +1 is the top-ranked feature as it is most positively correlated). Conversely, a negative value indicates that as the feature value decreases, the sentence is perceived as more complex (i.e., the feature ranked −1 is the most highly negatively correlated).

For both languages, the correlation between the top 20 ranked features and the judgment of complexity is substantial, ranging from 0.30 to 0.85 when considering sentences at agreement 14. At the two agreement thresholds for all lengths (columns $TOT10$ and $TOT14$ in the Figure), highly correlated features concern not only sentence length but also deep syntactic features, such as the depth of the whole parsed tree (*max_depth*), the length of dependency links (*links_len*), and features related to subordination (*n_subord_clauses*). In more detail, the 1st-ranked feature in Italian (*max_depth*, the maximum depth of the parsed tree) and the 1st-ranked feature in English (*n_tokens*, the length of the sentences) have a correlation of 0.64 and 0.84 respectively. Nominal modifications (*n_prep_chains* is also highly correlated (Italian $\rho = 0.59$, English $\rho = 0.54$) and equally ranked at the 3rd position in both languages. The distribution of *verbs_num_pers*, i.e., the distribution of the number and the person of the subjects of verbs, makes the sentence more complex only for Italian. This may be related to higher complexity in verbal morphology, as the use of third-person verbs in impersonal structures may increase sentence ambiguity with respect to the referent.

In English, sentence complexity is also influenced by the distribution of cardinal numbers (*cpos_NUM*) and the presence of a numeric modifier as a dependent (*dep_num*). This result is in line with the difficulty of numerical information shown in readability studies [26]. On the other hand, the relative ordering of subjects with respect to the verbal head and the verbal arity have a lower position in the negative ranking, suggesting that these features make a sentence easier. This outcome might be due to a more fixed predicate-argument structure and word order in the English language.

When focusing on sentences of the same length, features considered as a proxy of lexical complexity are in top positions in both languages. This is the case of the average word length (*char_tok*) and, only for English, of the lexical density (*lex_density*). While most features have similar rankings in all bins of same-length sentences for English, substantial differences can be observed between the ranking of features extracted from sentences that are $\leq$ 20 tokens and $\geq$ 20 tokens in length for Italian. In particular, when the average sentence length is $\geq$ 20 tokens, features related to subordination make the sentence more complex.

4.3.3 Prediction of sentence complexity judgments

This part of the work analyzes how linguistic features contribute to predicting the judgment of complexity assigned by humans to a sentence. To this end, a linear Support Vector Regression model (SVR) (see Section 3.2.2) is trained with default parameters using the Scikit-learn [220] implementation. A 3-fold cross-validation is performed over each subset of *agreed* sentences at degrees of agreement 10 and 14. The performance of the model is measured with two metrics: (1) the mean absolute error to evaluate the accuracy of the model in predicting the same complexity judgments assigned by humans to the sentences and (2) Spearman's

	IT-10	**IT-14**	**EN-10**	**EN-14**
mean abs err 1	0.77	0.78	0.71	0.68
Spearman 1	0.57	0.64	0.68	0.64
mean abs err 2	0.79	0.80	0.70	0.70
Spearman 2	0.55	0.63	0.67	0.73
mean abs err 3	0.85	0.75	0.77	0.60
Spearman 3	0.55	0.64	0.61	0.71
avg mean abs err	**0.80**	**0.78**	**0.72**	**0.66**
avg Spearman	**0.56**	**0.63**	**0.65**	**0.69**

Table 4.3: Performance of the linear SVM regression model and the avg score at different agreements.

correlation coefficient (Section 3.1.1) to evaluate the correlation between the ranking of features produced by the regression model and the ranking produced by the human judgments.

Table 4.3 shows the outcome of each cross-fold validation and the resulting average score of the two metrics. The model is very accurate and achieves a very high correlation of $\rho > 0.56$ with p-value < 0.001, and an average error difference (*avg mean abs err*) below 1. Particularly, the model obtained higher performance predicting the ranking of features extracted from sentences at the degree of agreement 14. This might be because these sentences are characterized by a more uniform distribution of linguistic phenomena and that these phenomena contribute to predicting the same judgment of complexity. These results are in line with the ones obtained by the Support Vector Classifier in predicting the degrees of agreement (Table 4.2 on page 87). This is more relevant for English, and it possibly suggests that the set of sentences similarly judged by humans are characterized by a lower variability of the values of the features.

4.4 Analysis of sentence complexity in context

This section of this case study is dedicated to the analysis performed on the data regarding the complexity of sentences in context. Here the objective is to pinpoint which linguistic phenomena extracted from the single sentences or the contextual sentences (i) influence the agreement on complexity between annotators, (ii) are correlated with complexity and complexity standard deviation in different contextual windows, and (iii) if these linguistic features are less or more effective than state-of-the-art language models in predicting complexity judgments in a low resource scenario.

4.4.1 Study of the agreement

As what has been discussed for single sentences in Section 4.3.1, the first step of this analysis was to look at the *degrees of agreement* between annotators. The definition of the degree of agreement is the same as in Section 4.3.1.

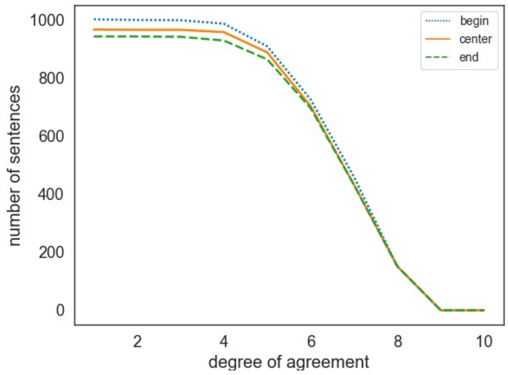


Figure 4.4: Number of sentences for each degree of agreement.

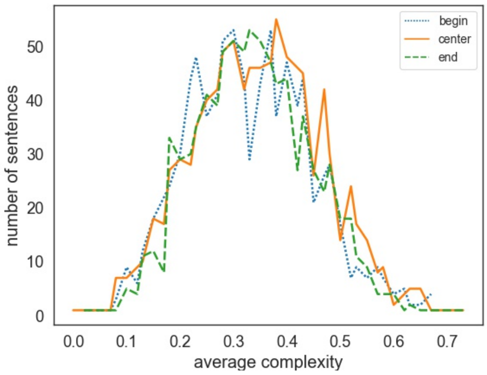


Figure 4.5: Number of sentences at different average complexity ratings.

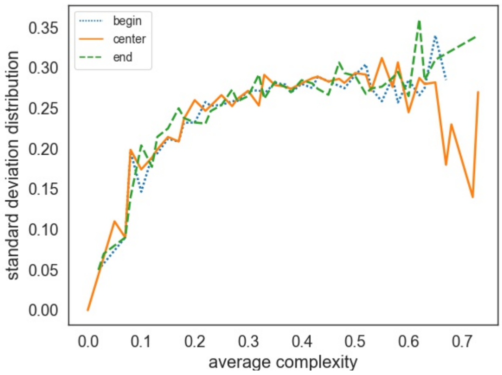


Figure 4.6: Mean standard deviation at different average complexity ratings.

Figure 4.4 reports the number of sentences for every degree of agreement, considering the different sentence positions within the context windows. A strong degree of agreement is found for all three windows, as most sentences have up to 5 annotators that have assigned a complexity judgment within the same range. As the degree of agreement increases, the number of sentences decreases consistently. This result is in line with what was detected when analyzing the complexity of single sentences out of context. The highest agreement is found at 8 annotators,

	begin		**center**		**end**	
	judg	*std*	*judg*	*std*	*judg*	*std*
Length 10	0.28	0.23	0.28	0.28	0.28	0.28
Length 15	0.27	0.23	0.32	0.28	0.30	0.28
Length 20	0.27	0.22	0.35	0.27	0.33	0.26
Length 25	0.26	0.21	0.36	0.26	0.35	0.26
Length 30	0.26	0.22	0.38	0.26	0.36	0.25
Length 35	0.25	0.21	0.39	0.26	0.38	0.26
All sents	0.26	0.22	0.35	0.27	0.33	0.27

Table 4.4: Mean complexity judgment and mean standard deviation on complexity for all sentences and at different lengths.

but on a small number of sentences (less than 200), while there are no sentences on which 9 or 10 annotators agree in any of the context windows. Also, this preliminary analysis suggests that the sentence position in the context window has little to no influence on the degree of agreement, as the numbers mostly follow the same trend in each context window. This assumption is confirmed by the results shown in Figure 4.5, in which the number of sentences that were assigned the same average complexity value is plotted against the same average complexity value. It appears clear that the average complexity follows a Gaussian distribution for all the windows of context, as most sentences received an average complexity between 0.2 and 0.4, regardless of their position in the context window.

To further address the relationship between complexity and agreement, it was calculated the standard deviation of the complexity judgments that were assigned to each sentence. In Figure 4.6 the standard deviation of each sentence is plotted against the average complexity assigned to the same sentence for the three windows of context. In the cases in which more than one sentence was assigned the same average complexity values, the average standard deviation of all the sentences is plotted. From the Figure, it is noticeable that the standard deviation tends to increase with the average complexity score assigned to sentences. This means that annotators agree more on rating a sentence as simple, suggesting that the perception of a sentence as more complex may be less homogeneous. This trend is overall similar for all contextual windows, although it appears to be a more uniform behavior in rating a sentence as more complex when surrounded by both contextual sentences (i.e., in the center window).

Besides the position of the sentence in the context window, also the length of a sentence may affect the perception of complexity (as described previously in Section 4.2.1). Thus, it was calculated the average complexity judgment assigned to sentences of the same length, for all three context windows, along with the average standard deviation. As visible in Table 4.4, the average complexity values tend to increase with the length of sentences for the center and the end window, as expected. On the other side, the standard deviation follows an opposite trend, marking that subjects agree more on the complexity of longer sentences (e.g., lengths 30 and 35); their perception of shorter sentences, instead, is more diversified. The data also show that when the sentence is at the beginning of the sentence, it is overall perceived as less complex. This could indicate that the following contextual sentences help the annotators in both the understanding and the processing of the first sentence.

Table 4.5 contains examples of sentences whose complexity scores vary the least or the most within the different context windows. In the case of *Zero Variance*, the sentence obtained the same average complexity, regardless of the relative position it had in the contextual window (begin, center, or end). Conversely, sentences with the highest variance received very diverse average values according to the position occupied by the sentence in the contextual windows. The Table also reports the actual average complexity values that the sentences with the highest variance got for each position.

4.4.2 Correlation between linguistic features and complexity

As described in Section 4.3.2, it was performed a correlation analysis also for complexity in context. The correlation was computed between the complexity score assigned to each sentence and the set of explicit linguistic features (Section 2.2.2) extracted from the sentence. This analysis is used to detect which linguistic phenomena are more involved in the assessment of sentence complexity and to verify whether these phenomena capture information about the sentence itself or about the context.

For each sentence, it was computed the Spearman's rank correlation coefficient (Section 3.1.1) between the average complexity score and the value of each linguistic feature extracted from *i*) the rated sentence, *ii*) its preceding one and *iii*) its following one, according to the contextual window. The correlation analysis was performed on the sentences altogether and then on the sentences divided into bins according to their length. The same process was repeated by correlating the standard deviation of complexity scores with the linguistic features of each sentence.

In what follows are discussed only the correlation results for the subset of sentences presented in the center window, since this is the only window in which the rated sentence is always surrounded by both a left and a right context sentence, allowing a comparison between the two context positions. In Appendix B are reported the full tables of correlation results for all contextual windows.

	Zero Variance	BCE	Highest Variance	B	C	E
Length 10	Tokyo's Nikkei index fell 84.15 points to 35442.40.	0.38	Nashua announced the Reiss request after the market closed.	0.22	0.42	0.63
Length 15	Elsewhere in Europe, share prices closed higher in Stockholm, Brussels and Milan.	0.23	Last year, the prisons' sales to the Pentagon totaled $336 million.	0.62	0.32	0.20
Length 20	Dow Jones industrials 2645.08, up 41.60; transportation 1205.01, up 13.15; utilities 219.19, up 2.45.	0.50	The cash dividend paid on the common stock also will apply to the new shares, the company said.	0.12	0.12	0.55
Length 25	In the nine months, Milton Roy earned $6.6 million, or $1.18 a share, on sales of $94.3 million.	0.38	Yesterday, Compaq plunged further, closing at $100 a share, off $8.625 a share, on volume of 2,633,700 shares.	0.25	0.67	0.42
Length 30	SsangYong, which has only about 3% of the domestic market, will sell about 18,000 of its models this year, twice as many as last year.	0.32	Though not reflected in the table, an investor should know that the cost of the option insurance can be partially offset by any dividends that the stock pays.	0.23	0.50	0.57
Length 35	In the nine months, net rose 35% to $120.1 million, or $1.64 a share, from $89.20 million, or $1.22 a share, a year earlier.	0.48	William Kaiser, president of the Kaiser Financial Group in Chicago, said the decline was almost certainly influenced by the early sell-off in the stock market, which partly reflected a weakening economy.	0.45	0.23	0.58
All sents	Dow Jones industrials 2645.08, up 41.60; transportation 1205.01, up 13.15; utilities 219.19, up 2.45.	0.50	The cash dividend paid on the common stock also will apply to the new shares, the company said.	0.12	0.12	0.55

Table 4.5: Sentences which vary the least or the most within context windows. B, C, and E respectively indicate the begin, center, and end windows.

Correlations with the average complexity score

Table 4.6 shows the top ten features ranked by the correlation score with average complexity, for all sentences and all groups of sentences of the same length. A

Features	L10	L15	L20	L25	L30	L35	All
B_dep_aux:pass	-	-	-	−1	-	-	-
B_dep_compound	-	-	5	-	-	-	-
B_dep_compound:prt	−4	-	-	-	-	-	-
B_dep_flat	-	-	-	−5	-	-	-
B_dep_nmod	-	-	-	5	-	-	-
B_dep_nsubj	-	−5	-	-	-	-	-
B_dep_nsubj:pass	-	-	-	−2	-	-	-
B_dep_nummod	-	-	3	-	-	-	-
B _princ_prop	-	-	−4	-	-	-	-
B_verb_root_perc	-	-	−3	-	-	-	-
C_aux_Fin	-	-	−1	-	−4	-	-
C_aux_num_pers_+	−5	-	−5	-	-	-	-
C_aux_Pres	-	-	-	-	−5	-	-
C_avg_max_depth	-	5	-	-	-	-	4
C_avg_max_link	-	-	-	-	-	-	8
C_avg_sub_chain	-	-	-	-	-	−1	-
C_avg_tok_clause	-	-	4	-	-	-	-
C_char_tok	-	-	-	-	-	−5	-
C_dep_aux	-	-	-	-	−2	-	-
C_dep_det	-	−3	-	-	-	-	-
C_dep_nmod	5	-	-	-	-	-	-
C_dep_nummod	-	4	2	-	2	2	5
C_dep_root	-	−1	-	-	-	-	−1
C_dep_xcomp	-	-	-	−3	-	-	-
C_max_link	-	-	-	-	-	-	7
C_n_prep_chain	-	-	-	-	-	-	6
C_n_tok	3	2	-	-	-	-	2
C_tok_sent	4	3	-	-	-	-	3
C_upos_ADJ	-	−4	-	-	-	-	-
C_upos_AUX	-	-	−2	-	−3	−2	−2
C_upos_DET	-	−2	-	-	-	-	-
C_upos_NUM	1	1	1	-	1	1	1
C_upos_PRON	-	-	-	-	-	−3	-
C_upos_SYM	2	-	-	-	-	3	-
C_verb_edge_1	-	-	-	-	−1	-	-
C_verb_Fin	-	-	-	-	3	-	-
C_verb_Ind	-	-	-	-	5	-	-
E_aux_Pres	−3	-	-	-	-	-	-
E_avg_link	−2	-	-	-	-	-	-
E_avg_max_depth	-	-	-	2	-	-	-
E_dep_ccomp	-	-	-	-	-	−4	-
E_dep_nummod	-	-	-	4	-	5	-
E_lexical_dens	-	-	-	−4	-	-	-
E_upos_NUM	-	-	-	1	-	4	-
E_upos_SYM	-	-	-	3	-	-	-
E_verb_edge_4	−1	-	-	-	-	-	-
E_verb_Fin	-	-	-	-	4	-	-

Table 4.6: Ranking of correlations between the top 10 linguistic features and the average complexity score for all sentences and for all length bins. The number indicates the position the feature occupies in the ranking: the higher the number(positive or negative), the higher the correlation. B_*, C_*,E_* mean that the features characterize the beginning, the central, and the ending sentence, respectively.

positive number indicates that the feature is linked to higher perceived complexity, meaning that linguistic phenomenon makes the sentence more complex in the eyes of annotators. Conversely, a negative number is linked to lower complexity, meaning the linguistic phenomenon helps annotators in the evaluation of sentence complexity. Examining the results for the correlations with the average complexity score in the Table, it emerges that statistically significant correlations (p-value < 0.05) with $\rho \geq \pm 0.20$ were found for 103 features out of the whole set. Among them, 44% belongs to the rated sentence (i.e., 45 features) and 56% to the contextual ones (i.e., 23 and 35 features to the left and the right sentence, respectively). Although it was expected that many features extracted from the rated sentence were correlated to complexity judgments, these results also suggest that humans have paid attention to the whole context when rating the middle sentence, and especially to the following sentence. The influence of context is also suggested by the fact that it is possible to observe much lower coefficients for all correlating features belonging to the rated sentence, unlike those reported in Section 4.3.2 for the same sentences evaluated in isolation.

When all sentences are considered, the first ten ones all belong to the middle sentence and refer to features modeling linguistic phenomena of different nature; nevertheless, two main groups are distinguishable, both positively correlated with the perception of sentence complexity. The first group is related to the presence of numerical information (i.e., literal numbers in the sentence), as conveyed by both POS and syntactic features (*C_upos_NUM, C_dep_nummod*). The second one, as more expected, concerns sentence length (*C_tok_sent, C_dep_root*) and features still related to length but capturing aspects of structural complexity, such as the depth of the whole parse tree and specific sub-trees, i.e., nominal chain headed by a preposition (*C_avg_max_depth, C_n_prep_chain*). Notably, the effect of sentence length is observed only for the middle sentence, while the length of contextual sentences is never correlated with complexity judgments. Even in this case, the correlation is much lower with respect to the one obtained by sentences judged in isolation (i.e., 0.31 vs. 0.84 reported in the previous study).

When analyzing bins of same-length sentences, it is noticeable that there is a more prominent role of features from the context, as suggested by the presence of features characterizing both the sentence preceding and following the rated one in the first ten positions of the ranking. Interestingly, for all bins, numerical information turned out to be the feature most correlated with complexity score, being it extracted from the rated or from contextual sentences (specifically, the right sentence, for the bin composed of sentences with 25 tokens).

Correlations with the standard deviation of complexity scores

Table 4.7 reports the first ten features more strongly correlated with the standard deviation of assigned complexity scores, for all rated sentences and all groups of

Features	L10	L15	L20	L25	L30	L35	All
B_aux_Inf	2	-	-	-	-	-	-
B_dep_compound:prt	-5	-	-	-	-	-	-
B_subj_pre	-	-	-	-5	-	-	-
B_upos_SYM	-	-	-	-3	-	-	-
B_verb_edge_1	-	-2	-	-	-	-	-
B_verb_Past	-	-1	-	-	-	-	-
C_avg_sub_chain	-	-	-	-	-	-	-1
C_char_tok	-	-	-	-	-	-	-5
C_dep_aux	-	-	-	-	-1	-	-
C_dep_nummod	-	-	-	-	-	-	2
C_dep_punct	-	-	-	-4	-	-	-
C_princ_prop	-	-	-	-	-	-2	-
C_sub_prop	-	-	-	-	-	3	-
C_upos_AUX	-	-	-	-	-	-	-2
C_upos_NUM	-	-	-	-	-	-	1
C_upos_PRON	-	-	-	-	-	-	-3
C_upos_PUNCT	-	-	-	-2	-	-	-
C_upos_SYM	-	-	-	-	-	-	3
C_verb_edge_1	-	-	-	-	-	2	-
C_verb_root_perc	-	-	-	-	-	-1	-
E_avg_link	-4	-	-	-	-	-	-
E_avg_max_link	-2	-	-	-	-	-	-
E_dep_aux	-6	-	-	-	-	-	-
E_dep_ccomp	-	-	-	-	-	-	-4
E_dep_nummod	-	-	-	-	-	-	5
E_dep_parataxis	-7	-	-	-	-	-	-
E_dep_root	1	-	-	-	-	-	-
E_max_link	-3	-	-	-	-	-	-
E_upos_ADV	-	-	1	-	-	-	-
E_upos_NUM	-	-	-	-	-	-	4
E_verb_edge_3	-	-	-	-1	-	1	-
E_verb_Past	3	-	-	-	-	-	-
E_verb_Pres	-1	-	-	-	-	-	-

Table 4.7: Ranking of correlations between the top 10 linguistic features and complexity standard deviation for all sentences and for all length bins. Feature labels and ranking numbers are used as in Table 4.6.

sentences of the same length. The Table is structured with the same rules as per Table 4.6. In this analysis, 29 statistically significant ($p < 0.05$) features with correlation $\rho \geq 0.20$ were found. These include 24% of features belonging to the rated sentence (i.e., 7 features), while the remaining features belong to the contextual sentences (i.e., 6 features for the left sentence, 15 for the right one). In this case, far fewer correlations were found if compared with the correlations with the average complexity, with most features being significant for the length bins but not when considering the sentences altogether. These results confirm that humans have paid attention to the whole context when evaluating the sentence, but also that standard deviation, and thus annotators' agreement, is a phenomenon harder to describe and subjective to factors that linguistic features cannot fully detect.

As noticeable, the ranking of features in Table 4.7 is mostly different than the one obtained correlating feature values and the average complexity scores, and it is harder to notice patterns in the rankings. Considering the entire group of sentences (*All* in the Table), numerical information is still predominant for sentences in the central window position (*C_upos_NUM*, *C_dep_nummod*) and sentences in the end window position (*E_upos_NUM*, *E_dep_nummod*) as a factor that is positively correlated with standard deviation – meaning that numbers in these positions tend to produce very different complexity judgments by annotators, resulting in a lower agreement on the complexity. On the other side, most features with a negative ranking belong to the central sentence and are related to aspects of structural complexity as the chains of subordinates (*C_avg_sub_chain*), the presence of pronouns and auxiliaries (*C_upos_AUX*, *C_upos_NUM*) which may indicate a more articulate verbal structure, and factors linked to the length of the sentence and the presence of longer (*C_char_tok*). A higher value for these features corresponds to a lower standard deviation, namely a higher agreement.

When considering groups of sentences of the same length, very few correlations appear, the majority resulting for short sentences (*L10*, sentences that are 10 tokens long) and being mostly negative correlations with the features of the end window sentences. Most of these features are related to structural complexity as the distance of syntactic links (*E_avg_max_link*, *E_avg_link*, *E_max_link*), coordination phenomena (*E_dep_parataxis*), or aspects that regard the main verb and the auxiliaries (*E_verb_Pres*, *E_dep_aux*).

4.4.3 Prediction of sentence complexity judgments

The results of the correlation analysis have shown that linguistic information of the context affects the perception of sentence complexity and the extent to which this perception is shared by annotators. Given this, and similarly to what has been shown in Section 4.3.3, this part of the work assesses the contribution of the context from a modeling standpoint.

Two regression tasks were built, one to predict the average complexity value assigned to each sentence and one to predict the standard deviation of complexity for each sentence. In both scenarios, two different models were employed: the first is a linear Support Vector Regression model (SVR) (Section 3.2.2) with standard parameters that leverages the explicit linguistic features as described in Section 2.2.2, and the second is obtained by fine-tuning the BERT base model (i.e., bert-base-uncased, detailed in Section3.3) on the dataset using the FARM[5] regression implementation. Both models were evaluated with a 5-fold cross-validation for each of the three windows of context.

For every window, different runs of the models were carried out, varying the number of contextual features to be considered. For the *begin window* and the *end*

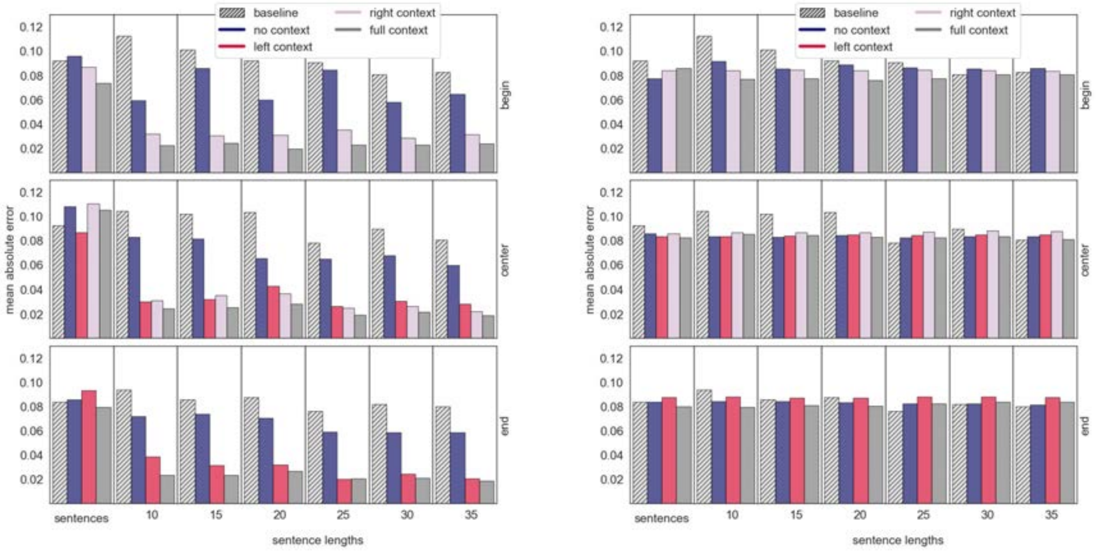
[5] `www.github.com/deepset-ai/FARM` (last visit 09/07/2023).

window models were run with *i*) the features of the single sentence (no context), *ii*) the features of the sentence + the features of the next sentence (right context) for the *begin window*, or + the features of the previous sentence (left context) for the *end window*, *iii*) the features of all the three sentences (full context, i.e. the whole window of context); for the *center window*, the models were trained with *i*) no context features, *ii*) left or right context features, *iii*) full context features.

The performance of the models was measured in terms of *mean absolute error* (MAE), evaluating their accuracy in predicting the same average judgment of complexity assigned by humans and the standard deviation of the complexity judgments. The same experiments were repeated, grouping the sentences according to their length. The baseline for the models' evaluation was calculated (i) in the case of all sentences by giving in input to the linear regression model only the length of the sentence as solely feature for the prediction, (ii) in the case of different lengths (binned sentences), by having the model always assigning the average complexity value (calculated on the whole set of sentences) to each sentence.

Figure 4.7 reports the results for the prediction of the average complexity, showing the average MAE obtained after the 5-fold validation, both for SVR and BERT models. The SVR models with linguistic features outperform BERT models overall. BERT models remain close to the baseline in all cases, despite the amount of context considered and the length of the sentences. Instead, the SVR models show significant differences as appropriate. In the case of all sentences, the model's performances are close to the baseline. Adding contextual features partially helps the model in the case of the begin and the end window, while performances worsen in the case of the center window. When considering sentences of the same length, the model's performance is always helped by the presence of contextual features, and the best results are achieved when the full context is taken into account for all the windows of context. This behavior confirms on one side that linguistic characteristics of the context are indeed very influential on complexity, and on the other side that the length of the sentence plays an important role in the perception of complexity, as it is only by binning the sentences that it is possible to exploit the effect of context in predicting complexity.

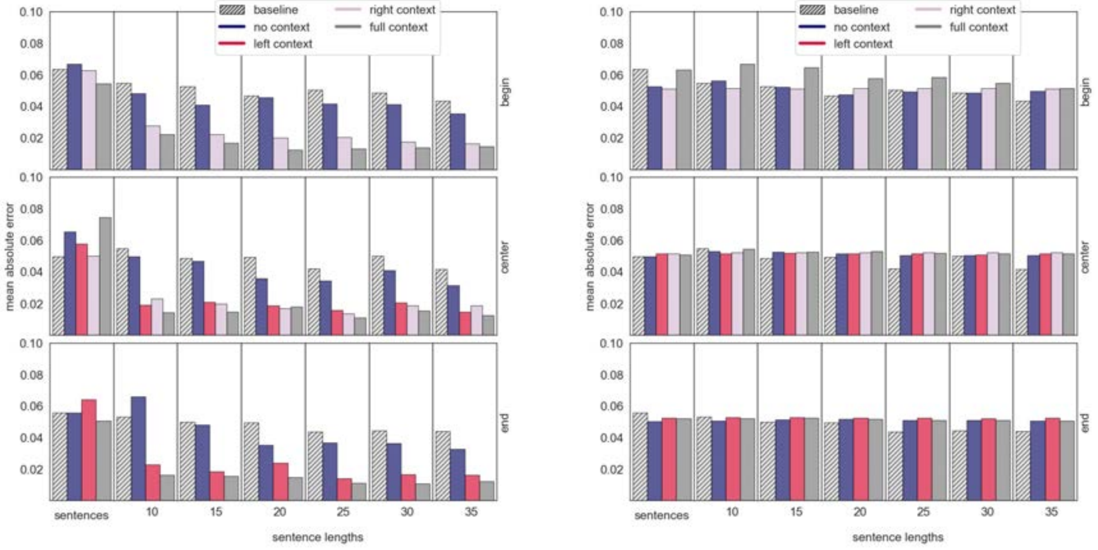
Figure 4.8 shows the results for the prediction of the standard deviation of complexity for SVR models and BERT models. As in the previous case, BERT models obtain results that are in line with the baseline and that are not influenced by different amounts of context. When looking at the results obtained with the explicit linguistic features, the outcome is quite different. For the all sentences case, the SVR model cannot predict the standard deviation of complexity, although the error gets lower for the begin window and the end window when the full context is used. Conversely, the model greatly improves when working on sentences of the same length. In all windows and for all lengths, using the features of the whole context significantly decreases the error in the prediction of standard deviation. When running the model with the features of the single



(a) SVR models

(b) BERT models

Figure 4.7: Performance (MAE) of Support Vector Machine Regression model on average complexity rating prediction. In different windows of context and with different context spans, for all sentences and at different sentence lengths.



(a) SVR models

(b) BERT models

Figure 4.8: Performance (MAE) of Support Vector Machine Regression models and BERT models in the prediction of complexity standard deviation. In different windows of context and with different context spans, for all sentences and at different sentence lengths.

sentence (i.e., no context), the model's performances are generally close to the ones of the baseline. This suggests that the context is particularly relevant in predicting how people will agree on their perception of complexity.

Overall, the results obtained here show that information about the complexity of a sentence is better encoded in its explicit linguistic features, thus its syntactic and morphosyntactic structures. On the other hand, although BERT has been proven to embed a wide range of linguistic properties, including syntactic ones [197, 271], the findings just discussed seem to suggest that this model does not exploit these kinds of features to solve a downstream task like the one here presented, for which few data are available. Indeed, it has been shown that BERT performs better on datasets larger than the one here employed [154]. Thus, it is fair to assume that more data may be needed for BERT to detect phenomena about perceived complexity.

Moreover, these results show that the presence of context plays an important role in complexity. As the SVR models are always helped by the contextual features, it is fair to assume that annotators have taken into account the whole context when expressing their judgment upon the complexity and that the presence of the context has strongly influenced their perception. Also, contextual linguistic phenomena are the ones that impact more on the variation of complexity perception between annotators as they are the ones that help more in the prediction of this variation.

Chapter 5

Analysis of Perceived Emotions in Response to Audiovisual Stimuli: a Study on Forrest Gump

"My mom always said life was like a box of chocolates. You never know what you're gonna get."

Forrest Gump

The previous case study was dedicated to how the complexity of syntactic and semantic levels in language can affect engagement in communication, creating emotions of boredom, frustration, or anger. While in the previous Chapter emotions are indirectly generated by possible perceived complexity, the case study detailed in this Chapter gives a closer look at how an external stimulus can actively elicit specific emotions in humans. In particular, this case study examines the emotive reaction that human beings have when exposed to an audio-visual stimulus that is meant to reproduce real-life content and context.

In the context of this study, the attention is centered on the film "Forrest Gump". Selected for its richness in life-like situations and its presentation of a wide range of emotions, this movie is an ideal candidate for fostering emotional resonance, i.e., the empathic responses in viewers. The movie's dialogue, teeming with emotion-related language, provides an avenue to examine the correlation between the explicit and implicit linguistic aspects and the emotions experienced by the observers.

The Chapter is organized as follows: Section 5.1 presents the background and motivations for this case study; Section 5.2 discusses the data used in the study, along with the composition of the final dataset and its statistics; Section 5.3 describes the method applied for predicting subjects' emotions from the text (i.e., the dialogue) extracted from the movie; finally, Section 5.4 presents the results obtained with the aforementioned approach and features.

5.1 Background and motivations

In psychological research, there is a well-established tradition of trying to elicit emotional states in laboratory settings for scientific purposes. Several procedures have been used to elicit emotions in the laboratory, including images [258, 276], sounds and music [268, 284, 283, 285], facial and body movements [90], drugs [296], relieved or imagined scenes [17, 136], and odors [49, 252]. During the last decades, a significant impact on the study of emotions was given by research on the affective processing of pictures, thanks also to the development of a few validated databases of pictures, such as the Pictures of Facial Affect (POFA) [83], the Karolinska Directed Emotional Faces (KDEF) [175], the Nencki Affective Picture System (NAPS) [186], or the International Affective Picture System [157], containing more than 1,000 exemplars of human experience and widely used for experimental research on emotions and attention.

The usage of affective images for emotion elicitation has been widespread in the past thank to the advantages they bring compared to other stimuli. Images are a class of stimuli with a clear, evocative ability and are easy to implement. They can be used in different experimental designs, from simple passive viewing to slide viewings mixed with more complex tasks. They are easy to edit, catalog, and are static: a desirable quality in some studies because dynamic changes may complicate the measure and interpretation of the affective response. Nevertheless, this last feature may also be interpreted as one of the limitations of pictures, as static image viewing does not yield an affective experience that is strong or ecological enough compared to the one prompted by a dynamic modality, such as video viewing [296].

Indeed, another effective and widespread method for emotion elicitation involves the use of movie clips. As emotion elicitors, movies have all the advantages of pictures, like their capability of being standardized, but they also offer the advantage of being dynamic and thus more similar to real life. Indeed, movie clips often represent a display of prototypical situations relevant to well-being and survival that make this stimulus high in ecological validity [114]. Movies also allow a multimodal stimulation of the viewer, simultaneously engaging both the visual and the auditory system. Throughout the years, movies have been used as elicitors to study all sorts of phenomena related to mood induction, showing they are suitable for eliciting basic emotions, such as fear or disgust, and also for the induction of more complex feelings [240, 248].

Movies can be particularly useful in studies regarding mood induction and emotion elicitation because they are also a source of intensive longitudinal data, i.e., data collected with repeated measurements or self-reports separated by relatively short time intervals [121]. Indeed, movies are rich in character dialogue, which offers the audience a direct window into what the character is feeling and thinking. The dialogue can be used as longitudinal text to gather information on emotion dynamics, i.e., the patterns of change and regularity in emotion [130, 155]. Movie dialogue is often rich in emotion-related words, e.g., if a character is angry, they are more likely to use anger-associated words [241]. The words and actions of the character then resonate with the viewer, whose emotions are elicited by an empathic reaction. Empathy is the ability to infer another person's feelings and share these feelings. Empathy is one of the fundamental social abilities that let humans interact among themselves, assuring successful communication and coordination of joint actions.

The work presented in this Chapter encourages empathic reactions and emotional contagion in a group of subjects, employing a naturalistic continuous stimulation paradigm – the movie *Forrest Gump* – and studies how the language presented in the movie's dialogue is related to the emotions experienced by the observers. This movie is rich in life-like situations and depicts a wide range of emotions that can be used to nurture emotional resonance [114, 223, 248], i.e., to foster emphatic responses in the observers, which are influenced by the narrative choices presented in the movie. Previous work confirmed the usefulness of this movie in the context of emotion elicitation, showing that, in a 60-second window, the emotions presented in the movie resemble the ones experienced in real life [163].

5.2 Data

The primary data source for this case study is *StudyForrest*[1], a research project centered around the use of the movie *Forrest Gump*. The project is built upon the contributions of multiple research groups and has the purpose of studying the human brain and the way it processes the vast amount of diverse information gathered by the senses. Most studies that tackle brain activity present subjects with simplified stimuli in a controlled environment, a setting that does not resemble the complex data that the brain collects in a natural environment. Indeed, a natural input is necessary for a deep understanding of brain functioning. The choice of the movie Forrest Gump is designed to provide for a complex sensory input that reproduces real-life-like content and contexts.

This open-source project includes data related to (*i*) brain structure and connectivity, (*ii*) behavior and brain function, and (*iii)* movie stimulus annotations. In (*i*) are arranged data that characterize the participant's brain function on various dimensions. A diverse set of stimulation paradigms (audio, audio-visual) and data acquisition setups were used (fMRI, physiological recordings of heartbeat

[1] `www.studyforrest.org` (last visit 09/07/2023).

and breathing). In (*ii*) are organized sets of structural brain images (MRI, angiography) that provide an in-vivo assessment of the participant's brain hardware. Lastly, (*iii*) gathers annotations of the content of the Forrest Gump movie, which is rich in visual and auditory features, but also facets of social communication (portrayed emotions, body contact, eye movements).

For this work, the focus was on two assets of data: the speech present in the movie (the dialogue pronounced by the movie characters) and the range of emotions elicited in a group of subjects when watching the movie. Hereafter the composition of the data is discussed, along with the dataset creation process.

5.2.1 Textual Data

The movie stimulus annotations section of the project provides a written transcription of the film dialogues and voice-overs, i.e., moments in which Forrest Gump acts as a central narrator and tells the story from his own perspective.

The data is composed of 985 rows, each representing a line of dialogue of a character. Each row is associated with the name of the character pronouncing the line and two timestamps t_{begin} and t_{end} with a *seconds.milliseconds* format. The former timestamp states when the character starts pronouncing a specific line, while the latter states the moment in which the character stops pronouncing the same line. An example of how the dataset is structured can be seen in Table 5.1, where the first five rows and the last five rows of the data are represented.

Among the totality of the rows, most lines are pronounced by the main character, Forrest: 224 lines are pronounced during the story, and 197 lines are pronounced as a voice-over.

5.2.2 Emotional Data

The collection of emotional data was carried out by the Molecular Mind Laboratory (MoMiLab), a research group of the IMT School for Advanced Studies of Lucca, Italy. Extensive details about the group's contributions to the StudyForrest project can be found in [163], along with information about emotional data acquisition procedures and analyses.

A group of 12 healthy subjects was selected (5 females, 7 males; mean age 26.6, age range 24-34), making sure none of the subjects had watched the movie in the year preceding the data acquisition. The subjects were asked to watch an edited version of the movie, cut into eight segments with a duration ranging from 11 to 18 minutes. The subjects were instructed to continuously rate the subjectively perceived intensity (scale $0 - 100$) of the emotions they were experiencing while watching, choosing among six basic emotions [84] (see also Section 1.1): happiness, surprise, fear, sadness, anger, and disgust. To do so, they pressed specific buttons on a keyboard, through which they could increase or decrease

row	start	end	character	text
1	189.8	194.0	FORREST	Hello. My name's Forrest Gump.
2	199.7	201.2	FORREST	You want a chocolate?
3	204.2	217.4	FORREST	I could eat about a million and a half of these. My momma always said, "Life was like a box of chocolates. You never know what you're gonna get."
4	225.2	237.8	FORREST	Those must be comfortable shoes. I'll bet you could walk all day in shoes like that and not feel a thing. I wish I had shoes like that.
5	238.6	240.0	BLACK WOMAN	My feet hurt.
...	...	...	...	...
981	6978.1	6985.5	FORREST	Hey, Forrest. Don't... I wanted to tell you I love you.
982	6986.3	6987.2	FORREST JR.	I love you too, Daddy.
983	6989.3	6991.0	FORREST	I'll be right here when you get back.
984	6996.3	6999.4	SCHOOL BUS DRIVER	You understand this is the bus to school now, don't you?
985	6999.4	7002.6	FORREST JR.	Of course, and you're Dorothy Harris and I'm Forrest Gump.

Table 5.1: Head data and tail data of the written transcriptions of dialogues from the movie "Forrest Gump".

the intensity of the emotions they were experiencing. Subjects were also allowed to report more than one emotion simultaneously.

5.2.3 Dataset creation

The ratings for emotional data were collected from a continuous output $z = (z_1, z_2, ..., z_n)$ from the keyboard, such that each z_i with $i \in [1, n]$ corresponds to an increment of 0.1 seconds (10Hz frequency) in the playing time of the movie ($z_i = 0.1$, $z_{i+1} = 0.2$, $z_{i+2} = 0.3$, ...). Each z_i is associated to a list $x_{i1}, x_{i2}, ..., x_{ij}$, with $x_j \in [0, 100]$ and $j \in$ [*happiness, surprise, fear, sadness, anger, disgust*], where each x_j indicates the intensity that one emotion assumes at a given timestamp. This results in a time series for each subject that collects the moment-by-moment perceived emotions and their intensity.

Once textual and emotional data were collected, they underwent some processing and manipulation to be temporally aligned. The first necessary step was a resampling of the emotional ratings. Indeed, the 10Hz frequency used for the collection of emotions is too detailed for the purpose of this study. The emotional time series were downsampled from 0.1 seconds to 2 seconds. New timestamps $s = (s_1, s_2, ..., s_m)$ were generated, such that each s_i corresponds to the sum of

20 consecutive z_i, thus to an increment of 2 seconds in the playing time of the movie. Each s_i is associated with a new list of emotional values, where each new value is the average of the values associated with the summed z_i.

After resampling, the text was aligned to the emotional data. In this alignment phase, two elements were taken into consideration to make sure the alignment was correct:

1. the emotive response and its consequent registration through the keyboard are not simultaneous to the movie's events. It may take a few moments for the subject to process their feelings and press the buttons on the keyboard;

2. the interest of this work is to understand how textual features (i.e., the linguistic features of the lines pronounced by the characters) are related to the emotive response. Given that the emotive response is delayed, it is impossible to know for sure if the emotion was caused by the text the subject listened to immediately before the emotion declaration or by the previous text.

The statements described in the two points above and, in particular, the one described in point 2, were tackled by aligning different amounts of text to a single timestamp. For each timestamp s_k in the data, three progressively larger time windows are considered, such that $window_i = [s_k - m, s_k]$, where $m = (2, 4, 6)$. For each sentence, its t_{end} is retrieved, and the sentence is aligned to the timestamp verifying if $s_k - m \leq t_{end} \leq s_k$, thus checking if the moment in which the sentence ends falls within the given time window. In this way, the larger the time window, the larger the amount of text that gets aligned with a specific timestamp. With this process, three different datasets are created, one for each time window. After, all the lines in which no text was aligned to s_k are removed. As a result, each dataset is composed of 898 timestamps associated with a line of text and six emotion declarations for each of the 12 subjects.

5.2.4 Data statistics and data selection

After aligning the data, their statistical distribution was analyzed. First, it was examined how many times each subject declared a specific emotion. Given a timestamp, if the subject assigned to an emotion a value different than zero, that emotion was considered present for that timestamp, regardless of the intensity declared. Along with the six categories of emotions existing in the dataset, two more categories were created, *emotion* and *neutral* to account for two specific cases:

1. if all six emotions were zero at the same time (all $x_j = 0$), this case was given the class *neutral*;

Subject	Happiness	Surprise	Fear	Sadness	Anger	Disgust	Neutral	Emotion
1	592	172	101	557	111	166	22	876
2	628	87	83	539	120	42	61	837
3	345	471	212	340	123	37	30	868
4	274	179	137	255	119	133	276	622
5	244	84	98	224	83	6	305	593
6	496	92	147	264	60	13	113	785
7	277	255	88	132	88	23	286	612
8	357	218	119	305	103	77	231	667
9	299	389	15	147	109	22	312	586
10	213	125	81	255	60	0	377	521
11	352	320	116	307	150	30	120	778
12	180	36	22	149	34	25	526	372
Total	4257	2428	1219	3474	1160	574	2659	8117

Table 5.2: Emotions distribution in the dataset.

2. if any emotion, among the six considered, was declared (at least one $x_j \neq 0$), this case was given the class *emotion*.

These two classes are meant to address the moments of the movie in which the subject did not allegedly perceive any emotion that they felt like declaring and the case of a generic emotional response without accounting for the specific perceived emotion.

Table 5.2 reports the distribution of the 6 basic emotions and the new *neutral* and *emotion* categories for all subjects, i.e., showing for how many timestamps across the movie that category of emotion was declared. It is necessary to remark that the category *emotion* is not the exact sum of all basic emotions, as there are cases in which two or more emotions were declared simultaneously.

What emerges from the statistics reported in Table 5.2 is that the distribution of the basic emotions in the dataset and among the subjects is quite uneven. Overall, the most represented emotions are happiness and sadness, while the least represented one is disgust. Furthermore, it appears clear that every subject had a different emotional experience while watching the movie. Some subjects declared all emotions evenly and frequently (e.g., subject 4, subject 8), while others declared emotions less frequently (e.g., subject 10, subject 12). This discrepancy happens primarily because emotive phenomena are intensely subjective, meaning that emotion processing is specific to each person and everyone experiences emotions at a different granularity [21].

To account for the inter-subject different emotional experiences, the level of agreement between the 12 subjects was measured using *Fleiss' Kappa* (Section 3.1.2). Table 5.3 reports the percentage of agreement for each basic emotion in the data. The lowest agreement is found on surprise and disgust, meaning the declaration of these emotions was not consistent across the subjects. This result

Emotion	**Agreement**
happiness	0.32
surprise	0.14
fear	0.41
sadness	0.31
anger	0.42
disgust	0.17

Table 5.3: Annotators agreement (Fleiss' Kappa) on all emotions

may seem in opposition to the supposed universalism of basic emotions [196], as the inconsistency and the low agreement indicate that participants did not experience surprise and disgust at the same points of the movie. However, the movie stimulus used in this study is not conceived to reflect the definition of six basic emotions. For instance, in [163] it is highlighted that some of the participants reported as disgusting some movie scenes that required further interpretation of the context of what was happening (e.g., the school's principal using his power to get sexual favors) rather than showing something that was repulsive. This interpretation of disgust was not present in all subjects, with some of them relying on the classical well-established definition of disgust. Consequently, they may have rated with *disgust* only the scenes they truly perceived as repulsive. This shows that even though research affirms the existence of six basic emotions, psychological constructs and cognitive interpretations need to be taken into account.

The most robust agreement is found on fear and anger, showing that these emotions are evoked in specific scenes of the movie and that subjects had a similar emotional response to those scenes. The agreement on happiness and sadness is slightly lower; nevertheless, this result is still significant, given these are the most declared emotions in the dataset, indicating what the general mood of the movie is.

Table 5.4 reports examples of sentences on which the subjects agreed the most for all six emotions. For every emotion, there are many sentences on which a large number of subjects agreed, meaning that there were various moments of the movie that elicited the same emotions in the subjects. In the case of disgust, the highest level of agreement was achieved in 8 subjects, only in one sentence. There were no other sentences for which 8 subjects (or more) agreed. This is justified by the fact that disgust is the least represented emotion in the data.

Given the results obtained from the agreement assessment and the distribution of emotional ratings, the analysis of underrepresented emotions was not addressed directly, even though the agreement of the subjects was high (as for fear and anger). In order to account for underrepresented emotions and not lose this information completely, the analyses relied on the general class *emotion*. Hence

Emotion	N subjs	Text
happiness	12	I had never seen anything so beautiful in my life. She was like an angel.
surprise	11	Jenny! Forrest!
fear	12	(into radio) Ah, Jesus! My unit is down hard and hurting! 6 pulling back to the blue line, Leg Lima 6 out! Pull back! Pull back!
sadness	12	Bubba was my best good friend. And even I know that ain't something you can find just around the corner. Bubba was gonna be a shrimpin' Boat captain, But instead he died right there by that river in Vietnam.
anger	12	Are you retarded, Or just plain stupid? Look, I'm Forrest Gump.
disgust	8	You don't say much, do you?

Table 5.4: Examples of sentences on which subjects agreed the most, for all emotions.

three different scenarios to examine were selected:

1. the presence of any kind of emotion (at least one $x_j \neq 0$),
2. the presence of happiness ($x_{happiness} \neq 0$),
3. the presence of sadness ($x_{sadness} \neq 0$).

Furthermore, the experiments were conducted only on two subjects, subject 4 and subject 8. These two subjects were chosen because they declared all emotions evenly, without neglecting any of them, and because the number of declarations for each emotion was quite similar between the two.

5.3 Emotions prediction from text

The three scenarios described in Section 5.2.4 were evaluated in contrast to the absence of any emotion (all $x_j = 0$), producing three binary classification tasks. The experiments rely on the use of automatically extracted explicit linguistic and lexical features described in Section 2.2.2, and implicit linguistic features, i.e., contextual word embeddings from a language model described in Section 3.3. In each task, the linguistic features, either explicit or implicit, are used as input to predict one of the binary options.

5.3.1 Prediction with explicit linguistic features

For the first set of features, the sentences were first POS tagged and parsed using UDPipe [263]. After, the set of explicit linguistic features was automatically extracted (hereafter *linguistic* features). For this task, some additional explicit features were added to the previous ones: namely features that can capture some lexical information (hereafter *lexical* features), as they identify sets of characters or words that appear more frequently within a sentence, i.e., bigrams, trigrams, and quadrigrams of characters, words, and lemmas.

Two SVM Classifier models (see Section 3.2.1) were trained, one on the linguistic features (hereafter called *SVMling*) and one on the lexical features (hereafter called *SVMlex*). The models were trained with a linear kernel and standard parameters (as per ScikitLearn[2] configurations), performing 10-cross-fold validation to evaluate the accuracy of the models.

5.3.2 Prediction with implicit linguistic features

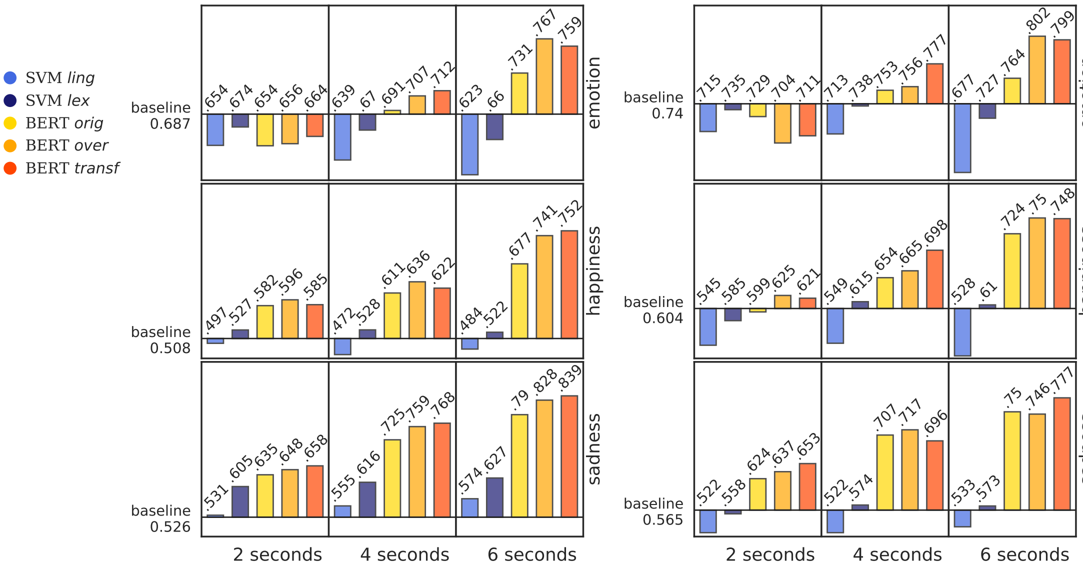
For the second set of features, the pre-trained BERT base model was retrieved and fine-tuned on the data of this study. The pre-trained BERT model already includes a lot of information about the language, as it has already been trained on a large amount of data. By fine-tuning the pre-trained model on the data of this study, it is possible to exploit the information already acquired by the model to solve the task of predicting emotions from the screenplay text.

Three different fine-tuning stages were performed to verify if the BERT pre-trained model is *di per se* sufficient to obtain accurate predictions or if it may benefit from some additional information. Specifically, the model was fine-tuned on:

1. the original data of this study as they have been discussed so far, without any kind of adjustment (from now on, this setting and the model derived from it will be referenced as *BERTorig*);
2. the original data with oversampling of the sentences that were assigned the neutral class to reach the same number of neutral entries as emotional entries (hereafter *BERTover*);
3. a transfer learning tuning step, followed by another tuning on the oversampled data (hereafter *BERTtransf).*

Transfer learning is a common machine learning technique in which a model trained on one task is re-purposed on a second related task. In this way, what was learned in the first task can be used to speed up the learning in the second task and grant greater generalization and better results. Applying a BERT pre-trained model in this study is already considered transfer learning, as the model

[2] `https://scikit-learn.org/` (last visit 10/07/2023).



(a) Subject 4

(b) Subject 8

Figure 5.1: Performance (accuracy) of SVM and BERT models in the prediction of emotion, happiness, and sadness, for every timespan window, and for both subject 4 and subject 8.

is trained on other data and for a different task. However, given that the tasks BERT is built upon are far from the one of this study, it is possible that the model may benefit from an additional round of tuning on an emotion-related task.

In the case of the transfer learning tuning described in point 3 above, the BERT pre-trained model underwent a first fine-tuning step on data different from the ones of this study but still conceived for a similar task. Notably, this step relied on data created for SemEval-2018 Task 1E-c [203], containing tweets annotated with 11 emotion classes.

After the tuning stages, the so fine-tuned models were used to perform the binary classification tasks on this study's data. The model accuracy was evaluated with 10-fold cross-validation.

5.4 Results and discussion

Figure 5.1 shows the accuracy scores for all the models, for both subjects and the three datasets. In all cases, the baseline was determined with a majority classifier, i.e., a classifier that always returns the most frequent label in the dataset. The trends in the results appear similar for both subjects.

SVM Classifier models are the worst-performing ones in all scenarios. In any case, *SVMling* is the model that gave the lowest performance, remaining below or around the baseline value. On the contrary, *SVMlex* tends to bring higher performance, despite remaining close to the baseline in most cases. The low scoring obtained by *SVMling* is due to the fact that features that look at the raw, morpho-syntactic, and syntactic aspects of text do not encode any relevant information regarding the emotional cues in the text. The syntactic aspects of language are not the main concurrent in the setting of this study. Nevertheless, in other scenarios, they can impact a person's emotionality. It is the case, for instance, when there is a complex syntactic structure in a sentence that makes the sentence itself less comprehensible, possibly causing feelings of boredom or frustration in an individual (as shown in the case study of Chapter 4). Indeed, in watching the movie, the subjects receive cues other than strictly syntactic ones (e.g., audio-visual) that may mitigate the effects of complex syntax if present or that convey better the emotional message. *SVMlex* always performs better than *SVMling* because the selected lexical features look at patterns of words and characters that are repeated in the input text and thus record information about the lexicon of the dataset. However, as this study's dataset is too small, it is hard for the model to retrieve the same lexical patterns in both the training and test set and gain points in its performance.

In all the datasets, BERT models outperform the SVM ones in both the prediction of happiness and sadness. In the case the prediction of the category emotion, BERT models are capable of excellent predictions only on the 6 seconds dataset. This last result can be explained by the fact that in the case of the prediction of the class emotion, all emotions are flattened into a single category. Thus, it may be difficult for the model to distinguish between general emotionally charged sentences and those not perceived as emotionally charged. When emotions are specific and separated, as in happiness and sadness cases, BERT can infer the perceived emotions even from small amounts of text (2 seconds and 4 seconds datasets).

BERTover and *BERTtransf* generally give better performances than what happens with *BERTorig*. This is especially true when a larger amount of text is used as input. The gain in accuracy when using *BERTover* when predicting the happiness and sadness classes is not outstanding if compared to *BERTorig* accuracy on the same classes. This result is not surprising because in these two cases, the classes to be predicted were already distributed quite evenly. Thus, an oversampling of the neutral class does not bring much additional information to the model. On the contrary, in the case of the prediction of the emotion class, the model achieves higher gains because it is helped by the higher representation of the neutral class.

With *BERTtransf*, the performances stay in line with the ones obtained with the bare oversampling. Moreover, in the prediction of the *emotion* class, the model performs worse than *BERTover* with the 6 seconds dataset. As SemEval data were too distant from the ones of the dataset of this study, fine-tuning the model on those data did not add any more helpful information. Therefore, even

	subject 4			subject 8		
	2 sec	4 sec	6 sec	2 sec	4 sec	6 sec
emotion	82.75	83.63	85.82	82.03	87.8	90.44
happiness	70.77	72.64	79.78	76.26	72.31	79.67
sadness	82.53	85.93	87.47	80.44	79.45	85.05

Table 5.5: Agreement (%) between *BERTover* and *BERTtransf* predictions.

though the SemEval task is similar to the one described here, the input text is too different from this study's sentences to contribute to the prediction substantially. Another form of transfer learning was also attempted by tuning the model on one subject and testing it on another one. However, the results obtained with this technique were not satisfactory and are not reported in this Chapter. This outcome is probably because emotion perception is a very personal phenomenon and it cannot be generalized from one individual to another one.

A further evaluation of the results of the models was realized by computing the percentage of agreement between the two best-performing models, *BERTover* and *BERTtransf*. The agreement was defined as the percentage of sentences for which the models fave the same output during the classification task. Table 5.5 reports the results for emotion, happiness, and sadness for every timespan window and both subject 4 and subject 8.

The agreement is quite high in all cases and it tends to get stronger with the amount of text on which models are trained (i.e., 6 seconds). A higher level in the agreement indicates that the models have similar behavior, thus making the same mistakes in the classification task. The lowest levels of agreement are encountered on the classification of happiness, showing that the two models work differently in this part of the task. Indeed, both *BERTover* and *BERTtransf* obtain high performances in predicting happiness, but the fact that their agreement is lower suggests that they differ in the mistakes they make in the classification. This information may be exploited to create systems that combine different classifiers, actually enhancing the classification accuracy. By doing this, it is possible to compare the cases in which two or more classifiers agree and the cases in which they make mistakes, thus choosing the best classification output accordingly.

The findings of this cause study indicate that specific perceived emotions can be accurately predicted using contextual embeddings derived from the dialogues uttered by the characters in a movie, even when the amount of text provided to the prediction model is minimal. However, when predicting general emotional elicitation (i.e., without targeting a specific emotion), predictive models necessitate a more substantial amount of text to yield accurate results.

This case study further revealed that the lexical, morpho-syntactic, and syntactic aspects of sentences are not effective predictors of the emotional responses experienced by viewers of a movie. This is primarily due to the fact that these features encapsulate minimal, if any, information about the emotional state and sentiment conveyed in the sentence. Conversely, contextual embeddings, which are aware of the word placement within a sentence, are capable of capturing this information as they encapsulate more semantic information about the sentence itself. Even though the stimuli coming from a movie are multiple (i.e., images, speech, music, contextual information), this study shows that the dialogue of the characters alone is already a good proxy for studying emotional elicitation and perception.

Chapter 6

Analysis of Bodily Response to Emotive Text

> *"Words mean more than what is set down on paper. It takes the human voice to infuse them with shades of deeper meaning."*
>
> *Maya Angelou*

The previous Chapters have shown the relationship between the linguistic structure of a text and the perceived level of complexity in a group of subjects, with complexity having an influence on the engagement and the emotions elicited during communication (Chapter 4); the relationship between the speech of a multimodal ecological stimulus and the emotions elicited in a group of subjects, showing how it is possible to use the transcription of the speech to predict some of the basic emotions experienced by the subjects (Chapter 5). While in the previous Chapters text was used to actively induce emotive responses into subjects, the case study of this Chapter employs emotively encoded texts to study how they influence bodily response in a group of subjects.

In particular, this case study focuses on the relationship between the linguistic profile of text and the acoustic properties and electrodermal activity of the readers. The structure of spoken language, which includes both semantic and syntactic components, is a crucial determinant of speech prosody, and by extensions, it substantially influences emotional expression. By exploring how this linguistic construction impacts speech prosody and physiological responses, such as those of the Autonomic Nervous System, it is possible to unravel the complex interplay between language, emotions, and communication.

The Chapter is structured as follows: Section 6.1 describes in depth the background and motivations for this case study; Section 6.2 describes the data, the experimental protocol, and the procedures for the extraction of the features used in the study; Section 6.3 details the different analyses applied on the data; finally, Section 6.4 delves into a discussion of the results obtained.

6.1 Background and motivations

Emotions significantly influence human spoken communication, having a considerable impact on the efficiency of speaking and reading tasks. In the past, evaluating emotions heavily depended on individual self-reported methods, which may be subject to personal biases and distinct personality traits. Nevertheless, modern technology has made it possible to incorporate speech prosody and autonomic nervous system (ANS) correlates, offering objective and reliable techniques to determine emotional conditions.

The ANS lays the physiological groundwork for emotional regulation, as it regulates bodily functions and is crucial in triggering emotional reactions [45]. Electrodermal activity (EDA) is one of the most thoroughly researched ANS correlates of emotional arousal. It measures variations in the skin's electrical conductance due to sweat gland activity, which is controlled by the sympathetic branch of the ANS. This provides objective assessments of emotional states, supplying invaluable information about the physiological expressions of emotions.

The intricate procedure involving ANS and somatic regulation also governs the production of speech. Human speech results from fine control of up to eighty muscles from respiratory, laryngeal, pharyngeal, palatal, and orofacial groups [72]. Such control is a complex process involving activity in the somatic and autonomic nervous systems (ANS). Alterations in the respiratory activity induced by the ANS manifest changes in the emotional state of the speaker by influencing the voice spectrum characteristics such as the fundamental frequency (F_0 - the frequency of vibration of the vocal folds), and its formants (F_1, F_2, F_3 - resonance frequencies of the vocal tract) [311] (also see Section 2.3). Hence, the analysis of speech prosody offers a crucial understanding of the emotional context and intentions concealed within spoken communication [131].

Previous studies have used various analytic methods to measure changes in fundamental frequency, loudness, speech rate, and pause in order to effectively characterize affective prosody [93] and to explore several psychological dimensions of the speaker: emotion [149], mood [66], stress [106, 98], and personality [116]. Nevertheless, inferring a speaker's emotional state from these features remains a challenging task.

However, the linguistic framework of a spoken text, encompassing both syntactic and semantic elements, significantly impacts speech prosody and, as a result, emotional expression. Therefore, an examination of how the linguistic structure of spoken text influences speech prosody and ANS correlates offers a meaningful approach to understanding the interplay between language, emotions, and communication. Specifically, investigating the effect of the linguistic structure of spoken text on speech prosody and ANS correlates, like EDA, uncovers the dynamics of how language and emotions interrelate. This understanding carries significant implications across various practical uses. For example, in the realm of human-computer interaction, discerning how linguistic signals influence emotional reactions can guide the creation of emotionally perceptive systems that adjust to the affective states of users. In healthcare scenarios, observing the emotional reactions of patients during speech or reading tasks could support the identification and management of emotional disorders.

In an effort to capture the various dimensions of information embedded in a text, including language, lexicon, and style, increasingly advanced Natural Language Processing (NLP) and machine learning methods have been conceived. The progress in these areas has led to the establishment of sophisticated techniques enabling the depiction of a text's linguistic profile by extracting a vast number of features that model underlying lexical, grammatical, and semantic phenomena [42].

Linguistic profiling (see Section 2.2.2) has been applied in various contexts, such as automatically classifying textual genres and registers [10] and modeling cognitive aspects of human language. For example, in [43], the authors have shown that linguistic features that capture lexical and (morpho-)syntactic properties of a sentence can be effectively used to predict the perception of its complexity by humans. This evidence has been further confirmed by a subsequent study [133], which also proved the reliability of linguistic features extracted from context in predicting human judgments of sentence complexity. A recent work [255], proposed a deep learning hierarchy for emotion recognition, combining text analysis computed by the language model ELMo [222] with prosody, voice quality, and spectral features. However, formal modeling of the relationship between features describing linguistic profiles, ANS response, and speech prosody could provide insights into the specific mechanisms that influence a speaker's emotional response.

The aim of this case study is to investigate the correlation between the linguistic structure of a text and the physiological and acoustic features commonly used to assess the activity of the autonomic nervous system (ANS) and speech production prosody. Participants were asked to read texts that were designed to elicit varying levels of emotional arousal and valence. Electrodermal activity (EDA) was analyzed as a widely used correlate of the sympathetic nervous system (SNS) to quantify the sympathetic reaction. Correlation and regression methods were applied to analyze the relationship between EDA-related features, speech prosodic and linguistic profiles of the texts to understand the extent to which the linguistic structure of a text interacts with the speech production and sympathetic response elicited by the same texts.

In addition, a complementary analysis was conducted to evaluate the strength of the relationship between the linguistic structure of a text and the physiological and acoustic features of speech production, but from the opposite perspective. Specifically, this analysis aimed to test the feasibility of using speech and physiological signals to predict a set of features that characterize the linguistic structure of the pronounced text. This approach aligns with recent efforts to use cognitive signals to improve the performance of natural language processing models in multi-modal settings, and to provide more cognitive-oriented benchmarks for their evaluation. While most of these studies have utilized eye-tracking data, which have been shown to be effective in various sequence labeling and sequence-to-sequence scenarios such as sentiment analysis, irony detection, Part-of-Speech tagging, Named Entity Recognition and relation extraction [129]. On the other hand, other sources of physiological data, such as ANS correlates, still require further investigation.

6.2 Data

This case study employs four texts, chosen to represent different levels of arousal and valence. The texts were chosen by following to the Circumplex Model of Affect [233] (refer to Section 1.1.2), which defines arousal as the intensity of perception and valence as the pleasantness or unpleasantness of the stimulus (the text, in this case). Two of the selected texts provide a detailed and graphic description of medieval torture practices and are classified as high arousal and negative valence (hereafter also referred to as *affective* or *emotive* texts). The two other text, classified as neutral, discuss text types and writing styles.

Before conducting any experiment, a group of 22 subjects (other than those who later participated in the study) evaluated the texts based on their levels of arousal and valence. Their evaluations confirmed the predetermined levels of arousal and valence based on the topic of the text. In particular, the texts were assigned the following valence and arousal rates:

- Neutral text 1 – Valence Rate: 0.30 ± 0.52; Arousal Rate: 1.31 ± 0.67
- Neutral text 2 – Valence Rate: 0.25 ± 0.55; Arousal Rate: 1.39 ± 0.77
- Emotive text 1 – Valence Rate: -1.31 ± 0.79; Arousal Rate: 3.26 ± 1.17
- Emotive text 2 – Valence Rate: -1.19 ± 0.84; Arousal Rate: 3.24 ± 1.09

The texts were also chosen to be of similar length, so that the reading task implemented for the study would have a similar duration for each of the subject.

The full texts assigned to the readers are reported in Appendix C, along with some statistics. As the texts chosen for the study are in Italian, the Appendix also contains their translation in English.

6.2.1 Subjects recruitment, experimental protocol and acquisition set-up

The study recruited 33 healthy individuals (17 females and 16 males) with an age range between 26.6 and 30.0. None of the participants had prior history of cardiovascular diseases, mental disorders, or phobias. All subjects provided written informed consent to participate in the study, which was approved by the Ethical Committee of the University of Pisa.

The experiment was divided into different sessions that were randomly assigned to the participants. Each session lasted approximately 2 minutes for each subject, followed by 40 seconds of rest and 1 minute of recovery. The study setup is described in more detail in [103]. During the sessions, each participant read aloud one *neutral* text and one *affective* text, chosen randomly from the available texts. The duration of the reading is consistent among the subjects as the texts are of similar lengths.

As the subjects read the texts, their speech signal and electrodermal activity were recorded. Following the reading task, each subject was asked to rate the texts they read on a scale of -2 to 2 for valence and 1 to 5 for arousal using the Self-Assessment Manikin (SAM) model.

6.2.2 Linguistic analysis

The texts were divided into sentences using full stops as a delimiter criterion. The resulting number of sentences in the neutral texts was 25, with an average length of 28 tokens. Affective texts had 40 sentences with an average of 21 tokens.

Each sentence was analyzed linguistically and represented as a vector of approximately 140 features, using the explicit linguistic features described in 2.2.2 to define a linguistic profile for each of the text of the study.

6.2.3 Speech signal processing

The speech time series recorded during the task were analyzed using the BioVoice toolbox [205] (see Section 2.3.2). Specifically, the F_0, F_1, F_2, and F_3 parameters described in Section 2.3 were calculated from the speech data.

First, the toolbox detected the voiced parts of each segment. Then, it calculated F_0, F_1, F_2, and F_3. For each voiced frame, F_0 is estimated with a two-step procedure:

1. Simple Inverse Filter Tracking (SIFT) is applied.

2. F_0 is adaptively estimated on signal frames of variable length inversely proportional to F_0. The estimation is done through the Average Magnitude Difference Function (AMDF) within the range provided by the SIFT [181].

The formants values over time were extracted by considering the Autoregressive Power Spectral Density (AR PSD). In addition to F_0 and the formants, the following features were extracted from each sentence:

- Signal Duration: the total time duration of reading, i.e., the duration in seconds of the time necessary to read the sentence, including pauses;
- Voiced Duration: the duration in seconds of the vocal emissions, excluding pauses;
- Mean Duration: the average voiced duration.

To account for the subject-dependency, the frequency features (F_0, F_1, F_2, and F_3) were scaled using the following formula: $F_i^{scaled} = F_i/\overline{F_{0neu}}$ where F_i represents the frequency feature of interest (in neutral or emotional test in each sentence) and $\overline{F_{0neu}}$ is the mean of the frequency of the corresponding neutral texts, computed for all time duration.

6.2.4 Electrodermal Activity signal processing

The cvxEDA algorithm [111] (detailed in Section 2.4.1) was used to decompose the Electrodermal Activity signal into its phasic and tonic components, as described in Section 2.4. After the decomposition process, the features described in Section 2.4 are extracted within the time window corresponding to each sentence, namely: the mean (*mean ph, mean ton*), standard deviation (*std ph, std ton*), and maximum value (*max ph, max ton*) of both components; the number of phasic peaks (*no pks*) and the sum of their amplitudes (*sum pks*); the power spectrum within the 0.045-0.25Hz interval (*edaSymp*), which reflect the sympathetic activity [232]. The features are then normalized according to the time window length.

6.3 Statistical analysis and modeling of the features

This case study aim to understand the relationship between the linguistic features of emotionally encoded texts and the acoustic and electrodermal responses generated during the reading of the texts. To this aim, different form of analysis and feature modeling were implemented.

The first statistical analysis of the study examines the relationship between (*i*) the explicit linguistic profile of the texts and (*ii*) speech and electrodermal activity features by implementing a correlation task, as detailed in Section 6.3.2.

A second statistical analysis aimed to examine the relationship between the linguistic features and the speech and Electrodermal Activity features from a modeling perspective. To achieve this, two complementary tasks were designed. The first scenario tackles the effectiveness of the linguistic features in the prediction of the speech and electrodermal activity ones (Section 6.3.3), while the second scenario tackles the effectiveness of the speech and electrodermal features in the prediction of the linguistic features (Section 6.3.4).

6.3.1 Self-Assessment Manikin statistical analysis

The Self-Assessment Manikin valence and arousal scores assigned by the participants at the end of each reading task were statistically compared between the a-priori neutral and negative texts using a Wilcoxon signed-rank test (see Section 3.1.3).

The Wilcoxon test confirmed that there were significant differences between the a-priori negative and neutral texts used in the experiments. Specifically, the scores for arousal and valence were significantly different between the two types of texts. The results showed that after reading the negative texts, the arousal score was significantly higher ($p < 0.01$), and the valence score was significantly lower ($p < 0.01$) as compared to the neutral texts.

6.3.2 Correlation analysis

This analysis aimed to identify which linguistic properties of the texts were most related to the subjects' physiological arousal and speech production. A study of these factors permits to discover any underlying interaction between the linguistic structure and profile of a text and the dynamics of the Sympathetic Nervous system and of speech.

To investigate this relationship, a correlation analysis task was set up. Each linguistic feature was correlated with every electrodermal activity feature and each speech feature using Spearman's correlation coefficient (Section 3.1.1) as the evaluation metric. False discovery rate correction was applied to account for multiple hypothesis testing [261]. Any correlations that were statistically significant (with a p-value < 0.05) and had a correlation coefficient different from zero were considered. The percentage of subjects for whom the pairwise correlation was significant was calculated for each feature, which allows to determine whether certain patterns were more stable across participants and to understand which phenomena they involve.

Table 6.1 and Table 6.2 provide an overview of the most significant results of the correlations between speech features and linguistic features, and between electrodermal activity features and linguistic features, respectively. The complete results, including the mean correlation values, can be found in Appendix C. In both tables, the linguistic features are grouped according to the phenomenon they describe. The cells in the tables show the percentage of subjects for which the lin-

	speech features					
linguistic features	**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 6.1: 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.

	Electrodermal Activity (EDA) features								
		phasic component					**tonic component**		
linguistic features	**eda symp**	**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

Table 6.2: Summary results of the correlations between Electrodermal Activity 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.

guistic features in the group were significantly correlated with the speech features and the electrodermal activity features, regardless of the correlation value.

Comparing the two tables, it can be observed that the linguistic features that belong to the same group were significant for a similar number of subjects and, in some cases, for the same number of subjects. Furthermore, correlations with features encoding syntactic-related phenomena were, on average, more significant for a higher number of subjects compared to correlations with lexical and morpho-syntactic features.

Table 6.1 shows the correlations between linguistic and speech features. It is noticeable that the *mean duration* and *signal duration* are the acoustic parameters that significantly correlate with almost all of the linguistic features for most subjects. The F_0 and F_3 features also show significant correlations for many subjects, while F_1 and F_2 are the least correlated features among the acoustic ones. When focusing on the different types of linguistic phenomena, it can be seen that acoustic features related to sentence length (*mean duration* and *signal duration*) are consistently correlated for most subjects with the linguistic features that encode aspects of sentence length.

Table 6.1 shows that there are also high correlations with the syntactic features that describe the use of subordination and the structure of the parsed tree, particularly for F_3, with up to 70% of the subjects achieving a significant correlation. Most linguistic features with significant correlations pertain to different aspects of linguistic complexity. Apart from sentence length, which is widely considered a shallow proxy of linguistic complexity and text readability [62], there are 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). However, features related to the use of lexicon, such as lexical density, are correlated with acoustic parameters for very few subjects.

The analysis of the correlations between Electrodermal Activity and Linguistic features in Table 6.2 reveals that the features with the most significant correlations with linguistic features for more than half of the subjects are *std ph*, *sum peaks* and *no peaks* from the phasic component. Conversely, *max peaks* and *mean pk* have fewer correlations and are significant for a smaller number of subjects. The feature from the tonic component that shows the most correlations is *std ton* (i.e., the standard deviation of the tonic component). This features has indeed a strong relationship with linguistic features that describe syntactic phenomena. The other tonic component features, *max ton*, *mean ton*, and the feature of the power spectrum, *edaSymp*, are significant for fewer subjects, and have the most significant correlations with linguistic features related to subordination phenomena. When looking at the different groups of linguistic features, it can be seen that features related to syntax, particularly subordination, have more correlation than lexical and morpho-syntactic features.

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

This part of the study aimed to evaluate the ability of the features describing the linguistic profile of texts to predict the Electrodermal Activity and Speech features. A Support Vector Regressor (Section 3.2.2) was used for this task, implemented with a Radial Basis Function kernel and standard parameters[1]. The regressor took all the linguistic features as input and used them to predict the Electrodermal Activity and Speech features. Leave-one-subject-out cross-validation was used to account for within-subject repetitions by training the model on all subjects except one and testing it on the left-out subject. The model performance was evaluated by comparing it with a baseline, calculated by running the model with only the length of sentences as an input feature.

After the completion of the task, a feature importance analysis was conducted. This involved selecting all the features for each participant that had a predicted value by the RBF-SVR, which correlated to at least ±0.30 with their actual value and was statistically significant (p-value < 0.05). Following this, an SVR model with a linear kernel was utilized to predict each of these selected features, i.e., the EDA and speech features were employed to predict the selected linguistic features. After, the coefficients that the linear-SVR assigned to the predicting features were extracted and used to construct the feature rankings.

The effectiveness of the implemented SVR models was assessed by comparing their predictions to the actual values of the features being analyzed. This was done by calculating the mean Spearman's correlation and variance for all subjects. The results for both signals (acoustic and electrodermal activity) are presented as percentages, which show the number of subjects for which the predictions were significantly correlated, regardless of the correlation values.

	% significant subjects	**mean correlation**	**correlation variance**	**baseline**
F_0	15%	0.4032	0.0027	0.3622
F_1	61%	**0.5419**	0.0181	-0.0272
F_2	97%	**0.5424**	0.0089	0.0524
F_3	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 6.3: 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.

[1] The standard parameters provided by Scikit Learn implementation in the function sklearn.svm.SVR

Table 6.3 shows the results for predicting acoustic features. The table contains the percentage of participants demonstrating a significant correlation between the predicted variable and the target variable. It also shows the mean and variance of the correlation, along with the correlation obtained by the baseline model. The model outperforms the baseline in all cases. The low variance in correlation across subjects indicates that the model's predictions for acoustic values are consistent among different subjects. As expected from the previous correlation results, the prediction of *mean duration* and *signal duration* is significant for almost all subjects, as these features are related to sentence length, which the model can identify in its input. However, the model, which uses the entire set of linguistic features, slightly surpasses the baseline, indicating that acoustic features are also influenced by other linguistic properties beyond sentence length. While the prediction of F_1 and F_2 is significant for a large number of subjects (>60%), F_0 and F_3 are significant for a smaller number of subjects (<30%). This is the opposite of what was seen in the correlation analysis, where the latter two features were significant for up to 70% of the subjects.

	% 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

Table 6.4: Regression results for the prediction of electrodermal activity features. Highlighted in bold are the features that obtain a mean correlation value across subjects > 0.50.

Table 6.4 shows the results for predicting Electrodermal Activity features. Similar to the prediction of acoustic features, the model's predictions in this case also outperform the baseline (which, as previously stated, is calculated using only sentence length as input). This is particularly evident in the prediction of *edasymp* and *sum pks*, which are among the best-predicted features when compared to the baseline. However, the variance of the predictions is higher for some features (e.g., *mean ph*, *mean ton*) compared to the relatively low variance obtained in predicting acoustic features. It is also notable that features describing the phasic component are overall predicted with higher accuracy, with the exception of *mean ph*. On the other hand, the maximum and mean values of the tonic component (*max ton*, *mean ton*) are less predictable by linguistic features. This result reflects the findings of the correlation analysis (see Table 6.2), where the pairwise correlation between these features and linguistic features was significant for a low percentage of subjects (on average, from 3% to 30%).

A close examination of the feature importance analysis results for speech feature prediction reveals a significant influence stemming from aspects related to sentence length. This includes not only the length of the sentence itself, but also other associated factors such as the count of verbal heads. This is because lengthier sentences typically incorporate more clauses through coordination or subordination. The study also found the distribution of subjects and their positioning (either pre-verbal or post-verbal) within the sentence to be highly indicative. Regarding the feature importance analysis for predicting the EDA, no clear influence pattern is discernible in the tonic component. Rather, the EDA phasic component prediction seems to correlate more with sentence length and its associated factors, like the count of prepositional chains and subordination occurrences. There also appears to be some influence from punctuation, which is thought to be linked to sentence length, as longer sentences generally have more punctuation.

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

This second task aimed to evaluate the effectiveness of the acoustic and physiological features in predicting the features underlying the internal structure of a text. To do this, a Support Vector Regression model was built using the Speech and Electrodermal Activity features to predict all the linguistic parameters. Like in the first task, the Support Vector Regressor was implemented with a Radial Basis Function kernel and standard parameters. The model performance was evaluated by comparing it to a baseline, which was created by training the model using only the Voiced Duration feature as an input feature.

As what happened in the task developed in Section 6.3.3, a feature importance analysis was implemented at the end of this task. The procedure is the same as the one detailed in the first task, with the difference that in this case the linguistic features were used to predict the EDA and speech features. Also in this case, the coefficients the linear-SVR assigned to the predicting features were extracted and used to construct the feature rankings.

As with the previous analysis, the effectiveness of the model was determined by comparing its predictions to the true values of the features being analyzed, using the mean Spearman's correlation and its variance for all subjects. The results of this analysis are presented in Table 6.5. The table only shows features for which the number of significant subjects was $\geq$15. The full table with all

linguistic features	**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
dependency links length	33 (100)	**0.6771**	0.0085	0.3486
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

Table 6.5: 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.

results can be found in Appendix C.

The predictions of the implemented model are always better than the baseline for all features. Additionally, the very low variance in the correlation coefficients among the different subjects confirms that the model's predictions are robust. The highest correlations are seen for *sentence length* and for features related to length, but modeling more complex properties of the global and local parsed tree structure. These include the average depth of the parsed tree, the average length of the dependency links, and the presence and internal structure of complex nominal complements headed by a preposition (i.e., *prepositional chains* *length*, *prepositional chains number*). These are also the features for which the correlations are significant for a high percentage of subjects ($\geq 90\%$).

When considering the division of linguistic features into different groups of phenomena, the best results are seen for features describing the use of subordination, with a mean correlation above 0.60 and predictions being significant for almost all subjects. On the other hand, Electrodermal Activity and Speech features have only a small impact on predicting morpho-syntactic properties. In terms of the distribution of grammatical categories, while the correlations are around 0.4 or higher, they are significant for only a few subjects. As seen in Table 6.5, the only exceptions are the presence of subordinating conjunctions and auxiliaries in the present tense, which are significantly correlated for a high number of subjects (i.e., 20 subjects out of 33 for subordinating conjunctions, 16 subjects out of 33 for auxiliaries in the present tense).

The feature importance analysis uncovers trends of influence across various types of linguistic features. For features related to lexical density, the most crucial predictors are F_1, F_3, signal duration, and the EDA tonic component. When it comes to morpho-syntactic features, F_2 and signal duration emerge as the most impactful predictors, while the EDA features don't demonstrate any consistent trend. Regarding features of syntactic relations and those pertaining to both local and global parsed tree structures, F_2 and signal duration are again recognized as the most significant predictors, although the EDA phasic component also exerts a strong influence. Finally, in relation to subordination phenomena, signal duration is highlighted as the most important predictor, but no clear trend can be discerned from the EDA features.

6.4 Discussion

The aim of this case study was to examine the relationship between the linguistic characteristics of neutral and emotional texts and the emotional response of the reader, as measured by electrodermal activity and speech signal analysis. A combination of correlation and regression analysis was used to investigate how the linguistic structure of the texts relates to these signals. The assumption was that both electrodermal activity and speech signal would indicate the emotional response elicited by the task, as assessed by the Self-Assessment Manikin method.

The correlation analysis of Section 6.3.2 revealed a statistically significant relationship between certain linguistic properties of the text and speech and electrodermal activity features. Specifically, significance was found between linguistic features related to aspects of syntactic complexity, such as the use of subordination and verbal predicate structure, and speech features that describe some prosodic aspects of speech often associated with emotional states (e.g., F_0, F_3 variation over time). The findings further illustrate how speech features, such as signal duration, could serve as markers of linguistic complexity due to their strong association with sentence length. Indeed, sentence length acts as a complexity indicator, as lengthier sentences usually contain more complex dependencies and syntactic formations, including numerous subordinate clauses. This complexity elevates the cognitive exertion needed to understand the sentence and its structural basis. Consequently, given the direct correlation between sentence length and signal duration, the latter can also be perceived as a complexity metric. Additionally, electrodermal activity 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 [18].

The strong significant relationship between the linguistic characteristics and the acoustic and EDA features was further confirmed by the strong prediction performance of the linguistic-driven SVR models. The combination of linguistic features showed a significant and relevant ability to predict ANS-related features, both when they described characteristics of the voice spectrum (i.e., fundamental frequency and formants) that could be affected by respiratory activity, and when they described the physiological arousal manifested by sweat gland activity.

The SVR model also demonstrated an exceptional ability to predict *edasymp* values in addition to the EDA features previously identified by the correlation analysis. This feature is a dependable indicator of sympathetic system activity and a well-established stress marker, which supports the idea of a connection between features commonly considered as proxies of linguistic complexity, particularly at the syntactic level, and stress reactions in the subject [232].

However, this outcome could raise a double possibility of interpretation. This result could be interpreted in two ways. On one hand, the linguistic structure of the spoken sentence may obscure the true impact of voice prosody and EDA in determining a speaker's emotional state. The variations in prosody and EDA dynamics could be caused by mechanical changes in respiratory activity associated with speech, which is known to affect both acoustic and EDA characteristics. On the other hand, the linguistic structure itself could directly impact a subject's emotional state, which would be accurately captured by the speech and EDA features.

The last hypothesis is supported by previous research, introducing the idea that a combination of speech processing features and linguistic features can be used to accurately recognize emotional state [13, 255]. However, these studies typically focus on lexical and contextual aspects of language and do not take into account other important features such as morpho-syntactic or syntactic information. These features have been shown to have a significant impact on an individual's emotional state as they are related to various psycholinguistic phenomena and can affect cognitive load and language processing difficulty. The results of the study presented in this manuscript align with previous studies, particularly [43, 133], which have shown that the same set of linguistic features used in our study are highly correlated with conscious judgments of perceived sentence complexity given by native speakers.

This study uncovered further evidence that acoustic and physiological signals can accurately predict a wide range of linguistic characteristics, which play a role in shaping the grammatical and syntactic structure of language. This supports the idea that there is a strong connection between a speaker's emotional state and the way they use language. Additionally, incorporating cognitive signals into natural language processing research could lead to more sophisticated models, and provide deeper insight into the distinctions between human and machine language understanding.

Conclusions

This thesis has tackled the relationship between natural language and emotions, trying to explain which linguistic phenomena can elicit emotive reactions and which of them are related to the emotional aspects of the speaker and of the reader. This relationship has been examined from different perspectives and by introducing different modalities in the analysis. Natural language has been the leitmotiv that has guided all the analyses presented in the three case studies. By collecting a wide set of linguistically-motivated features, it is possible to represent any text with its linguistic profile, capturing information about various levels of linguistic phenomena and stylistic elements of language. To assess whether these features are sufficient for describing the relationship of language with emotions, they were paired and compared also with implicit vectorial representations generated by a Language Model, which is able to encode lexical and semantic relational aspects that the linguistic profile may fail to detect. When available, language was paired with other modalities, specifically speech and electrodermal activity, to examine how emotionally encoded language impacts the bodily parameters of a reader.

The first case study of this thesis presented a novel method for modeling human perception of sentence complexity. Playing an essential role in effective communication, the study of language complexity is fundamental to understanding which linguistic phenomena facilitate feelings of positive engagement in a human-computer or a human-human interaction. To tackle these aspects, a group of subjects recruited through a crowdsourcing task was asked to annotate a corpus of sentences in terms of perceived complexity. The analyses of the case study were divided into two main parts, based on whether complexity was studied for sentences in isolation or within a context.

The first part of this case study focused on the complexity of sentences presented in isolation for two languages, English and Italian. As first analysis, a Support Vector Classifier was applied to assess the role of linguistic features in predicting how much annotators agree on expressing the level of complexity of the sentences. It was shown that deep syntactic features, such as the use of subordination and nominal modification, play a significant role in predicting the level of agreement of human annotators. Interestingly, the classifier required only a few of the linguistic features to predict the agreement level when more than half of the annotators report the same judgment of complexity. When there is no consensus among the annotators, the performance of the classifier in the prediction of the agreement lowers significantly. This is a clear sign that there are some specific linguistic phenomena that most annotators take into account for the evaluation of complexity. This assumption is confirmed by the following analysis, which tackled the correlation between the explicit linguistic features and the complexity judgments assigned by annotators. The analysis revealed that syntactic phenomena related to sentence structure are among the top-ranked features characterizing sentences rated highly complex by a given number of agreeing annotators. Moreover, the set of selected explicit features significantly contributes to automatically predicting the human judgment of sentence complexity, meaning that the examined linguistic factors play an important role in the perception of complexity.

The second part of this first case study examined how the context surrounding a sentence influences the perception of its complexity by humans. Indeed, although a sentence by itself may be perceived as highly complex, it is possible that additional context helps the reader of a text (or, in general, the receiver of a message) in understanding the sentences with less cognitive effort. From the dataset created in the first part of the study, only English sentences were extracted and enriched with the preceding and following sentences, representing the context of the original sentence. Three different windows of context were studied according to the position the original sentence occupied within them (begin, center, end).

The results of the first analysis revealed a strong agreement between annotators, regardless of the position the evaluated sentence occupies with respect to the other contextual sentences. This analysis also showed that annotators reach a stronger agreement when the sentence is rated as simple, while the level of agreement decreases when the sentence is rated as complex. This discrepancy could indicate that the processing of complex sentence is subjective and possibly more dependent on the knowledge of each annotator, even when the context is provided. Subsequently, an attempt was made to try to predict the values of complexity assigned to the sentences by the annotators. Differently from the first part of the study, in which only explicit linguistic features were taken into consideration, this second part of the study also leveraged implicit linguistic features obtained from a pre-trained and fine-tuned language model (BERT). The results have shown that models using explicit linguistic features achieve higher accuracy than BERT in the prediction of the scores of complexity assigned to sentences. This was especially true when the models used explicit linguistic features from all contextual sentences, in addition to the ones of the sole rated sentence. However, this result is highly dependent on the fact that very few data were available for this task, while language models notoriously require training on larger datasets.

As done when studying complexity of sentences in isolation, this study assessed the correlation between the explicit linguistic features and the judgments of complexity assigned by annotators also for the sentences presented with context. Contrary to what was done in the first part of the study, this second part also assessed the correlation with the standard deviation of the complexity scores assigned to sentences. From the analysis of the correlation between the features and the judgments of complexity, it emerged that the annotators paid attention during the task both to the sentence to be evaluated and the context in which it was presented. The analyses also evidenced that the presence of numbers and of phenomena capturing aspects of structural complexity, especially in long sentences, contribute to higher perceived complexity. When the explicit linguistic features are correlated to the standard deviation of the complexity judgments, fewer correlations emerge. Still, it is noticeable from the results that humans have paid attention to the context in which sentences appear. Although fewer correlations are present, also in this case the presence of numerical information and elements that mark structural complexity has a strong influence on the standard deviation of the complexity judgments.

The second case study tackled the analysis of emotion elicitation from audiovisual stimuli, based on a dataset of sentences extracted from the movie Forrest Gump and annotated with the emotions perceived by a group of subjects. The main goal of this case study was to explore the interactions between the linguistic aspects of the movie dialogues and the emotions elicited in participants that watched the movie in a controlled setting. To achieve this goal, the study leveraged both the explicit linguistic features from the linguistic profile and the implicit contextual features generated by the language model BERT. Both the set of features were used to predict the emotions experienced by the participants in the study.

The results showed that contextual embeddings extracted from the sentences pronounced by the movie characters can be leveraged to accurately predict specific perceived emotions, even with a small amount of text as input. However, for predicting generic emotional elicitation, a larger amount of text is required for predictive models to yield accurate results. On the other hand, lexical, morpho-syntactic, and syntactic aspects of the dialogue sentences are not strong predictors of the emotional elicitation subjects experienced during the view of the movie. This outcome is expected, as these features encode little to no information regarding the emotive state and sentiment of what is expressed in the sentence. Instead, contextual embeddings can capture this information because they are aware of how the words within a sentence are co-located and encode more information about the semantics of the sentence itself. It is also important to notice that the stimuli coming from a movie are multiple (i.e., images, speech, music, other contextual cues), thus linguistic aspects by themselves cannot be the sole emotion elicitors. However, the results of this study imply the importance of dialogue in studying emotional elicitation and perception, even in the presence of multiple stimuli. The findings suggest that contextual embeddings can be a useful tool for predicting emotional responses to audiovisual stimuli.

The last case study highlights the importance of linguistic features in shaping the emotional response of readers and supports the idea that emotional and cognitive signals play a significant role in natural language processing. This study explored the relationship between the explicit linguistic features of emotionally encoded and neutral texts and the bodily emotional response of the reader, as measured by electrodermal activity and speech signal analysis. The results revealed a significant relationship between certain linguistic properties of the texts and speech and electrodermal activity features. In particular, linguistic features related to syntactic complexity, such as subordination and the verbal predicate structure, were strongly correlated with the speech features related to emotional states, such as the variation in fundamental frequency and the voice formants. Additionally, electrodermal activity features that reflect arousing states, such as fear and anxiety, were correlated with most linguistic properties of the texts. Furthermore, a linguistic-driven Support Vector Regression model demonstrated that the selected set of linguistic features can be used to accurately predict both autonomic nervous system-related features and *edasymp* values, which are reliable indicators of the activity of the sympathetic system and a well-established stress marker. These findings suggest that there is a strong connection between a speaker's emotional state and the way they use language. The study aligns with previous research, demonstrating that a combination of speech processing features and linguistic features can accurately recognize a human's emotional state. However, previous studies often focus on lexical and contextual aspects of language and neglect other important features such as morpho-syntactic or syntactic information. The findings of this study provide evidence that acoustic and physiological signals can predict a wide range of linguistic characteristics, which play a role in shaping the grammatical and syntactic structure of language.

Future applications

The findings reported in this thesis present a wealth of potential applications across diverse fields, ranging from communication studies, to human-computer interactions, multimedia analysis, and content-based recommendations. It opens the opportunity for further exploration and utilization in areas where understanding the interaction between language and emotion is critical, not only in the field of research but also in real-world applications.

In the field of educational technology, these results could help in designing intelligent tutoring systems that are more sensitive to the cognitive state of the learner. The tutoring system could adapt its instructions or exercises based on the level of complexity a student can handle, as determined by the linguistic features of their responses. Moreover, in language learning platforms, the understanding of perceived sentence complexity could help to create more efficient learning content tailored to individual student's proficiency.

In the area of human-computer interaction, these findings could be used to enhance algorithms for natural language understanding and sentiment analysis. For instance, customer support bots could be trained to adjust their responses according to the emotional state of the customer, inferred by analyzing the customer's speech and text. Additionally, the results of the study could be applied for creating AI chatbots with a higher degree of empathy, giving them the ability to respond more sensitively to the emotional content of human language.

The conclusions of the second case study could be applied to the entertainment industry, in particular to scriptwriting and filmmaking. Understanding how language in dialogue elicits emotional responses can guide writers in designing narratives that effectively engage audiences. Furthermore, platforms like Netflix could use these findings to refine their recommendation algorithms, linking emotional responses to dialogue, and providing viewers with content that resonates more emotionally.

As for the last case study, the relationship between the linguistic features of emotionally encoded and neutral texts and the bodily emotional response of the reader could have clinical applications in mental health. In psychotherapy, for example, this research could inform the development of diagnostic tools that analyze patients' language use and physiological responses, providing insights into their emotional states. The correlation found between linguistic properties in texts and electrodermal activity could be valuable in biofeedback therapy, where patients learn to modulate their physiological reactions.

Moreover, the findings of the last case study could be applied to voice-operated virtual assistants or voice user interface devices, like Amazon's Alexa or Google Home. These devices could be optimized to not only understand the content of the command but also the emotional context of the speaker, leading to more fitting responses.

References

[1] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249, 2021.

[2] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829, 2021.

[3] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, page e12189, 2020.

[4] Zishan Ahmad, Raghav Jindal, Asif Ekbal, and Pushpak Bhattachharyya. Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. *Expert Systems with Applications*, 139:112851, 2020.

[5] Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.

[6] David M Alexander, Chris Trengove, P Johnston, Tim Cooper, JP August, and Evian Gordon. Separating individual skin conductance responses in a short interstimulus-interval paradigm. *Journal of neuroscience methods*, 146(1):116–123, 2005.

[7] Arif Ali, Abdur Rasheed, Afaq Ahmed Siddiqui, Maliha Naseer, Saba Wasim, and Waseem Akhtar. Non-parametric test for ordered medians: The jonckheere terpstra test. *International Journal of Statistics in Medical Research*, 4(2):203, 2015.

[8] Merav Allouch, Amos Azaria, Rina Azoulay, Ester Ben-Izchak, Moti Zwilling, and Ditza A Zachor. Automatic detection of insulting sentences in conversation. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pages 1–4. IEEE, 2018.

[9] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586, 2005.

[10] Shlomo Argamon. Computational register analysis and synthesis. *Register Studies, Forthcoming*, 2019.

[11] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text & talk*, 23(3):321–346, 2003.

[12] Ivon Arroyo, David G Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. Emotion sensors go to school. In *AIED*, volume 200, pages 17–24. Citeseer, 2009.

[13] Bagus Tris Atmaja and Masato Akagi. Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm. *Speech Communication*, 126:9–21, 2021.

[14] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[15] Dominik R Bach. A head-to-head comparison of scralyze and ledalab, two model-based methods for skin conductance analysis. *Biological psychology*, 103:63–68, 2014.

[16] Dominik R Bach and Karl J Friston. Model-based analysis of skin conductance responses: Towards causal models in psychophysiology. *Psychophysiology*, 50(1):15–22, 2013.

[17] Richard C Baker and Daniel O Guttfreund. The effects of written autobiographical recollection induction procedures on mood. *Journal of Clinical Psychology*, 49(4):563–568, 1993.

[18] Andrea Baldini, Sergio Frumento, Danilo Menicucci, Angelo Gemignani, Enzo Pasquale Scilingo, and Alberto Greco. Subjective fear in virtual reality: a linear mixed-effects analysis of skin conductance. *IEEE Transactions on Affective Computing*, 2022.

[19] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[20] Nayan Banik and Md Hasan Hafizur Rahman. Evaluation of naïve bayes and support vector machines on bangla textual movie reviews. In *2018 international conference on Bangla speech and language processing (ICBSLP)*, pages 1–6. IEEE, 2018.

[21] Lisa Feldman Barrett. Valence is a basic building block of emotional life. *Journal of Research in Personality*, 40(1):35–55, 2006.

[22] Robert J Barry, Sabine Feldmann, Evian Gordon, Kathryn I Cocker, and Chris Rennie. Elicitation and habituation of the electrodermal orienting response in a short interstimulus interval paradigm. *International journal of psychophysiology*, 15(3):247–253, 1993.

[23] Marian S Bartlett, Bjorn Braathen, Gwen Littlewort-Ford, John Hershey, Ian Fasel, Tim Marks, Evan Smith, Terrence J Sejnowski, and Javier R Movellan. Automatic analysis of spontaneous facial behavior: A final project report. Technical report, Technical Report UCSD MPLab TR 2001.08, University of California, San Diego, 2001.

[24] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 568–573. IEEE, 2005.

[25] Anton Batliner, Kerstin Fischer, Richard Huber, Jörg Spilker, and Elmar Nöth. How to find trouble in communication. *Speech communication*, 40(1-2):117–143, 2003.

[26] Susana Bautista and Horacio Saggion. Can numerical expressions be simpler? implementation and demostration of a numerical simplification system for spanish. In *LREC*, pages 956–962, 2014.

[27] Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*, 2018.

[28] Christopher Beedie, Peter Terry, and Andrew Lane. Distinctions between emotion and mood. *Cognition & Emotion*, 19(6):847–878, 2005.

[29] Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 190(1):80–91, 2010.

[30] Mathias Benedek and Christian Kaernbach. Decomposition of skin conductance data by means of nonnegative deconvolution. *psychophysiology*, 47(4):647–658, 2010.

[31] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[32] Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, 2012.

[33] Jean-Philippe Bernardy, Shalom Lappin, and Jey Han Lau. The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Melbourne, Australia, 2018. Association for Computational Linguistics.

[34] Cindy L Bethel, Kristen Salomon, Robin R Murphy, and Jennifer L Burke. Survey of psychophysiology measurements applied to human-robot interaction. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 732–737. IEEE, 2007.

[35] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345, 2001.

[36] Wolfram Boucsein. *Electrodermal activity.* Springer Science & Business Media, 2012.

[37] Elizabeth A Boyle, Anne H Anderson, and Alison Newlands. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and speech*, 37(1):1–20, 1994.

[38] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

[39] Margaret M Bradley and Peter J Lang. Measuring emotion: Behavior, feeling, and physiology. *Cognitive neuroscience of emotion*, 25:49–59, 2000.

[40] Claude Breault and Raymond Ducharme. Effect of intertrial intervals on recovery and amplitude of electrodermal reactions. *International journal of psychophysiology*, 14(1):75–80, 1993.

[41] Assaf Breska, Keren Maoz, and Gershon Ben-Shakhar. Interstimulus intervals for skin conductance response measurement. *Psychophysiology*, 48(4):437–440, 2011.

[42] Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7145–7151, 2020.

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

[44] John P. Burg. A new analysis technique for time series data. *SIAM Journal on Applied Mathematics*, 28(2):263–273, 1975.

[45] Alejandro Luis Callara, Laura Sebastiani, Nicola Vanello, Enzo Pasquale Scilingo, and Alberto Greco. Parasympathetic-sympathetic causal interactions assessed by time-varying multivariate autoregressive modeling of electrodermal activity and heart-rate-variability. *IEEE Transactions on Biomedical Engineering*, 2021.

[46] Rafael A Calvo and Sidney D'Mello. Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.

[47] Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer, 2012.

[48] Erik Cambria, Robyn Speer, Catherine Havasi, and Amir Hussain. Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI fall symposium series*, 2010.

[49] C Estelle Campenni, Edward J Crawley, and Michael E Meier. Role of suggestion in odor-induced mood change. *Psychological Reports*, 94(3_suppl):1127–1136, 2004.

[50] Lea Canales and Patricio Martínez-Barco. Emotion detection from text: A survey. In *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pages 37–43, 2014.

[51] Walter B Cannon. The james-lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology*, 39(1/4):106–124, 1927.

[52] Justine Cassell and Timothy Bickmore. External manifestations of trustworthiness in the interface. *Communications of the ACM*, 43(12):50–56, 2000.

[53] Peter McFaul Chapman. *Models of engagement: intrinsically motivated interaction with multimedia learning software.* PhD thesis, University of Waterloo, 1997.

[54] Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317, 2019.

[55] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, 2019.

[56] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

[57] Girija Chetty, Michael Wagner, and Roland Goecke. A multilevel fusion approach for audiovisual emotion recognition. In *AVSP*, pages 115–120, 2008.

[58] Donald G Childers and Chih K Lee. Vocal quality factors: Analysis, synthesis, and perception. *the Journal of the Acoustical Society of America*, 90(5):2394–2410, 1991.

[59] Andry Chowanda, Rhio Sutoyo, Sansiri Tanachutiwat, et al. Exploring text-based emotions recognition machine learning techniques on social media conversation. *Procedia Computer Science*, 179:821–828, 2021.

[60] Arthur Charles Clarke. *2001: A Space Odyssey.* New American Library, New York, NY, 1968.

[61] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[62] Kevin Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(1):97–135, 2014.

[63] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

[64] William Jay Conover. *Practical nonparametric statistics*, volume 350. john wiley & sons, 1999.

[65] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[66] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[67] Walter Daelemans. Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14*, pages 451–462. Springer, 2013.

[68] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals.* Oxford University Press, USA, 1998.

[69] Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. *Handbook of Psychophysiology*, 2017.

[70] Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Lrec*, volume 6, pages 449–454, 2006.

[71] Tullio De Mauro. *Grande dizionario italiano dell'uso (GRADIT).* Torino, UTET, 2000.

[72] John R Deller Jr. Discrete-time processing of speech signals. In *Discrete-time processing of speech signals*, pages 908–908. Prentice Hall PTR, 1993.

[73] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[74] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):43, 2015.

[75] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *IEEE Transactions on pattern analysis and machine intelligence*, 21(10):974–989, 1999.

[76] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. Emotional speech: Towards a new generation of databases. *Speech communication*, 40(1-2):33–60, 2003.

[77] Dr.speech. `https://www.drspeech.com/`. Accessed: 26-06-2023.

[78] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020.

[79] Robert Edelberg. Electrical activity of the skin: Its measurement and uses in psychophysiology. *Handbook of psychophysiology*, pages 367–418, 1972.

[80] Robert Edelberg. Electrodermal mechanisms: A critique of the two-effector hypothesis and a proposed replacement. *Progress in electrodermal research*, pages 7–29, 1993.

[81] Robert Edelberg and Michael Muller. Prior activity as a determinant of electrodermal recovery rate. *Psychophysiology*, 18(1):17–25, 1981.

[82] Panteleimon Ekkekakis. *The measurement of affect, mood, and emotion: a guide for health-behavioral research.* Cambridge University Press, 2013.

[83] Paul Ekman. Pictures of facial affect. *Consulting Psychologists Press*, 1976.

[84] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[85] Paul Ekman. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological bulletin*, 1994.

[86] Paul Ekman and Wallace V Friesen. *Facial action coding system: Investigator's guide.* Consulting Psychologists Press, 1978.

[87] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*, volume 11. Elsevier, 2013.

[88] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[89] Paul Ekman and Dacher Keltner. Universal facial expressions of emotion. *California Mental Health Research Digest*, 8(4), 1970.

[90] Paul Ekman, Robert W Levenson, and Wallace V Friesen. Autonomic nervous system activity distinguishes among emotions. *science*, 221(4616):1208–1210, 1983.

[91] Paul Ekman and Harriet Oster. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554, 1979.

[92] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[93] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, mar 2011.

[94] Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005.

[95] Irfan A. Essa and Alex Paul Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997.

[96] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

[97] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.

[98] Raul Fernandez and Rosalind W Picard. Modeling drivers' speech under stress. *Speech communication*, 40(1-2):145–159, 2003.

[99] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[100] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

[101] Don C Fowles, Margaret J Christie, Robert Edelberg, William W Grings, David T Lykken, and Peter H Venables. Publication recommendations for electrodermal measurements. *Psychophysiology*, 18(3):232–239, 1981.

[102] Nico Henri Frijda. Emotions are functional, most of the time. In Paul Ekman and Richard J Davidson, editors, *The nature of emotion*, pages 112–122. New York: Oxford University Press, 1994.

[103] Shadi Ghiasi, Getano Valenza, Maria Sole Morelli, Matteo Bianchi, Enzo Pasquale Scilingo, and Alberto Greco. The role of haptic stimuli on affective reading: a pilot study. In *2019 41st Annual EMBC*, pages 4938–4941. IEEE, 2019.

[104] Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 24(11):1–76, 1998.

[105] Edward Gibson. The dependency locality theory: A distance–based theory of linguistic complexity. *W.O.A. Marants and Y. Miyashita (Eds.), Image, Language and Brain*, Cambridge, MA: MIT Press:95–126, 2000.

[106] Cheryl L Giddens, Kirk W Barron, Jennifer Byrd-Craven, Keith F Clark, and A Scott Winter. Vocal indices of stress: a review. *Journal of voice*, 27(3):390–e21, 2013.

[107] Daniela Girardi, Filippo Lanubile, and Nicole Novielli. Emotion detection using noninvasive low cost sensors. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 125–130. IEEE, 2017.

[108] Daniel Goleman. *Emotional intelligence*. Bantam, 2006.

[109] Dhruvi D Gosai, Himangini J Gohil, and Hardik S Jayswal. A review on a emotion detection and recognization from text using natural language processing. *International Journal of Applied Engineering Research*, 13(9):6745–6750, 2018.

[110] Alberto Greco, Antonio Lanata, Gaetano Valenza, Enzo Pasquale Scilingo, and Luca Citi. Electrodermal activity processing: A convex optimization approach. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2290–2293. IEEE, 2014.

[111] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, 2015.

[112] Alberto Greco, Gaetano Valenza, Jesús Lázaro, Jorge Mario Garzón-Rey, Jordi Aguiló, Concepcion De-la Camara, Raquel Bailón, and Enzo Pasquale Scilingo. Acute stress state classification based on electrodermal activity modeling. *IEEE Transactions on Affective Computing*, 2021.

[113] Mark K Greenwald, Edwin W Cook, and Peter J Lang. Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. *Journal of psychophysiology*, 1989.

[114] James J Gross and Robert W Levenson. Emotion elicitation using films. *Cognition & emotion*, 9(1):87–108, 1995.

[115] Haisong Gu and Qiang Ji. An automated face reader for fatigue detection. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 111–116. IEEE, 2004.

[116] Andrea Guidi, Claudio Gentili, Enzo Pasquale Scilingo, and Nicola Vanello. Analysis of speech features and personality traits. *Biomedical Signal Processing and Control*, 51:1–7, 2019.

[117] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46:389–422, 2002.

[118] Mousannif Hajar et al. Using youtube comments for text-based emotion recognition. *Procedia Computer Science*, 83:292–299, 2016.

[119] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the NAACL*, pages 159–166, 2001.

[120] H van Halteren. Linguistic profiling for authorship recognition and verification. 2004.

[121] Ellen L Hamaker and Marieke Wichers. No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, 26(1):10–15, 2017.

[122] John HL Hansen, Sahar E Bou-Ghazale, Ruhi Sarikaya, and Bryan Pellom. Getting started with susas: a speech under simulated and actual stress database. In *Eurospeech*, volume 97, pages 1743–46, 1997.

[123] Zellig S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[124] Daniel Haryadi and Gede Putra Kusuma. Emotion detection in text using nested long short-term memory. *International Journal of Advanced Computer Science and Applications*, 10(6), 2019.

[125] Maryam Hasan, Elke Rundensteiner, and Emanuel Agu. Emotex: Detecting emotions in twitter messages. 2014 ase bigdata. In *Socialcom/Cybersecurity Conference*, 2014.

[126] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, 7(1):35–51, 2019.

[127] John A Hawkins. An efficiency theory of complexity and related phenomena. In Geoffrey Sampson, David Gil, and Peter Trudgill, editors, *Language Complexity as an Evolving Variable. Studies in the Evolution of Language*, volume 13, pages 252–68. Oxford University Press, 2009.

[128] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[129] Nora Hollenstein, Maria Barrett, and Lisa 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*, pages 15–27, Marseille, France, May 2020. European Language Resources Association.

[130] Tom Hollenstein. This time, it's real: Affective flexibility, time scales, feedback loops, and the regulation of emotion. *Emotion Review*, 7(4):308–315, 2015.

[131] Clifford S Hopkins, Roy J Ratley, Daniel S Benincasa, and John J Grieco. Evaluation of voice stress analysis technology. In *Proceedings of the 38th annual Hawaii international conference on system sciences*, pages 20b–20b. IEEE, 2005.

[132] Chenyang Huang, Amine Trabelsi, and Osmar R Zaïane. Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert. *arXiv preprint arXiv:1904.00132*, 2019.

[133] Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. Sentence complexity in context. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 186–199, 2021.

[134] Anil K Jain and Stan Z Li. *Handbook of face recognition*, volume 1. Springer, 2011.

[135] Cynthia L Janes, Barbara D Strock, David G Weeks, and Julien Worland. The effect of stimulus significance on skin conductance recovery. *Psychophysiology*, 22(2):138–145, 1985.

[136] Joris H Janssen, Paul Tacken, JJG de Vries, Egon L van den Broek, Joyce HDM Westerink, Pim Haselager, and Wijnand A IJsselsteijn. Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection. *Human–Computer Interaction*, 28(6):479–517, 2013.

[137] R Jayakrishnan, Greeshma N Gopal, and MS Santhikrishna. Multi-class emotion detection and annotation in malayalam novels. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5. IEEE, 2018.

[138] Dan Jurafsky. *Speech & language processing.* Pearson Education India, 2000.

[139] James F Kaiser. On a simple algorithm to calculate the'energy'of a signal. In *International conference on acoustics, speech, and signal processing*, pages 381–384. IEEE, 1990.

[140] James F Kaiser. Some useful properties of teager's energy operators. In *1993 IEEE international conference on acoustics, speech, and signal processing*, volume 3, pages 149–152. IEEE, 1993.

[141] Ashish Kapoor, Winslow Burleson, and Rosalind W Picard. Automatic prediction of frustration. *International journal of human-computer studies*, 65(8):724–736, 2007.

[142] Mounika Karna, D Sujitha Juliet, and R Catherine Joy. Deep learning based text emotion recognition for chatbot applications. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, pages 988–993. IEEE, 2020.

[143] William R Kennedy, Gwen Wendelschafer-Crabb, and T Clark Brelje. Innervation and vasculature of human sweat glands: an immunohistochemistry-laser scanning confocal fluorescence microscopy study. *Journal of Neuroscience*, 14(11):6825–6833, 1994.

[144] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019.

[145] W. Kintsch, E. Kozminsky, W.J. Streby, G. McKoon, and J.M. Keenan. Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior, 14(2)*, 1975.

[146] Paul R Kleinginna and Anne M Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379, 1981.

[147] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction.* Cengage Learning, 2013.

[148] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018.

[149] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[150] Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, 1970.

[151] William H Kruskal. Historical notes on the wilcoxon unpaired two-sample test. *Journal of the American Statistical Association*, 52(279):356–360, 1957.

[152] Swarna Kuchibhotla, Hima Deepthi Vankayalapati, RS Vaddi, and Koteswara Rao Anne. A comparative analysis of classifiers in emotion recognition through acoustic features. *International Journal of Speech Technology*, 17:401–408, 2014.

[153] Puneet Kumar and Balasubramanian Raman. A bert based dual-channel explainable text emotion recognition system. *Neural Networks*, 150:392–407, 2022.

[154] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, 2020.

[155] Peter Kuppens, Zita Oravecz, and Francis Tuerlinckx. Feelings change: accounting for individual differences in the temporal dynamics of affect. *Journal of personality and social psychology*, 99(6):1042, 2010.

[156] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[157] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual.* NIMH, Center for the Study of Emotion & Attention Gainesville, FL, 2005.

[158] Nancy K Lankton, D Harrison McKnight, and John Tripp. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10):1, 2015.

[159] Brenda Laurel. *Computers as theatre.* Addison-Wesley, 2013.

[160] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[161] Chul Min Lee, Shrikanth S Narayanan, et al. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303, 2005.

[162] Suk Kyu Lee, Mungyu Bae, Woonghee Lee, and Hwangnam Kim. Cepp: Perceiving the emotional state of the user based on body posture. *Applied Sciences*, 7(10):978, 2017.

[163] Giada Lettieri, Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Paolo Papale, Monica Betta, Pietro Pietrini, and Luca Cecchetti. Emotionotopy in the human right temporo-parietal cortex. *Nature communications*, 10(1):1–13, 2019.

[164] Jasy Suet Yan Liew and Howard R Turtle. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL Student Research Workshop*, pages 73–80, 2016.

[165] Chong L Lim, Chris Rennie, Robert J Barry, Homayoun Bahramali, Ilario Lazzaro, Barry Manor, and Evian Gordon. Decomposing skin conductance into tonic and phasic components. *International Journal of Psychophysiology*, 25(2):97–109, 1997.

[166] Jen-Chun Lin, Chung-Hsien Wu, and Wen-Li Wei. Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. *IEEE Transactions on Multimedia*, 14(1):142–156, 2011.

[167] Yuan-Pin Lin, Yi-Hsuan Yang, and Tzyy-Ping Jung. Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening. *Frontiers in neuroscience*, 8:94, 2014.

[168] Gwen Littlewort, Ian Fasel, M Stewart Bartlett, and Javier R Movellan. Fully automatic coding of basic expressions from video. *University of California, San Diego, San Diego, CA*, 92093:14, 2002.

[169] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012.

[170] Huaping Liu, Yong Fang, and Qinghua Huang. Music emotion recognition using a variant of recurrent neural network. In *2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018)*. Atlantis Press, 2019.

[171] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273:271–280, 2018.

[172] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2010.

[173] Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1):5–18, 2005.

[174] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.

[175] Daniel Lundqvist, Anders Flykt, and Arne Öhman. Karolinska directed emotional faces. *Cognition and Emotion*, 1998.

[176] Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. The paisa'corpus of italian web texts. In *9th Web as Corpus Workshop (WaC-9)@ EACL 2014*, pages 36–43. EACL (European chapter of the Association for Computational Linguistics), 2014.

[177] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[178] Luyao Ma, Long Zhang, Wei Ye, and Wenhui Hu. Pkuse at semeval-2019 task 3: emotion detection with emotion-oriented neural attention network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 287–291, 2019.

[179] Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70, 2010.

[180] Aditya Malte and Pratik Ratadiya. Multilingual cyber abuse detection using advanced transformer architecture. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 784–789. IEEE, 2019.

[181] Claudia Manfredi, Leonardo Bocchi, and Giovanna Cantarella. A multipurpose user-friendly tool for voice analysis: Application to pathological adult voices. *Biomedical Signal Processing and Control*, 4(3):212–220, jul 2009.

[182] Claudia Manfredi, Andrea Giordano, Jean Schoentgen, Samia Fraj, Leonardo Bocchi, and Philippe H Dejonckere. Perturbation measurements in highly irregular voice signals: Performances/validity of analysis software tools. *Biomedical signal processing and control*, 7(4):409–416, 2012.

[183] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[184] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[185] Mika Mäntylä, Bram Adams, Giuseppe Destefanis, Daniel Graziotin, and Marco Ortu. Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity? In *Proceedings of the 13th international conference on mining software repositories*, pages 247–258, 2016.

[186] Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. The nencki affective picture system (naps): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior research methods*, 46(2):596–610, 2014.

[187] Simi Maria, Cristina Bosco, and Montemagni Simonetta. Less is more? towards a reduced inventory of categories for training a parser for the italian stanford dependencies. In *Language Resources and Evaluation 2014*, pages 83–90. European Language Resources Association (ELRA), 2014.

[188] Mohammad Mavadati, Peyten Sanger, and Mohammad H Mahoor. Extended disfa dataset: Investigating posed and spontaneous facial expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.

[189] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30, 2017.

[190] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, 2013.

[191] Patrick E McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1, 2010.

[192] Danielle S. McNamara. Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, 55(1):51–62, mar 2001.

[193] John McWhorter. The world's simplest grammars are creole grammars. *Linguistic typology*, 5(2):125–66, 2001.

[194] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[195] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61, 2016.

[196] Batja Mesquita and Robert Walker. Cultural differences in emotions: A context for interpreting emotional experiences. *Behaviour research and therapy*, 41(7):777–793, 2003.

[197] Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. Linguistic profiling of a neural language model. *arXiv preprint arXiv:2010.01869*, 2020.

[198] Matti Miestamo et al. Grammatical complexity in a cross-linguistic perspective. *Language complexity: Typology, contact, change*, pages 23–41, 2008.

[199] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[200] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[201] George A Miller. *WordNet: An electronic lexical database.* MIT press, 1998.

[202] Saif M Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier, 2016.

[203] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

[204] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013.

[205] Maria Sole Morelli, Silvia Orlandi, and Claudia Manfredi. Biovoice: A multipurpose tool for voice analysis. *Biomedical Signal Processing and Control*, 64:102302, 2021.

[206] Donn Morrison, Ruili Wang, and Liyanage C De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech communication*, 49(2):98–112, 2007.

[207] Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[208] Ryohei Nakatsu. Toward the creation of a new medium for the multimedia era. *Proceedings of the IEEE*, 86(5):825–836, 1998.

[209] Markus Neuhäuser. *Wilcoxon–Mann–Whitney Test*, pages 1656–1658. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[210] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593, 2016.

[211] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, 2016.

[212] Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932, 2007.

[213] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Detection of stress and emotion in speech using traditional and fft based log energy features. In *Fourth International Conference on Information, Communications and*

Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, volume 3, pages 1619–1623. IEEE, 2003.

[214] Heather L O'Brien and Elaine G Toms. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955, 2008.

[215] O'Shaughnessy. *Speech communications: Human and machine*. Universities press, 1999.

[216] Jahna Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 369–378, 2010.

[217] Maja Pantic and Leon JM Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881–905, 2000.

[218] Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[219] Seo-Hui Park, Byung-Chull Bae, and Yun-Gyung Cheong. Emotion recognition from text stories using an emotion embedding model. In *2020 IEEE international conference on big data and smart computing (BigComp)*, pages 579–583. IEEE, 2020.

[220] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[221] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[222] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[223] Pierre Philippot. Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cognition and emotion*, 7(2):171–193, 1993.

[224] Rosalind W Picard. *Affective computing*. MIT press, Cambridge, MA, 1997.

[225] Rosalind W Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1-2):55–64, 2003.

[226] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[227] Incorporated Plural Publishing. *Vocal health and pedagogy: Science, assessment, and treatment*. Plural Publishing, 2017.

[228] R Plutchik. Emotion: A psychoevolutionary synthesis. New York: Ed, 1980.

[229] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[230] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[231] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38, 2013.

[232] Hugo F Posada-Quintero, John P Florian, Alvaro D Orjuela-Cañón, Tomas Aljama-Corrales, Sonia Charleston-Villalobos, and Ki H Chon. Power spectral density analysis of electrodermal activity for sympathetic function assessment. *Annals of biomedical engineering*, 44(10):3124–3135, 2016.

[233] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[234] Whitney Quesenbery and Whitney Interactive Design. Dimensions of usability: Defining the conversation, driving the process. In *UPA 2003 Conference*, pages 23–27, 2003.

[235] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pretraining. 2018.

[236] Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Identifying emotions from walking using affective and deep features. *arXiv preprint arXiv:1906.11884*, 2019.

[237] Byron Reeves and Clifford Ivar Nass. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University press, 1996.

[238] Bernhard Riedl, Mark Nischik, Frank Birklein, Bernhard Neundörfer, and Hermann O Handwerker. Spatial extension of sudomotor axon reflex sweating in human skin. *Journal of the autonomic nervous system*, 69(2-3):83–88, 1998.

[239] Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146, 1976.

[240] Jonathan Rottenberg, RD Ray, JJ Gross, JA Coan, and JJB Allen. The handbook of emotion elicitation and assessment. *JJB Allen & JA Coan (Eds.)*, pages 9–28, 2007.

[241] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

[242] James A Russell. Affective space is bipolar. *Journal of personality and social psychology*, 37(3):345, 1979.

[243] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[244] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[245] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.

[246] Peter Salovey and John D Mayer. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211, 1990.

[247] Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, and Gholamreza Anbarjafari. Emotion recognition from skeletal movements. *Entropy*, 21(7):646, 2019.

[248] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and emotion*, 24(7):1153–1172, 2010.

[249] Elliot Schumacher, Maxine Eskenazi, Gwen Frishkoff, and Kevyn Collins-Thompson. Predicting the relative difficulty of single sentences with and without surrounding context. In *Conference on Empirical Methods in Natural Language Processing, pages 1871–1881.*, 2016.

[250] William A Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325, 1955.

[251] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Multimodal approaches for emotion recognition: a survey. In *Internet Imaging VI*, volume 5670, pages 56–67. SPIE, 2005.

[252] Janina Seubert, Amy F Rea, James Loughead, and Ute Habel. Mood induction with olfactory stimuli reveals differential affective responses in males and females. *Chemical senses*, 34(1):77–84, 2009.

[253] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2020.

[254] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074, 2018.

[255] Prabhav Singh, Ridam Srivastava, KPS Rana, and Vineet Kumar. A multi-modal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, 229:107316, 2021.

[256] Ilse Smits, Piet Ceuppens, and Marc S De Bodt. A comparative study of acoustic voice measurements by means of dr. speech and computerized speech lab. *Journal of Voice*, 19(2):187–196, 2005.

[257] Speech analyzer - sil language technology. `https://software.sil.org/speechanalyzer/`. Accessed: 26-06-2023.

[258] Calandra Speirs, Zorry Belchev, Amanda Fernandez, Stephanie Korol, and Christopher Sears. Are there age differences in attention to emotional images following a sad mood induction? evidence from a free-viewing eye-tracking paradigm. *Aging, Neuropsychology, and Cognition*, 25(6):928–957, 2018.

[259] Jacopo Staiano and Marco Guerini. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*, 2014.

[260] Ioanna-Ourania Stathopoulou and George A Tsihrintzis. Emotion recognition from body movements and gestures. In *Intelligent interactive multimedia systems and services*, pages 295–303. Springer, 2011.

[261] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[262] Milan Straka, Jan Hajic, and Jana Straková. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, 2016.

[263] Milan Straka and Jana Straková. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[264] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[265] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Lisbon, Portugal, 2004.

[266] Matla Suhasini and Badugu Srinivasu. Emotion detection framework for twitter data using supervised classifiers. In *Data Engineering and Communication Technology*, pages 565–576. Springer, 2020.

[267] Johan Sundberg and R Sataloff. Vocal tract resonance. *Vocal Health and Pedagogy: Science, Assessment, and Treatment*, 2005.

[268] Gillian Sutherland, Bobby Newman, and Stanley J. Rachman. Experimental investigations of the relations between mood and intrusive unwanted cognitions. *The British journal of medical psychology*, 55 Pt 2:127–38, 1982.

[269] KR Tan, ML Villarino, and C Maderazo. Automatic music mood recognition using russell's two-dimensional valence-arousal space from audio and lyrical data as classified using svm and naïve bayes. In *IOP Conference Series: Materials Science and Engineering*, volume 482, page 012019. IOP Publishing, 2019.

[270] HM Teager and SM Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. *Speech production and speech modelling*, pages 241–261, 1990.

[271] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, 2019. Association for Computational Linguistics.

[272] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[273] Ying-li Tian, Takeo Kanade, and Jeffrey F Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 229–234. IEEE, 2002.

[274] Ingo R Titze. The myoelastic aerodynamic theory of phonation. *(No Title)*, 2006.

[275] Donald W Trim. Calculus. *(No Title)*, 1993.

[276] Meike K Uhrig, Nadine Trautmann, Ulf Baumgärtner, Rolf-Detlef Treede, Florian Henrich, Wolfgang Hiller, and Susanne Marschall. Emotion elicitation: A comparison of pictures and films. *Frontiers in psychology*, 7:180, 2016.

[277] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 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, 2016.

[278] Vladimir Vapnik. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780, 1963.

[279] Vladimir Vapnik. Estimation of dependences based on empirical data: Springer series in statistics (springer series in statistics), 1982.

[280] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[281] Vladimir Vapnik and Alexey Chervonenkis. On a perceptron class. *Automation and Remote Control*, 25:112–120, 1964.

[282] Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

[283] Daniel Västfjäll. Emotion induction through music: A review of the musical mood induction procedure. *Musicae Scientiae*, 5(1_suppl):173–211, 2001.

[284] Daniel Västfjäll. Emotional reactions to product sounds. *Human Engineering*, 2003.

[285] Daniel Västfjäll and Mandel Kleiner. Emotion in product sound design. *Proceedings of Journées Design Sonore*, pages 1–17, 2002.

[286] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[287] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181, 2006.

[288] Vocal analysis tool — splab. `http://splab.cz/en/download/software/software-pro-analyzu-fonace`. Accessed: 26-06-2023.

[289] User manual – sygyt software. `https://www.sygyt.com/en/documentation/`. Accessed: 26-06-2023.

[290] Robert Walecki, Ognjen Rudovic, V Pavlovic, B Schuller, and Maja Pantic. Deep structured learning for facial expression intensity estimation. *Image Vis. Comput*, 259:143–154, 2017.

[291] Harald G Wallbott. Bodily expression of emotion. *European journal of social psychology*, 28(6):879–896, 1998.

[292] Nannan Wang, Xinbo Gao, Dacheng Tao, Heng Yang, and Xuelong Li. Facial feature point detection: A comprehensive survey. *Neurocomputing*, 275:50–65, 2018.

[293] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[294] Tmh kth: Wavesurfer. `http://www.speech.kth.se/wavesurfer/`. Accessed: 26-06-2023.

[295] Zhen Wen et al. Capturing subtle facial motions in 3d face tracking. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1343–1350. IEEE, 2003.

[296] Rainer Westermann, Kordelia Spies, Günter Stahl, and Friedrich W Hesse. Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of social psychology*, 26(4):557–580, 1996.

[297] Wevosys home english. `https://www.wevosys.com/`. Accessed: 26-06-2023.

[298] Liza Wikarsa and Sherly Novianti Thahir. A text mining application of emotion classifications of twitter's users using naive bayes method. In *2015 1st International Conference on Wireless and Telematics (ICWT)*, pages 1–6. IEEE, 2015.

[299] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[300] Windows tool for speech analysis. `https://www.phon.ucl.ac.uk/resource/sfs/wasp/`. Accessed: 26-06-2023.

[301] Eddie Wong and Sridha Sridharan. Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489)*, pages 95–98. IEEE, 2001.

[302] Chung-Hsien Wu and Wei-Bin Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21, 2010.

[303] Chung-Hsien Wu, Jui-Feng Yeh, and Ze-Jing Chuang. Emotion perception and recognition from speech. In *Affective Information Processing*, pages 93–110. Springer, 2009.

[304] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[305] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In

Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pages 1480–1489, 2016.

[306] Paul Thomas Young. Feeling and emotion. In Benjamin Binem Wolman, editor, *Handbook of general psychology.* Englewood Cliffs, New Jersey, Prentiche-Hall, 1973.

[307] G. Udny Yule. On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 226(636-646):267–298, 1927.

[308] Jerrold H Zar. *Biostatistical analysis.* Pearson Education India, 1999.

[309] Semir Zeki and John Paul Romaya. Neural correlates of hate. *PloS one*, 3(10):e3556, 2008.

[310] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008.

[311] Zhaoyan Zhang. Mechanics of human voice production and control. *The Journal of the Acoustical Society of America*, 140(4):2614, 2016.

[312] Zhengyou Zhang, Michael Lyons, Michael Schuster, and Shigeru Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proceedings Third IEEE International Conference on Automatic face and gesture recognition*, pages 454–459. IEEE, 1998.

[313] Guojun Zhou, John HL Hansen, and James F Kaiser. Nonlinear feature based classification of speech under stress. *IEEE Transactions on speech and audio processing*, 9(3):201–216, 2001.

Appendices

Appendix A

List of abbreviations used for the explicit linguistic features

This appendix reports the list of the abbreviations for the explicit linguistic features described in Section 2.2.2. Features for lexical density that belong to De Mauro's "Vocabolario Fondamentale" [71] are valid only for Italian.

Raw text properties

- *char_per_tok*: number of characters per token;
- *n_tokens*: number of tokens per sentence;
- *n_sentences*: number of sentences;
- *token_per_clause*: number of tokens per proposition (average clause length).

Lexical density

- *in_dict*: percentage of tokens in the "Vocabolario Fondamentale";
- *in_dict_types*: percentage of types in the "Vocabolario Fondamentale";
- *in_FO*: percentage of tokens in the fundamental words of the "Vocabolario Fondamentale";
- *in_AD*: percentage of tokens in the high availability words of the "Vocabolario Fondamentale";
- *in_AU*: percentage of tokens in the high usage words of the "Vocabolario Fondamentale";

- *in_FO_types*: percentage of types in the fundamental words of the "Vocabolario Fondamentale";
- *in_AD_types*: percentage of types in the high availability words of the "Vocabolario Fondamentale";
- *in_AU_types*: percentage of types in the high usage words of the "Vocabolario Fondamentale";
- *ttr_form*: Type/Token Ratio calculated on forms;
- *ttr_lemmi*: Type/Token Ratio calculated on lemmas;
- *lexical_density*: ratio of content words over the total number of words in a text.

Morpho-syntactic information

- *cpos_dist_X*: distribution of grammatical categories;
- *verbs_mood_dist_X*: distribution of verbal moods;
- *verbs_tense_dist_X*: distribution of verbal tenses;
- *verbs_num_pers_dist_X*: distribution of verbal person and number.

Verbal predicate structure

- *avg_verb_edges*: average verbal arity;
- *verb_edges_dist*: verbal arity distribution;
- *verb_edges_freq*: frequency of verbal arity;
- *verbal_head*: number of verbal heads;
- *verbal_head_per_sent*: number of verbal heads per sentence;
- *verbal_root*: number of verbal roots;
- *verbal_root_per_sent*: number of verbal roots per sentence.

Distributions of syntactic relations

- *dep_dist_X*: distribution of dependency relations;
- *dep_freq_X*: frequency of dependency relations.

Global and local parsed tree structure

- *avg_links_len*: average length of syntactic links;
- *avg_max_depth*: average depth of syntactic trees;
- *avg_prepositional_chain_len*: average length of prepositional chains;
- *max_links_len*: length of the longest syntactic link;
- *n_prepositional_chains*: number of embedded complement chains governed by a nominal head;
- *obj_post*: percentage of post-verbal objects;
- *obj_pre*: percentage of pre-verbal objects;
- *prep_dist_X*: distribution of prepositions;
- *prep_freq_X*: frequency of prepositions;
- *prepositional_chain_distribution_X*: distribution of prepositional chains per their length;
- *prepositional_chain_freq_x*: frequency of prepositional chains per their length;
- *subj_post*: percentage of post-verbal subjects;
- *subj_pre*: percentage of pre-verbal subjects.

Use of subordination

- *avg_subordinate_chain_len*: average length of chains of subordinate propositions;
- *n_subordinate_chain*: total length of chains of subordinate propositions;
- *n_subordinate_proposition*: number of subordinate propositions;
- *principal_proposition_dist*: distribution of principal propositions;
- *subordinate_dist_X*: distribution of subordinate propositions;
- *subordinate_freq_X*: frequency of subordinate propositions;
- *subordinate_post*: number of subordinate proposition after the principal proposition;
- *subordiante_pre*: number of subordinate proposition begore the principal proposition;

- *subordinate_proposition_dist*: distribution of subordinate propositions;
- *total_subordinate_chain_len*: total lenght of chains of subordinate propositions.

Appendix B

Complete results of the correlations between Linguistic Features and Complexity Average Scores and between Linguistic Features and Complexity Standard Deviation

In this Appendix are shown the complete results for the correlations between Linguistic Features and Perceived Complexity in context as discussed in Section 4.4.2. More specifically, the Tables reported in this Appendix show the correlations between the values of the Linguistic Features and the Complexity Average Scores (marked as *judg*) and the Standard Deviation of the Complexity Scores (marked as *std*).

Table B.1 shows the correlations between Linguistic Features and Perceived Complexity for the *Begin Context Window*; Table B.2 shows the correlations between Linguistic Features and Perceived Complexity for the *Center Context Window*; Table B.3 shows the correlations between Linguistic Features and Perceived Complexity for the *End Context Window.*

Correlations between Linguistic Features and Complexity Average Scores (*judg*) and between Linguistic Features and Complexity Standard Deviation (*std*) for the *Begin Context Window.*

Table B.1: Values of correlation for statistically significant (p-value$<$ 0.05) linguistic features with $\rho \geq 0.20$ that correlate with either the average judgment of complexity or the complexity standard deviation. For the *begin context window*, for all sentences and for sentences divided according to their length.

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_aux_+	−0.20	·	·	·	·	·	·	·	·	·	·	·	·	·
B_aux_Fin	−0.29	·	·	·	−0.25	·	·	0.22	·	·	·	·	·	·
B_aux_Ind	−0.27	·	·	·	·	·	·	·	·	·	·	·	·	·
B_avg_link	0.31	·	·	·	·	·	·	·	·	·	·	·	0.21	·
B_avg_max_depth	0.25	·	0.23	·	·	·	·	·	·	·	·	·	0.29	·
B_avg_max_link	0.36	·	·	·	·	·	·	·	·	·	·	·	0.26	·
B_avg_prep_chain	-	·	·	·	·	·	·	·	·	0.20	·	·	·	·
B_avg_sub_chain	-	·	·	·	·	·	−0.23	·	·	·	·	·	·	·
B_avg_tok_clause	−0.20	·	0.25	·	·	·	·	·	·	·	·	·	·	·
B_char_tok	·	·	−0.24	·	·	·	·	·	·	·	·	·	·	·
B_dep_advmod	·	·	·	·	0.20	·	·	·	·	·	·	·	·	·
B_dep_amod	−0.25	·	·	·	·	·	·	·	·	·	·	·	·	·
B_dep_appos	0.54	·	·	·	·	·	·	·	·	·	·	·	·	·
B_dep_compound	0.27	·	·	·	·	−0.22	·	·	0.20	·	0.21	·	0.22	·
B_dep_cop	−0.25	·	·	·	·	·	·	·	·	·	·	·	·	·
B_dep_det	−0.33	·	·	·	·	·	·	·	·	·	·	·	−0.21	·
B_dep_nsubj	−0.43	·	·	·	·	·	·	·	·	·	·	·	−0.22	·
B_dep_nummod	0.39	·	0.20	·	0.30	·	0.23	·	0.33	·	0.35	·	0.33	·
B_dep_obl:tmod	·	·	·	·	·	·	·	·	·	−0.26	·	·	·	·
B_dep_punct	·	·	·	·	0.20	·	·	·	·	·	·	·	·	·
B_dep_root	−0.34	·	−0.33	·	·	·	·	·	·	·	·	·	−0.32	·
B_dep_xcomp	·	·	·	·	·	·	−0.25	·	·	·	·	·	·	·
B_lexical_dens	·	·	·	·	−0.22	·	−0.21	·	·	·	·	·	·	·

continued on next page

continued from previous page

linguistic features	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
	judg	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_max_link	0.36	·	·	·	·	·	·	·	·	·	·	·	0.26	·
B_n_prep_chain	·	·	·	·	·	·	0.20	·	·	·	·	·	0.24	·
B_n_tok	0.34	·	0.33	·	·	·	·	·	·	·	·	·	0.32	·
B_obj_post	·	·	0.22	·	·	·	·	·	·	·	·	·	·	·
B_princ_prop	−0.27	·	·	·	·	·	·	·	·	·	·	·	·	·
B_sub_1	−0.24	·	·	·	·	·	·	·	·	·	·	·	·	·
B_sub_prop	·	·	·	·	·	·	−0.22	·	·	·	·	·	·	·
B_subj_pre	−0.42	·	·	·	·	·	·	·	·	·	·	·	·	·
B_tok_sent	0.34	·	0.33	·	·	·	·	·	·	·	·	·	0.32	·
B_ttr	−0.20	·	·	·	·	·	·	·	·	·	·	·	·	·
B_ttr_lemma	−0.21	·	·	·	·	·	·	−0.20	·	·	·	·	·	·
B_upos_ADJ	−0.26	·	·	·	−0.24	·	·	·	·	·	·	·	·	·
B_upos_ADP	−0.20	·	·	·	·	·	·	·	·	·	·	·	·	·
B_upos_AUX	−0.29	·	·	·	·	·	·	0.23	·	·	·	·	·	·
B_upos_DET	−0.33	·	·	·	·	·	·	·	·	·	·	·	−0.21	·
B_upos_NUM	0.40	·	0.30	·	0.33	·	0.30	·	0.34	·	0.30	·	0.34	·
B_upos_PART	·	·	·	·	·	·	·	·	·	·	−0.20	·	·	·
B_upos_PRON	−0.25	·	−0.24	·	·	·	·	·	·	·	·	·	·	·
B_upos_PUNCT	·	·	·	·	0.20	·	·	·	·	·	·	·	·	·
B_upos_SYM	0.30	·	0.22	·	·	·	0.27	·	0.29	·	0.31	·	0.28	·
B_upos_VERB	−0.30	·	·	·	·	·	·	·	·	·	·	·	·	·
B_verb_edge_0	·	·	−0.25	·	·	·	·	·	·	·	·	·	·	·
B_verb_head_sent	−0.42	·	·	·	·	·	−0.21	·	·	·	·	·	·	·
B_verb_root_perc	−0.43	−0.22	·	·	·	·	·	·	·	·	·	·	·	·
C_aux_+	·	·	·	·	·	−0.21	·	·	·	·	·	·	·	·
C_aux_Fin	−0.31	·	·	·	−0.21	·	·	·	·	·	·	·	·	·
C_aux_Ind	−0.32	·	·	·	·	·	·	·	·	·	·	·	·	·
C_aux_Pres	−0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
C_aux_Sing+3	−0.29	·	·	·	·	·	·	·	·	·	·	·	·	·
C_avg_link	−0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
C_avg_sub_chain	−0.24	·	·	·	·	·	·	·	·	·	·	·	·	·
C_avg_tok_clause	−0.33	·	·	·	·	·	·	·	·	·	·	·	·	·
C_avg_verb_edge	−0.30	·	·	·	−0.21	·	·	·	·	·	·	·	·	·

continued on next page

continued from previous page

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
C_char_tok	0.28	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_aux	·	·	·	·	−0.21	·	·	·	·	·	·	·	·	·
C_dep_aux:pass	·	·	·	·	·	·	·	·	·	0.23	·	·	·	·
C_dep_cc	−0.23	·	·	·	−0.20	·	·	·	·	·	·	·	·	·
C_dep_ccomp	−0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_compound	0.21	·	·	·	0.24	·	·	·	·	·	·	·	·	·
C_dep_nmod:poss	−0.24	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_nsubj	·	·	·	·	−0.31	·	·	·	·	·	·	·	·	·
C_dep_nsubj:pass	·	·	·	·	·	·	·	·	·	0.23	·	·	·	·
C_dep_nummod	·	·	·	·	0.28	·	0.27	·	0.22	·	0.26	·	0.23	·
C_dep_obj	−0.21	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_obl	−0.25	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_root	0.30	·	·	·	·	·	·	·	·	·	·	·	·	·
C_n_tok	−0.30	·	·	·	·	·	·	·	·	·	·	·	·	·
C_obj_post	−0.24	·	·	·	−0.21	·	·	·	·	·	·	·	·	·
C_prep_3	0.37	·	·	·	·	·	·	·	·	·	0.22	·	·	·
C _princ_prop	−0.27	·	·	·	·	·	·	·	·	·	·	·	·	·
C_sub_1	−0.30	·	·	·	·	·	·	·	·	·	·	·	·	·
C_sub_post	−0.28	·	·	·	·	·	·	·	·	·	·	·	·	·
C_sub_pre	·	·	·	·	−0.24	·	·	·	·	·	·	·	·	·
C_sub_prop	−0.22	·	·	·	·	·	·	·	·	·	·	·	·	·
C_subj_pre	−0.39	·	·	·	·	·	·	·	·	·	·	·	·	·
C_tok_sent	−0.30	·	·	·	·	·	·	·	·	·	·	·	·	·
C_upos_AUX	−0.22	·	·	·	−0.23	·	·	·	·	·	·	·	·	·
C_upos_CCONJ	−0.21	·	·	·	·	·	·	·	·	·	·	·	·	·
C_upos_DET	−0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
C_upos_NUM	·	·	·	·	0.24	·	0.35	·	0.23	·	0.26	·	0.24	·
C_upos_PART	−0.25	·	·	·	·	·	·	·	·	·	·	·	·	·
C_upos_PRON	−0.27	·	−0.23	·	·	·	·	·	·	·	·	·	·	·
C_upos_VERB	−0.29	·	·	·	−0.27	·	·	·	·	·	·	·	·	·
C_verb_edge_5	−0.31	·	·	·	·	·	·	·	·	·	·	·	·	·
C_verb_head_sent	−0.38	·	·	·	−0.28	·	·	·	·	·	·	·	·	·
C_verb_Ind	−0.23	·	·	·	·	·	·	·	·	·	·	·	·	·

continued on next page

continued from previous page

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
C_verb_Inf	−0.21	·	·	·	·	−0.20	·	·	·	·	·	·	·	·
C_verb_Part	−0.24	·	·	·	·	·	·	·	·	·	·	·	·	·
C_verb_Past	−0.29	·	·	·	·	·	·	·	·	·	·	·	·	·
C_verb_Pres	·	·	·	·	−0.26	·	·	·	·	·	·	·	·	·
C_verb_root_perc	−0.44	·	·	·	−0.29	·	·	·	·	·	·	·	·	·
C_verb_Sing+3	·	·	·	·	−0.25	·	·	·	·	·	·	·	·	·
E_aux_Fin	−0.29	·	·	·	−0.21	·	·	·	·	·	·	·	·	·
E_aux_Ind	−0.22	·	·	·	−0.22	·	·	·	·	0.22	·	·	·	·
E_aux_Pres	·	·	·	·	−0.24	·	−0.21	·	·	·	·	·	·	·
E_avg_link	0.32	·	·	·	0.31	·	·	·	·	·	·	·	·	·
E_avg_max_link	-	·	·	·	0.29	·	·	·	·	·	·	·	·	·
E_avg_prep_chain	0.26	·	·	·	·	·	·	·	·	·	0.23	·	·	·
E_avg_verb_edge	−0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
E_dep_advmod	·	·	−0.25	−0.24	·	·	·	·	·	·	·	·	·	·
E_dep_appos	0.37	·	·	·	0.21	·	0.23	·	·	·	·	·	·	·
E_dep_det	−0.25	·	·	·	·	·	·	·	·	·	·	·	·	·
E_dep_list	·	·	·	·	0.22	·	·	·	·	·	·	·	·	·
E_dep_nmod	0.35	·	·	·	·	·	·	·	·	·	0.25	·	·	·
E_dep_nsubj	−0.29	·	·	·	−0.25	·	·	·	·	·	·	·	·	·
E_dep_nummod	0.40	·	·	·	·	·	·	·	·	·	·	·	0.21	·
E_dep_obj	−0.31	·	·	·	·	·	·	·	·	·	·	·	·	·
E_lexical_dens	−0.30	·	·	·	·	·	·	·	·	·	·	·	·	·
E_max_link	-	·	·	·	0.29	·	·	·	·	·	·	·	·	·
E_n_prep_chain	0.32	·	·	·	0.22	·	·	·	·	·	0.20	·	·	·
E_obj_post	−0.28	·	·	·	·	·	·	·	·	·	·	·	·	·
E_prep_1	0.29	·	·	·	·	·	·	·	·	·	·	·	·	·
E _princ_prop	−0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
E_sub_pre	·	−0.21	·	·	·	·	·	·	·	·	·	·	·	·
E_subj_pre	−0.45	·	·	·	·	·	·	·	·	·	·	·	·	·
E_ttr	−0.31	·	·	·	−0.29	·	·	·	·	·	·	·	·	·
E_ttr_lemma	−0.30	·	·	·	−0.29	·	·	·	·	·	·	·	·	·
E_upos_ADP	0.22	·	·	·	·	·	·	·	·	·	·	·	·	·
E_upos_AUX	−0.24	·	·	·	−0.27	·	·	·	·	·	·	·	·	·

continued on next page

continued from previous page

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
E_upos_DET	−0.27	·	·	·	·	·	·	·	·	·	·	·	·	·
E_upos_NUM	0.39	·	·	·	0.23	·	·	·	·	·	0.21	·	0.23	·
E_upos_PRON	−0.34	·	·	·	·	·	·	·	·	·	·	·	·	·
E_upos_VERB	−0.38	·	·	·	·	·	·	·	·	·	−0.24	·	·	·
E_verb_edge_1	−0.29	·	·	·	·	·	·	·	·	·	·	·	·	·
E_verb_edge_3	−0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
E_verb_Ger	·	·	·	·	0.20	·	·	·	·	·	·	·	·	·
E_verb_head_sent	−0.30	·	·	·	·	·	·	·	·	·	·	·	·	·
E_verb_root_perc	−0.41	·	·	·	−0.21	·	·	·	·	·	·	·	·	·

Correlations between Linguistic Features and Complexity Average Scores (*judg*) and between Linguistic Features and Complexity Standard Deviation (*std*) for the *Center Context Window*.

Table B.2: Values of correlation for statistically significant (p-value$<$ 0.05) linguistic features with $\rho \geq 0.20$ that correlate with either the average judgment of complexity or the complexity standard deviation. For the *center context window*, for all sentences and for sentences divided according to their length.

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_aux_+	−0.21	·	·	·	·	·	·	·	·	·	·	·	·	·
B_aux_form_Ger	·	·	0.21	·	·	·	·	·	·	·	·	·	·	·
B_aux_form_Inf	·	0.20	·	·	·	·	·	·	·	·	·	·	·	·
B_aux_Pres	·	·	·	·	·	·	·	·	−0.21	·	·	·	·	·
B_avg_prep_chain	-	·	·	·	·	·	0.23	·	·	·	·	·	·	·
B_avg_sub_chain	−0.24	·	·	·	·	·	·	·	·	·	·	·	·	·
B_dep_aux	·	·	·	·	·	·	·	·	·	·	−0.28	·	·	·
B_dep_aux:pass	·	·	·	·	·	·	−0.32	·	·	·	·	·	·	·
B_dep_compound	·	·	0.20	·	0.21	·	·	·	·	·	0.22	·	0.21	·
B_dep_flat	·	·	·	·	·	·	−0.22	·	·	·	·	·	·	·
B_dep_nmod	·	·	·	·	·	·	0.25	·	·	·	·	·	·	·
B_dep_nsubj	·	·	−0.24	·	·	·	·	·	·	·	−0.21	·	·	·
B_dep_nsubj:pass	·	·	·	·	·	·	−0.29	·	·	·	·	·	·	·
B_dep_nummod	·	·	0.27	·	0.23	·	·	·	·	·	0.26	·	·	·
B_n_prep_chain	·	·	·	·	·	·	0.23	·	·	·	·	·	·	·
B_princ_prop	·	·	·	·	−0.24	·	·	·	·	·	·	·	·	·
B_sub_post	−0.22	·	·	·	·	·	·	·	·	·	·	·	·	·
B_subj_pre	·	·	·	·	·	·	·	−0.20	·	·	·	·	·	·
B_upos_NUM	·	·	0.22	·	·	·	·	·	·	·	0.29	·	·	·
B_upos_PRON	·	·	·	·	·	·	·	·	−0.22	·	·	·	·	·
B_upos_PROPN	·	·	0.21	·	·	·	·	·	·	·	·	·	·	·
B_upos_SYM	·	·	·	·	·	·	·	−0.21	·	·	0.21	·	·	·
B_upos_VERB	·	·	−0.21	·	−0.22	·	·	·	·	·	·	·	·	·

continued on next page

continued from previous page

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_verb_edge_1	·	·	·	−0.20	·	·	·	·	·	·	·	·	·	·
B_verb_Past	·	·	·	−0.21	·	·	·	·	·	·	·	·	·	·
B_verb_root_perc	·	·	·	·	−0.25	·	·	·	·	·	·	·	·	·
C_aux_+	−0.24	·	·	·	−0.24	·	·	·	·	·	−0.20	·	·	·
C_aux_form_Fin	−0.21	·	·	·	−0.28	·	·	·	−0.22	·	·	·	·	·
C_aux_Ind	·	·	·	·	−0.23	·	·	·	·	·	·	·	·	·
C_aux_Past	·	·	·	·	−0.21	·	·	·	·	·	·	·	·	·
C_aux_Pres	−0.21	·	·	·	·	·	·	·	−0.22	·	−0.24	·	·	·
C_avg_max_depth	0.21	·	0.31	·	·	·	·	·	·	·	·	·	0.29	·
C_avg_max_link	-	·	·	·	·	·	·	·	·	·	·	·	0.25	·
C_avg_sub_chain	-	·	·	·	·	·	·	·	−0.20	·	−0.38	·	·	·
C_avg_tok_clause	·	·	·	·	0.22	·	·	·	·	·	0.26	·	·	·
C_char_tok	·	·	·	·	·	·	·	·	·	·	−0.30	·	·	·
C_dep_amod	·	·	·	·	−0.24	·	·	·	·	·	·	·	·	·
C_dep_aux	·	·	−0.22	·	−0.21	·	·	·	−0.28	−0.25	−0.29	·	·	·
C_dep_case	·	·	·	·	·	·	·	·	·	·	0.22	·	·	·
C_dep_ccomp	·	·	·	·	·	·	·	·	·	·	−0.27	·	·	·
C_dep_det	·	·	−0.28	·	−0.22	·	·	·	·	·	·	·	·	·
C_dep_mark	·	·	·	·	·	·	·	·	·	·	−0.23	·	·	·
C_dep_nmod	0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_nsubj	−0.22	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_nummod	0.21	·	0.32	·	0.25	·	·	·	0.26	·	0.35	·	0.29	·
C_dep_punct	·	·	·	·	·	·	·	−0.21	·	·	·	·	·	·
C_dep_root	−0.24	·	−0.33	·	·	·	·	·	·	·	·	·	−0.31	·
C_dep_xcomp	·	·	·	·	·	·	−0.26	·	·	·	−0.28	·	·	·
C_lexical_dens	·	·	−0.23	·	−0.21	·	·	·	·	·	−0.27	·	·	·
C_max_link	-	·	·	·	·	·	·	·	·	·	·	·	0.25	·
C_n_prep_chain	0.23	·	·	·	·	·	·	·	·	·	·	·	0.25	·
C_n_tok	0.24	·	0.33	·	·	·	·	·	·	·	·	·	0.31	·
C_princ_prop	·	·	·	·	·	·	·	·	·	·	0.23	−0.21	·	·
C_sub_2	·	·	·	·	·	·	·	·	·	·	−0.20	·	·	·
C_sub_4	·	·	·	·	·	·	·	·	·	·	−0.24	·	·	·
C_sub_post	·	·	·	·	·	·	·	·	·	·	−0.28	·	·	·

continued on next page

continued from previous page

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
C_sub_prop	·	·	·	·	·	·	·	·	·	·	−0.29	0.20	·	·
C_tok_sent	0.24	·	0.33	·	·	·	·	·	·	·	·	·	0.31	·
C_upos_ADJ	−0.21	·	−0.25	·	−0.22	·	·	·	·	·	−0.26	·	·	·
C_upos_AUX	−0.24	·	·	·	−0.27	·	·	·	−0.23	·	−0.32	·	−0.23	·
C_upos_DET	·	·	−0.28	·	−0.21	·	·	·	·	·	·	·	·	·
C_upos_NUM	0.30	·	0.41	·	0.31	·	·	·	0.28	·	0.39	·	0.33	·
C_upos_PRON	−0.21	·	·	·	−0.21	·	·	·	·	·	−0.31	·	·	·
C_upos_PUNCT	·	·	·	·	·	·	·	−0.21	·	·	·	·	·	·
C_upos_SYM	0.26	·	0.30	·	·	·	·	·	·	·	0.34	·	0.24	·
C_upos_VERB	·	·	·	·	·	·	·	·	·	·	−0.24	·	·	·
C_verb_+	·	·	·	·	·	·	0.22	·	·	·	·	·	·	·
C_verb_edge_1	·	·	·	·	·	·	·	·	−0.28	·	·	0.20	·	·
C_verb_edge_2	·	·	·	·	·	·	·	·	·	·	−0.26	·	·	·
C_verb_form_Fin	·	·	·	·	·	·	·	·	0.24	·	·	·	·	·
C_verb_form_Inf	·	·	·	·	·	·	·	·	·	·	−0.27	·	·	·
C_verb_head_sent	·	·	·	·	−0.23	·	·	·	·	·	−0.28	·	·	·
C_verb_Ind	·	·	·	·	·	·	·	·	0.21	·	·	·	·	·
C_verb_root_perc	·	·	·	·	·	·	·	·	·	·	·	−0.22	·	·
E_aux_Pres	−0.27	·	−0.21	·	·	·	·	·	·	·	·	·	·	·
E_avg_link	−0.29	−0.23	·	·	·	·	·	·	·	·	·	·	·	·
E_avg_max_depth	·	·	·	·	·	·	0.30	·	·	·	·	·	·	·
E_avg_max_link	−0.23	−0.25	·	·	·	·	·	·	·	·	·	·	·	·
E_avg_sub_chain	−0.21	·	·	·	·	·	·	·	·	·	−0.26	·	·	·
E_avg_tok_clause	·	·	·	·	·	·	·	·	·	·	0.25	·	·	·
E_avg_verb_edge	−0.21	·	·	·	·	·	·	·	·	·	·	·	·	·
E_dep_advmod	−0.24	·	−0.20	·	·	·	·	·	·	·	·	·	·	·
E_dep_aux	·	−0.22	·	·	·	·	·	·	·	·	·	·	·	·
E_dep_case	·	·	·	·	·	·	0.20	·	·	·	·	·	·	·
E_dep_ccomp	·	·	·	·	·	·	·	·	·	·	−0.31	·	·	·
E_dep_nummod	·	·	·	·	·	·	0.28	·	·	·	0.33	·	0.22	·
E_dep_parataxis	·	−0.22	·	·	·	·	·	·	·	·	·	·	·	·
E_dep_root	0.21	0.21	·	·	·	·	·	·	·	·	·	·	·	·
E_dep_xcomp	·	−0.21	−0.23	·	·	·	·	·	·	·	·	·	·	·

continued on next page

continued from previous page

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
E_lexical_dens	·	·	·	·	·	·	−0.25	·	·	·	−0.22	·	·	·
E_max_link	−0.23	−0.25	·	·	·	·	·	·	·	·	·	·	·	·
E_n_tok	−0.21	−0.21	·	·	·	·	·	·	·	·	·	·	·	·
E_prep_1	·	·	·	·	−0.22	·	·	·	·	·	·	·	·	·
E_prep_2	·	−0.20	·	·	·	·	0.20	·	·	·	·	·	·	·
E_sub_post	·	·	·	·	·	·	·	·	·	·	−0.22	·	·	·
E_sub_pre	·	·	−0.22	·	·	·	·	·	·	·	·	·	·	·
E_sub_prop	−0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
E_tok_sent	−0.21	−0.21	·	·	·	·	·	·	·	·	·	·	·	·
E_upos_ADV	·	·	−0.23	·	·	0.22	·	·	·	·	·	·	·	·
E_upos_NUM	·	·	·	·	·	·	0.33	·	·	·	0.34	·	0.22	·
E_upos_PART	·	·	·	·	·	·	·	·	·	·	−0.23	·	·	·
E_upos_PRON	−0.22	·	·	·	·	·	·	·	·	·	·	·	·	·
E_upos_SYM	·	·	·	·	·	·	0.28	·	·	·	0.30	·	·	·
E_upos_VERB	·	·	·	·	·	·	·	·	·	·	−0.21	·	·	·
E_verb_edge_3	·	·	·	·	·	·	·	−0.22	·	·	·	0.21	·	·
E_verb_edge_4	−0.30	·	·	·	·	·	·	·	·	·	·	·	·	·
E_verb_form_Fin	·	·	·	·	·	·	·	·	0.21	·	·	·	·	·
E_verb_form_Inf	·	·	·	·	·	·	·	·	·	·	−0.22	·	·	·
E_verb_head_sent	−0.24	·	·	·	·	·	·	·	·	·	·	·	·	·
E_verb_Past	·	0.20	·	·	·	·	0.23	·	·	·	·	·	·	·
E_verb_Pres	−0.20	−0.30	·	·	·	·	·	·	·	·	·	·	·	·
E_verb_Sing+3	−0.20	−0.21	·	·	·	·	·	·	·	·	·	·	·	·

Correlations between Linguistic Features and Complexity Average Scores (*judg*) and between Linguistic Features and Complexity Standard Deviation (*std*) for the *End Context Window.*

Table B.3: Values of correlation for statistically significant (p-value$<$ 0.05) linguistic features with $\rho \geq 0.20$ that correlate with either the average judgment of complexity or the complexity standard deviation. For the *end context window*, for all sentences and for sentences divided according to their length.

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_aux_Fin	·	·	·	·	−0.23	·	·	·	·	·	·	·	·	·
B_aux_Ind	·	·	·	·	−0.21	·	·	·	·	·	·	·	·	·
B_avg_link	-	·	·	·	−0.25	·	·	·	·	·	·	·	·	·
B_avg_max_link	-	·	·	·	−0.24	·	·	·	·	·	·	·	·	·
B_dep_acl	·	·	·	·	·	−0.23	·	·	·	·	·	·	·	·
B_dep_advcl	·	·	·	·	·	·	·	·	−0.20	·	·	·	·	·
B_dep_case	·	·	·	·	−0.21	·	·	·	·	·	·	·	·	·
B_dep_ccomp	·	·	·	·	·	·	·	·	·	·	−0.21	·	·	·
B_dep_nmod:poss	·	·	·	·	·	·	·	−0.22	·	·	·	·	·	·
B_dep_obj	·	·	−0.25	·	·	·	·	·	·	·	·	·	·	·
B_dep_obl	·	·	·	·	−0.26	·	·	·	·	·	·	·	·	·
B_dep_xcomp	·	·	−0.21	·	·	·	·	·	·	·	·	·	·	·
B_max_link	-	·	·	·	−0.24	·	·	·	·	·	·	·	·	·
B_prep_3	·	·	·	·	·	·	·	−0.20	·	·	·	·	·	·
B_sub_1	·	·	·	·	−0.23	·	·	−0.25	·	·	·	·	·	·
B_subj_pre	·	·	·	·	−0.21	·	·	·	·	·	·	·	·	·
B_ttr	·	·	·	·	−0.25	·	−0.20	·	·	·	·	·	·	·
B_ttr_lemma	·	·	·	·	−0.22	·	−0.22	·	·	·	·	·	·	·
B_upos_ADP	·	·	·	·	−0.23	·	·	·	·	·	·	·	·	·
B_upos_AUX	·	·	·	·	−0.26	·	·	·	·	·	·	·	·	·
B_upos_NOUN	·	·	·	·	·	·	·	·	·	·	0.22	·	·	·
B_upos_SYM	0.23	·	·	·	·	·	·	·	·	·	·	·	·	·
B_upos_VERB	·	·	·	·	−0.23	·	·	·	−0.24	·	·	·	·	·

continued on next page

continued from previous page

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
B_verb_head_sent	·	·	·	·	−0.21	·	·	·	·	·	·	·	·	·
B_verb_Part	·	·	·	·	−0.26	·	·	·	·	·	·	·	·	·
B_verb_root_perc	·	·	·	·	·	·	·	·	·	·	·	−0.20	·	·
C_aux_Fin	−0.21	·	−0.21	·	−0.26	·	·	·	·	·	·	·	·	·
C_char_tok	·	·	·	·	·	·	·	·	·	·	−0.20	·	·	·
C_dep_appos	·	·	·	·	0.26	·	·	·	·	·	·	·	·	·
C_dep_aux	·	·	−0.27	·	·	·	·	·	·	·	·	·	·	·
C_dep_case	0.22	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_compound	·	·	0.22	·	0.22	·	·	·	·	·	·	·	·	·
C_dep_det	−0.21	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_fixed	·	·	·	·	·	−0.21	·	·	·	·	·	·	·	·
C_dep_nmod	0.22	·	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_nsubj	·	·	·	·	−0.20	·	·	·	·	·	·	·	·	·
C_dep_nummod	·	·	·	·	0.26	·	·	·	0.20	·	·	·	·	·
C_dep_obl	·	0.20	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_obl:tmod	·	−0.25	·	·	·	·	·	·	·	·	·	·	·	·
C_dep_punct	·	·	·	·	0.23	·	·	·	·	·	·	·	·	·
C_sub_2	·	0.22	·	·	·	·	·	·	·	·	·	·	·	·
C_sub_post	·	0.27	·	·	·	·	·	·	·	·	·	·	·	·
C_sub_pre	−0.22	−0.23	·	·	·	·	·	·	·	·	·	·	·	·
C_sub_prop	·	0.22	·	·	·	·	·	·	·	·	·	·	·	·
C_subj_pre	·	·	·	·	−0.27	·	·	·	·	·	·	·	·	·
C_ttr	·	·	·	·	−0.24	·	·	·	·	·	·	0.23	·	·
C_ttr_lemma	·	·	0.22	·	−0.27	·	·	·	·	·	·	0.22	·	·
C_upos_AUX	−0.25	·	−0.21	·	−0.21	·	·	·	·	·	·	·	·	·
C_upos_DET	−0.23	·	·	·	·	·	·	·	−0.22	·	·	·	·	·
C_upos_NUM	·	·	·	·	0.26	·	·	·	0.21	·	·	·	·	·
C_upos_PRON	·	·	−0.21	·	·	·	·	·	·	·	·	·	·	·
C_upos_PROPN	·	·	0.25	·	·	·	·	·	·	·	·	·	·	·
C_upos_PUNCT	·	·	·	·	0.23	·	·	·	·	·	·	·	·	·
C_upos_SYM	·	·	·	·	·	·	·	·	·	·	0.26	·	·	·
C_verb_Past	·	·	0.28	·	·	·	·	·	·	·	0.23	·	·	·
C_verb_Pres	·	·	−0.20	·	·	·	·	·	·	·	·	·	·	·

continued on next page

continued from previous page

linguistic features	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
	judg	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
C_verb_root_perc	·	·	·	·	−0.28	·	·	·	·	·	·	·	·	·
E_aux_Fin	·	·	·	·	−0.23	·	·	·	·	·	·	·	·	·
E_aux_Inf	·	·	·	·	·	·	·	·	·	·	−0.25	·	·	·
E_aux_Pres	−0.20	·	·	·	·	·	·	·	−0.21	·	·	·	·	·
E_avg_link	-	·	·	·	·	·	·	·	·	·	·	·	0.24	·
E_avg_max_depth	0.21	·	0.22	·	·	·	·	·	·	·	·	·	0.27	·
E_avg_max_link	-	·	·	·	·	·	·	·	·	·	·	·	0.28	·
E_avg_sub_chain	-	·	·	·	·	·	·	·	·	·	−0.28	·	·	·
E_avg_tok_clause	·	·	·	·	·	·	·	·	0.20	·	·	·	·	·
E_avg_verb_edge	−0.28	−0.21	·	·	·	·	·	·	·	·	·	·	·	·
E_char_tok	·	·	·	·	·	·	−0.22	·	·	·	·	·	·	·
E_dep_acl:relcl	·	·	·	·	·	·	0.21	·	·	·	·	·	·	·
E_dep_advcl	·	·	·	·	·	·	·	·	−0.20	·	·	·	·	·
E_dep_advmod	·	·	−0.23	·	·	·	·	·	·	·	·	·	·	·
E_dep_amod	·	·	−0.23	·	·	·	·	·	·	·	·	·	·	·
E_dep_appos	0.28	·	·	·	·	·	·	·	·	0.23	·	·	·	·
E_dep_aux	·	·	·	·	·	·	·	·	·	·	−0.32	·	·	·
E_dep_compound	0.20	·	0.27	·	·	·	·	·	0.22	·	·	·	0.21	·
E_dep_det	·	·	−0.30	·	−0.33	·	·	·	·	·	·	·	·	·
E_dep_mark	·	·	·	·	·	·	·	·	·	·	−0.29	·	·	·
E_dep_nmod	0.20	·	·	·	·	·	·	·	·	·	·	·	·	·
E_dep_nsubj	·	·	·	·	·	·	·	·	·	·	·	·	−0.21	·
E_dep_nummod	·	·	·	·	0.27	·	0.23	·	0.21	·	0.25	·	0.22	·
E_dep_obj	·	−0.22	·	·	·	·	·	·	·	·	·	·	·	·
E_dep_obl	·	·	·	·	·	·	·	·	·	·	−0.27	·	·	·
E_dep_parataxis	·	·	·	·	·	·	0.22	·	·	·	·	·	·	·
E_dep_punct	·	·	·	·	0.22	·	·	·	·	·	·	·	·	·
E_dep_root	·	·	−0.33	·	·	·	·	·	·	·	·	·	−0.33	·
E_lexical_dens	·	·	·	·	·	·	−0.29	·	·	·	·	·	·	·
E_max_link	-	·	·	·	·	·	·	·	·	·	·	·	0.28	·
E_n_tok	·	·	0.33	·	·	·	·	·	·	·	·	·	0.33	·
E_obj_post	·	−0.23	·	·	·	·	·	·	·	·	·	·	·	·
E_sub_2	·	·	·	·	·	·	−0.21	·	·	·	−0.23	·	·	·

continued on next page

continued from previous page

linguistic	**length 10**		**length 15**		**length 20**		**length 25**		**length 30**		**length 35**		**all sents**	
features	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*	*judg*	*std*
E_sub_post	·	·	·	·	·	·	·	·	·	·	−0.25	·	·	·
E_subj_pre	−0.32	−0.23	·	·	·	·	·	·	·	·	·	·	·	·
E_tok_sent	·	·	0.33	·	·	·	·	·	·	·	·	·	0.33	·
E_ttr	·	·	·	·	−0.22	·	−0.21	·	·	·	−0.23	·	−0.20	·
E_ttr_lemma	·	·	·	·	−0.22	·	·	·	·	·	−0.20	·	·	·
E_upos_ADV	−0.21	·	·	·	·	·	·	·	·	·	·	·	·	·
E_upos_AUX	·	·	·	·	−0.24	·	·	·	·	·	·	·	·	·
E_upos_DET	·	·	−0.30	·	−0.33	·	·	·	·	·	·	·	·	·
E_upos_NOUN	·	·	·	·	−0.25	·	·	·	·	·	·	·	·	·
E_upos_NUM	·	·	0.21	·	0.28	·	0.27	·	·	·	0.28	·	0.25	·
E_upos_PART	·	·	·	·	·	·	·	·	·	·	−0.23	·	·	·
E_upos_PRON	−0.22	·	·	·	·	·	·	·	−0.21	·	−0.24	·	·	·
E_upos_PROPN	·	·	·	·	·	·	·	·	·	0.24	·	·	·	·
E_upos_PUNCT	·	·	·	·	0.22	·	·	·	·	·	·	·	·	·
E_upos_SYM	·	−0.23	·	·	·	·	0.23	·	·	·	0.27	·	0.21	·
E_upos_VERB	·	·	·	·	·	·	·	·	−0.24	·	−0.25	·	·	·
E_verb_edge_2	·	·	·	·	·	·	·	·	·	·	−0.24	·	·	·
E_verb_edge_3	−0.24	·	·	·	·	·	·	·	·	·	·	·	·	·
E_verb_edge_6	·	·	·	·	·	−0.21	·	·	·	·	·	·	·	·
E_verb_Fin	·	·	·	·	0.22	·	·	·	·	·	·	·	·	·
E_verb_Ger	·	·	·	·	·	·	·	·	−0.25	·	·	·	·	·
E_verb_head_sent	·	·	·	·	−0.20	·	·	·	−0.20	·	·	·	·	·
E_verb_Inf	·	·	·	·	−0.23	·	·	·	·	·	−0.22	·	·	·
E_verb_Pres	−0.22	·	·	·	·	·	·	·	·	·	·	·	·	·
E_verb_root_perc	−0.20	·	·	·	·	·	·	·	·	·	·	·	·	·
E_verb_Sing+3	·	·	·	·	0.23	·	·	·	·	·	·	·	·	·

Appendix C

In this Appendix are reported the additional materials for Chapter 6.

Table C.1 shows the four Italian texts on which are based the analyses discussed in Chapter 6. As described in Section 6.2, the four texts are chosen to represent different levels of arousal and valence. In particular, two of the selected texts are chosen to have high arousal and negative values (Emotional text No. 1 and No. 2) as they provide a detailed and graphic description of medieval torture practices, while the other two are chosen to have low arousal and neural valence (Neutral text No. 1 and No. 2) and describe text types and writing styles. The Table shows the texts in their Italian version and with an English translation. For each text, the Table also reports the average sentence length, the average word length for each sentence, and the valence and arousal of the texts.

Table C.2 shows the complete results of the correlation analysis between Speech Features and Linguistic Features, while Table C.3 shows the complete results of the correlation analysis between Electrodermal Activity Features and Linguistic Features. In these Tables, for each pairwise correlation, each number in the rows corresponds to the percentage of subjects for which the correlation was statistically significant (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 and different from zero.

Stimuli

Table C.1: List of texts presented to the subjects as stimuli. The Table shows the original Italian version of the texts with their corresponding English translation. For each text is reported: its valence and arousal ratings, its average sentence length, and the average words length per sentence.

Neutral text No. 1	
	Average sentence length: 37.88 ± 23.13; Average words length per sentence: 5.05 ± 0.56 Valence Rate: 0.30 ± 0.52; Arousal Rate: 1.31 ± 0.67
Italian:	Il testo descrittivo ha lo scopo di informare il destinatario sulle caratteristiche (spaziali, temporali, sensoriali ecc.) di un determinato referente (oggetti, ambienti, persone, situazioni ecc.). A tal fine si ricorre opportunamente a un lessico ricco e preciso, a un'attenta selezione dei dati e a un'ordinata struttura del discorso. La sintassi ha frasi brevi ed essenziali. In taluni casi l'emittente può allegare anche le proprie valutazioni, opinioni o giudizi, al fine di stimolare le emozioni del lettore: si passa così da una descrizione oggettiva e distaccata a una descrizione soggettiva e valutativa. Tipici testi descrittivi sono, per esempio, le guide turistiche o le voci di enciclopedia. Descrivere un oggetto è un'operazione apparentemente semplice, che richiede, però, molta precisione nella selezione dei dati, nella disposizione delle osservazioni e nella scelta del lessico. Il testo narrativo presenta al lettore la dinamica di svolgimento di una determinata vicenda o serie di vicende, reali o immaginarie. Esso risponde a una domanda elementare, "che cosa è successo?", in vista di obietti o scopi differenti: per divertire, nel caso di un romanzo o racconto di intrattenimento; per trasmettere idee o valori, nel caso di un testo a forte impegno ideologico; per informare, nel caso di un articolo di cronaca, di un rapporto ufficiale al termine di una missione o dopo un incidente. Quando non persegue particolari finalità artistiche, la narrazione espone gli eventi secondo l'ordine cronologico della loro successione, dall'inizio alla conclusione (fabula); tuttavia, per colpire le emozioni del lettore, si possono disporre i fatti in un intreccio che non corrisponde alla successione logica e temporale, con fenomeni di suspense o flashback.
English:	The descriptive text aims to inform the recipient about the characteristics (spatial, temporal, sensory, etc.) of a given referent (objects, environments, people, situations, etc.). To this end, rich and precise vocabulary, careful selection of data, and orderly discourse structure are used appropriately. Syntax has short and essential sentences. In some cases the issuer may also attach his or her own evaluations, opinions or judgments, in order to stimulate the reader's emotions: thus moving from an objective and detached description to a subjective and evaluative one. Typical descriptive texts are, for example, travel guides or encyclopedia entries. Describing an object is a seemingly simple task, which requires, however, much precision in the selection of data, the arrangement of observations and the choice of vocabulary. Narrative text presents the reader with the unfolding dynamics of a given event or series of events, real or imaginary. It answers an elementary question, "what happened?" in view of different objectives or purposes: to entertain, in the case of a novel or entertainment story; to convey ideas or values, in the case of a text with a strong ideological commitment; to inform, in the case of a news article, an official report at the end of a mission or after an incident. When not pursuing any particular artistic purpose, the narrative sets out events in the chronological order of their succession, from beginning to conclusion (fabula); however, in order to strike the reader's emotions, facts may be arranged in a plot that does not correspond to logical and temporal succession, with suspense or flashback phenomena.

continued from previous page	
Neutral text No. 2	
	Average sentence length: 22.31 ± 7.33; Average words length per sentence: 4.85 ± 0.61 Valence Rate: 0.25 ± 0.55; Arousal Rate: 1.39 ± 0.77
Italian:	Scrivere un testo argomentativo significa esporre la propria idea, la propria tesi, in merito ad un determinato tema, basandola su dati certi e su un ragionamento logico dichiarato. Un testo argomentativo generalmente si pone l'obbiettivo di convincere chi lo legge della validità della tesi esposta. Prima di esporre la propria tesi, in un buon testo argomentativo, è bene introdurre il lettore al tema che si tratterà. Il nostro inizio, quindi, conterrà una descrizione generale dell'argomento, in cui vengono presentati alcuni dettagli e alcuni aspetti che possono essere utili ad avvalorare la nostra tesi. Questo primo passo è utile per far sentire a proprio agio il lettore durante le lettura del testo argomentativo. Per prima cosa esponi la tua tesi, chiarendo in modo dettagliato la tua opinione. Evita i periodi eccessivamente lunghi e complessi, disponi le frasi in ottica causa-effetto. Seguendo le buone norme del giornalismo, verifica che il tuo testo funziona se risponde alle domande chi/dove/quando/come/perché. In seguito, esponi tutti i dati e gli argomenti che sono a tuo favore, sottolineando il legame con la tua tesi. Accertati della veridicità dei dati e delle affermazioni che riporterai, ed esponi gli argomenti in modo tale che la logicità del tuo ragionamento sia chiara al lettore. In una fase successiva, il tuo testo argomentativo deve raccontare le idee e le opinioni opposte alla tua. Anche in questo caso devi essere preciso e dettagliato nell'esposizione. La validità della tua tesi deve basarsi su dati certi, non su un inganno intellettuale. Per discutere degli argomenti che contrastano la tua tesi e sostengono le tesi opposte, usa la tecnica del contrasto. Scomponi le tesi contrarie in più punti e confutali uno dopo l'altro, contrapponendogli i dati in tuo favore. Il ragionamento logico sarà così valorizzato.
English:	To write an argumentative text means to state one's idea, one's thesis, about a given topic, basing it on certain data and stated logical reasoning. An argumentative text generally aims to convince the reader of the validity of the thesis stated. Before stating one's thesis, in a good argumentative text, it is a good idea to introduce the reader to the topic that will be discussed. Our beginning, therefore, will contain a general description of the topic, in which some details and some aspects are presented that may be useful to substantiate our thesis. This first step is useful to make the reader feel comfortable while reading the argumentative text. First, state your thesis, clarifying your opinion in detail. Avoid excessively long and complex periods, arrange sentences in a cause-and-effect perspective. Following the good rules of journalism, verify that your text works if it answers the questions who/where/when/how/why. Next, lay out all the data and arguments that are in your favor, emphasizing the link to your thesis. Make sure that the data and statements you report are true, and state the arguments in such a way that the logicality of your argument is clear to the reader. At a later stage, your argumentative text must recount the ideas and opinions opposed to yours. Again, you must be precise and detailed in your exposition. The validity of your argument must be based on hard data, not intellectual deception. To discuss arguments that counter your thesis and support opposing theses, use the contrast technique. Break down the opposing theses into several points and refute them one after another, countering them with data in your favor. Logical reasoning will thus be enhanced.

continued from previous page

Emotional text No. 1	
	Average sentence length: 21.04 ± 10.37; Average words length per sentence: 4.94 ± 0.59 Valence Rate: -1.31 ± 0.79; Arousal Rate: 3.26 ± 1.17
Italian:	Una delle condanne a morte più crudeli mai esistite, ed ancora in vigore in alcune parti del mondo, è certamente la lapidazione. Questa tecnica di esecuzione è praticata principalmente nei paesi islamici ed è riservata agli adulteri di entrambi i sessi. La vittima viene avvolta in un sudario e seppellita parzialmente. Dopo di che le vengono tirati addosso dei sassi che cominciano a distruggerne la carne e le ossa, fino al sopraggiungere della morte. Le regole della lapidazione sono agghiaccianti ed assurde. Vengono scelti solo i sassi abbastanza piccoli da causare una morte il più lenta possibile. La pratica del rogo invece, è una forma di esecuzione applicata, nei tempi passati, per punire coloro che venivano giudicati colpevoli di eresia. In Sud-Africa, nei tempi recenti, un copertone viene riempieto di benzina ed infilato a forza attorno al busto della vittima, in maniera tale da bloccarne anche le braccia. Dopo di che viene appiccato il fuoco. La morte sopraggiunge dopo circa 20 minuti. Venti minuti nei quali il condannato a morte è costretto a subire ustioni tremende che, poco per volta, ne distruggono il corpo tramutandolo in un ammasso di carne carbonizzata e gomma fusa. Una tecnica di tortura estremamente dolorosa. Una condanna a morte disumana che, tristemente,è attuata ancora oggi. Infine, uno dei metodi di esecuzione più dolorosi, crudeli e sanguinolenti della storia è; stato certamente l'Aquila di Sangue. Una forma di esecuzione leggendaria, tipica delle popolazioni antiche del nord Europa. Il condannato veniva spogliato ed immobilizzato su un altare in posizione prona. Il carnefice, armato di un affilato coltello, praticava un'incisione molto profonda lungo la sua schiena. Nonostante questo primo passo della tortura fosse già; di per se estremamente doloroso, non era che l'inizio. L'aguzzino infilava a forza le mani all'interno dell'incisione praticata e, con violenza, spezzava le costole della vittima e le faceva fuoriuscire della schiena, in modo tale che ricordassero un paio di ali. Dopo di che, sempre con le mani, estraeva i polmoni dell'agonizzante condannato e li posizionava sulle sue spalle. La vittima moriva entro poco tempo per soffocamento circondata dal suo stesso sangue, tra i più atroci ed inimmaginabili dolori.
English:	One of the cruelest death sentences ever, and still in force in some parts of the world, is certainly stoning. This execution technique is practiced mainly in Islamic countries and is reserved for adulterers of both sexes. The victim is wrapped in a shroud and partially buried. After that, stones are thrown at her that begin to destroy her flesh and bones until death occurs. The rules of stoning are chilling and absurd. Only stones small enough to cause the slowest possible death are chosen. The practice of burning at the stake, on the other hand, is a form of execution applied in earlier times to punish those found guilty of heresy. In South Africa, in recent times, a tire is filled with gasoline and forced around the victim's torso in such a way that the victim's arms are also blocked. After that, a fire is set. Death occurs after about 20 minutes. Twenty minutes in which the condemned man is forced to suffer tremendous burns that, little by little, destroy his body, turning it into a pile of charred flesh and melted rubber. An extremely painful torture technique. An inhumane death sentence that, sadly, is still carried out today. Finally, one of the most painful, cruel and bloody methods of execution in history is; certainly was the Blood Eagle. A legendary form of execution, typical of the ancient peoples of northern Europe. The condemned person was stripped and immobilized on an altar in a prone position. The executioner, armed with a sharp knife, would make a very deep incision along his back. Although this first step of torture was already; in itself extremely painful, it was only the beginning. The torturer would forcefully thrust his hands inside the made incision and violently break the victim's ribs and spill them out of his back in such a way that they resembled a pair of wings. After that, again with his hands, he would extract the lungs of the agonizing condemned man and place them on his shoulders. The victim would die within a short time from suffocation surrounded by his own blood, amid the most excruciating and unimaginable pains.

continued from previous page

Emotional text No. 2	
	Average sentence length: 21.57 ± 11.41; Average words length per sentence: 4.81 ± 0.52 Valence Rate: -1.19 ± 0.84; Arousal Rate: 3.24 ± 1.09
Italian:	La divisione del corpo è un metodo di esecuzione veramente raccapricciante. Quanto di più doloroso e crudele voi possiate immaginare. La sega era utilizzata quando si voleva dare una morte lenta alla vittima, ed allo stesso tempo la si voleva privare della sua dignità. Il condannato veniva appeso a testa in giù per le caviglie. Le sue mani erano saldamente bloccate affinché non potesse in alcun modo disturbare il lavoro dei suoi aguzzini. Il loro compito era tanto semplice quanto sadico. Una grossa sega veniva posizionata sull’inguine del condannato che veniva diviso in due nella maniera più lenta possibile. La morte sopraggiungeva spesso dopo molto tempo. Il posizionamento a testa in giù della vittima faceva si che tutto il sangue scorresse verso il cervello, amplificando la sensazione di dolore e ritardando la morte per dissanguamento. Una tortura violentissima, che portava il condannato a trascorrere gli ultimi attimi della sua vita urlando di dolore e pregando i suoi carnefici per una morte più rapida. La morte per scuoiamento raggiunge l’apice del sadismo umano. Oltre al compito di scuoiare il condannato a morte, premurandosi di tenerlo in vita il più a lungo possibile, il boia doveva anche preservare la pelle della sua vittima in modo da poterla appendere per le vie della città come monito alla popolazione. Questa tecnica di esecuzione era molto diffusa in tutto il mondo e le varianti erano molte. In Cina , si toglieva la pelle solamente dal volto della vittima. In Persia, tutto il corpo veniva spellato, portandone la carne viva a vista. Immaginate il dolore ed il terrore di un condannato allo scuoiamento mentre il suo aguzzino, lentamente e con perizia, gli asportava la pelle, pezzo dopo pezzo. Immaginate di vedere parti del vostro corpo che una dopo l’ altra vengono posizionate su un tavolo davanti a voi mentre vi tramutate lentamente in una maschera di sangue e carne. Normalmente la morte per dissanguamento sopraggiungeva dopo qualche ora, ma se il boia era esperto riusciva a compiere il lavoro senza provocare eccessive perdite di sangue. In questo caso l’agonia era notevolmente prolungata ed era lo shock ad uccidere il condannato.
English:	Body division is a truly gruesome method of execution. As painful and cruel as you can imagine. The saw was used when you wanted to give a slow death to the victim, and at the same time you wanted to deprive him of his dignity. The condemned man was hung upside down by his ankles. His hands were firmly locked so that he could in no way disturb the work of his tormentors. Their task was as simple as it was sadistic. A large saw was placed on the condemned man’s groin, which was split in two in the slowest way possible. Death often came after a long time. Positioning the victim upside down caused all the blood to flow to the brain, amplifying the sensation of pain and delaying death by exsanguination. A very violent torture, which led the condemned person to spend the last moments of his life screaming in pain and begging his executioners for a quicker death. Death by flaying reaches the height of human sadism. In addition to the task of flaying the condemned man to death, taking care to keep him alive as long as possible, the executioner also had to preserve the skin of his victim so that he could hang it on the streets of the city as a warning to the populace. This execution technique was widespread throughout the world and there were many variations. In China, skin was removed only from the victim’s face. In Persia, the whole body was skinned, bringing the living flesh into view. Imagine the pain and terror of a condemned person being skinned as his torturer slowly and expertly removed his skin, piece by piece. Imagine seeing parts of your body being placed one after ’the other on a table in front of you as you slowly morphed into a mask of blood and flesh. Normally death by exsanguination would come after a few hours, but if the executioner was skilled he could get the job done without causing excessive blood loss. In this case the agony was considerably prolonged and it was the shock that killed the condemned man.

Correlations between Speech Features and Linguistic Features

Table C.2: Results of correlation analysis 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 (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 and different from zero.

	speech features					
linguistic features	**F0**	**F1**	**F2**	**F3**	**mean duration**	**signal duration**
raw text properties						
average clause length	33 (-0.103)	12 (-0.021)	3 (0.04)	45 (-0.06)	55 (0.004)	100 (-0.03)
sentence length	24 (-0.144)	9 (-0.015)	3 (0.43)	58 (0.033)	73 (0.054)	100 (0.075)
lexical density						
lexical density	·	·	·	12 (-0.027)	18 (0.007)	100 (-0.017)
morpho-syntactic information						
auxiliary form: finite	30 (-0.095)	9 (-0.062)	·	42 (-0.073)	64 (-0.011)	100 (-0.026)
auxiliary form: infinite	30 (-0.087)	9 (-0.014)	·	42 (-0.07)	55 (-0.009)	100 (-0.029)
auxiliary mood: indicative	33 (-0.04)	9 (-0.052)	·	39 (-0.027)	52 (-0.025)	100 (-0.014)
auxiliary mood: subjunctive	30 (-0.118)	9 (-0.019)	·	36 (-0.065)	58 (-0.016)	100 (-0.028)
auxiliary person: 3rd plural	30 (-0.039)	12 (-0.05)	3 (-0.044)	45 (-0.031)	58 (-0.01)	100 (-0.016)
auxiliary person: 2nd singular	30 (-0.038)	12 (-0.05)	3 (-0.045)	45 (-0.031)	58 (-0.012)	100 (-0.013)
auxiliary person: 3rd singular	30 (-0.039)	12 (-0.05)	3 (-0.046)	45 (-0.031)	58 (-0.003)	100 (-0.008)
auxiliary tense: future	30 (-0.059)	9 (-0.05)	3 (-0.047)	39 (-0.027)	52 (-0.011)	100 (-0.014)
auxiliary tense: imperative	30 (-0.035)	9 (-0.051)	3 (-0.047)	39 (-0.037)	52 (-0.011)	100 (-0.011)
auxiliary tense: present	30 (-0.028)	9 (-0.016)	·	42 (-0.02)	58 (-0.01)	100 (-0.011)
verb from: finite	21 (0.004)	6 (-0.059)	·	36 (-0.034)	45 (-0.012)	100 (-0.012)
verb form: gerundive	18 (0.032)	6 (-0.055)	·	33 (-0.019)	45 (-0.019)	100 (-0.021)
verb form: infinite	24 (-0.007)	9 (-0.016)	·	36 (-0.02)	52 (-0.002)	100 (-0.006)
verb form: participe	18 (-0.088)	6 (-0.055)	·	33 (-0.049)	42 (0.001)	100 (-0.025)
verb mood: imperative	·	·	·	18 (-0.017)	27 (0.002)	100 (-0.008)
verb mood: indicative	·	·	·	24 (-0.024)	36 (0.01)	100 (-0.016)

continued on next page

continued from previous page

	speech features					
linguistic features	**F0**	**F1**	**F2**	**F3**	**mean duration**	**signal duration**
verb mood: subjunctive	·	·	·	12 (-0.057)	27 (0.003)	100 (-0.018)
verb person: 2nd plural	30 (-0.105)	9 (-0.052)	·	36 (-0.081)	55 (-0.02)	100 (-0.031)
verb person: 3rd plural	30 (-0.094)	9 (-0.052)	3 (0.049)	36 (-0.081)	55 (-0.02)	100 (-0.031)
verb person: 2nd singular	27 (-0.102)	9 (-0.061)	·	36 (-0.062)	52 (-0.011)	100 (-0.025)
verb person: 3rd singular	30 (-0.094)	9 (-0.016)	3 (0.049)	36 (-0.064)	52 (-0.004)	100 (-0.024)
verb tense: future	·	·	·	30 (-0.029)	39 (0.006)	100 (-0.021)
verb tense: imperative	12 (-0.013)	·	·	27 (-0.032)	42 (-0.01)	100 (-0.016)
verb tense: past	12 (-0.014)	6 (0.192)	·	27 (-0.026)	42 (0.032)	100 (-0.007)
verb tense: present	6 (-0.105)	·	·	30 (-0.017)	42 (-0.011)	100 (-0.026)
adjective	·	·	·	6 (-0.066)	·	18 (0.102)
adjective (possessive)	·	·	·	9 (-0.021)	12 (0.004)	88 (-0.019)
adverb	·	·	·	6 (-0.01)	9 (-0.017)	70 (-0.036)
adverb (negation)	·	·	·	9 (-0.022)	12 (-0.022)	79 (-0.017)
conjunction (coordinative)	·	·	·	6 (-0.067)	9 (-0.018)	79 (-0.014)
conjunction (subordinative)	·	·	·	6 (-0.008)	12 (0.006)	79 (-0.03)
determiner (demonstrative)	·	·	·	6 (-0.067)	·	3 (0.075)
determiner (indefinite)	·	·	·	6 (-0.006)	·	39 (-0.006)
preposition	·	·	·	6 (-0.005)	9 (-0.017)	61 (-0.006)
punctuation (balanced)	·	·	·	6 (-0.007)	·	12 (0.042)
punctuation (clause boundary)	·	·	·	9 (-0.022)	12 (0.005)	79 (-0.024)
punctuation (comma, hyphen)	·	·	·	12 (-0.029)	18 (0.006)	100 (-0.016)
punctuation (sentence boundary)	·	·	·	9 (-0.022)	12 (-0.022)	79 (-0.022)
number (cardinal)	·	·	·	6 (-0.006)	·	·
number (ordinal)	·	·	·	6 (-0.005)	3 (-0.069)	45 (-0.019)
pronoun (clitic)	·	·	·	9 (-0.063)	15 (-0.006)	91 (-0.014)
pronoun (demonstrative)	·	·	·	12 (-0.059)	18 (-0.014)	100 (-0.015)
pronoun (personal)	·	·	·	6 (-0.066)	·	18 (0.004)
pronoun (indefinite)	·	·	·	9 (-0.02)	12 (-0.02)	79 (-0.027)
pronoun (possessive)	·	·	·	6 (-0.068)	12 (-0.021)	79 (-0.019)
pronoun (interrogative)	·	·	·	9 (-0.062)	15 (-0.028)	91 (-0.017)
pronoun (relative)	·	·	·	6 (-0.066)	3 (-0.069)	45 (0.015)

continued on next page

continued from previous page

	speech features					
linguistic features	**F0**	**F1**	**F2**	**F3**	**mean duration**	**signal duration**
article (determinative)	·	·	·	12 (-0.029)	18 (0.007)	100 (-0.02)
article (indeterminative)	·	·	·	18 (-0.015)	30 (0.01)	100 (-0.03)
noun (common)	·	·	·	6 (-0.004)	·	36 (-0.006)
noun (proper)	·	·	·	6 (-0.008)	12 (0.005)	85 (-0.023)
predeterminer	·	·	·	9 (-0.023)	12 (-0.022)	85 (-0.029)
verb (main)	·	·	·	12 (-0.058)	21 (-0.001)	100 (0.019)
verb (auxiliary)	·	·	·	9 (-0.024)	12 (-0.022)	85 (-0.037)
verb (modal)	·	·	·	6 (-0.006)	·	3 (0.076)
verbal predicate structure						
verb arity (average)	58 (-0.052)	36 (-0.003)	21 (0.001)	73 (-0.033)	97 (-0.023)	100 (-0.01)
verb arity (0 dependency links)	58 (-0.077)	30 (-0.052)	21 (0.015)	73 (-0.054)	97 (-0.039)	100 (-0.029)
verb arity (1 dependency link)	58 (-0.039)	30 (0.005)	18 (0.006)	73 (-0.033)	97 (-0.02)	100 (-0.011)
verb arity (2 dependency links)	61 (-0.076)	30 (-0.065)	21 (0.015)	73 (-0.054)	97 (-0.043)	100 (-0.033)
verb arity (3 dependency links)	55 (-0.075)	30 (-0.037)	21 (0.016)	73 (-0.047)	97 (-0.034)	100 (-0.021)
verb arity (4 dependency links)	61 (-0.069)	33 (-0.053)	21 (0.015)	73 (-0.049)	97 (-0.04)	100 (-0.026)
verb arity (5 dependency links)	55 (-0.039)	30 (0.003)	15 (-0.006)	73 (-0.034)	97 (-0.021)	100 (-0.012)
verb arity (6 dependency links)	55 (-0.069)	30 (-0.04)	21 (0.032)	73 (-0.044)	97 (-0.032)	100 (-0.022)
verbal head per sentence	30 (-0.027)	12 (-0.048)	3 (-0.044)	45 (-0.022)	58 (-0.01)	100 (-0.016)
verbal roots percentage	33 (-0.039)	12 (-0.024)	3 (0.04)	45 (-0.022)	55 (-0.003)	100 (-0.016)
syntactic relations (distributions)						
clausal modifier of noun	39 (-0.039)	15 (-0.034)	9 (0.024)	58 (-0.033)	88 (-0.018)	100 (-0.013)
relative clause modifier	42 (-0.033)	15 (-0.052)	9 (-0.053)	67 (-0.037)	88 (-0.027)	100 (-0.018)
adverbial clause modifier	36 (-0.027)	18 (0.007)	9 (0.023)	61 (-0.042)	82 (-0.014)	100 (-0.008)
adverbial modifier	39 (-0.039)	15 (-0.033)	9 (-0.054)	64 (-0.031)	88 (-0.02)	100 (-0.015)
adjectival modifier	39 (-0.039)	15 (-0.005)	12 (-0.056)	67 (-0.027)	88 (-0.015)	100 (-0.011)
appositional modifier	33 (-0.039)	12 (-0.024)	3 (0.042)	42 (-0.028)	73 (-0.016)	100 (-0.019)
auxiliary	33 (-0.038)	12 (-0.024)	3 (0.045)	45 (-0.015)	61 (-0.007)	100 (-0.015)
passive auxiliary	36 (-0.057)	18 (-0.009)	9 (0.023)	64 (-0.043)	88 (-0.019)	100 (-0.014)
case marking	42 (-0.033)	15 (-0.053)	9 (-0.053)	58 (-0.035)	88 (-0.027)	100 (-0.019)
coordinating conjunction	39 (-0.083)	21 (-0.002)	9 (0.023)	61 (-0.065)	82 (-0.035)	100 (-0.029)
clausal complement	39 (-0.031)	15 (-0.032)	9 (-0.055)	64 (-0.031)	91 (-0.021)	100 (-0.015)

continued on next page

continued from previous page

	speech features					
linguistic features	**F0**	**F1**	**F2**	**F3**	**mean duration**	**signal duration**
compound	33 (-0.038)	12 (-0.048)	3 (0.042)	42 (-0.036)	73 (-0.021)	100 (-0.022)
conjunct	39 (-0.038)	15 (0.066)	12 (0.125)	64 (-0.047)	85 (-0.017)	100 (-0.001)
copula	39 (-0.031)	15 (-0.052)	9 (-0.053)	67 (-0.032)	88 (-0.027)	100 (-0.019)
clausal subject	42 (-0.033)	15 (-0.053)	9 (-0.053)	61 (-0.031)	88 (-0.027)	100 (-0.019)
determiner	33 (-0.038)	12 (-0.048)	3 (0.044)	45 (-0.028)	64 (-0.019)	100 (-0.022)
possessive determiner	33 (-0.037)	15 (-0.03)	9 (0.024)	55 (-0.033)	76 (-0.02)	100 (-0.016)
predeterminer	36 (-0.057)	18 (-0.032)	9 (0.023)	64 (-0.043)	88 (-0.023)	100 (-0.018)
expletive	33 (-0.04)	12 (-0.047)	3 (0.04)	42 (-0.036)	55 (-0.009)	100 (-0.022)
reflexive pronoun in reflexive passive	33 (-0.039)	12 (-0.024)	3 (0.043)	45 (-0.022)	73 (-0.016)	100 (-0.018)
fixed multiword expression	39 (-0.03)	18 (-0.015)	9 (0.023)	64 (-0.043)	88 (-0.013)	100 (-0.012)
names	36 (-0.048)	18 (-0.033)	9 (0.023)	64 (-0.043)	88 (-0.025)	100 (-0.02)
indirect object	36 (-0.037)	12 (-0.034)	6 (0.012)	55 (-0.039)	82 (-0.034)	100 (-0.028)
marker	33 (-0.039)	12 (-0.051)	3 (0.043)	45 (-0.031)	73 (-0.016)	100 (-0.015)
nominal modifier	36 (-0.047)	12 (-0.029)	6 (-0.0)	58 (-0.036)	82 (-0.019)	100 (-0.017)
nominal subject	33 (-0.04)	12 (-0.028)	3 (0.042)	42 (-0.031)	55 (0.004)	100 (-0.013)
passive nominal subject	36 (-0.028)	21 (-0.018)	9 (-0.054)	55 (-0.021)	82 (-0.022)	100 (-0.015)
numeric modifier	33 (-0.1)	12 (-0.05)	9 (0.025)	52 (-0.061)	73 (-0.03)	100 (-0.032)
object	33 (-0.038)	12 (-0.024)	3 (0.043)	42 (-0.028)	64 (-0.014)	100 (-0.018)
oblique nominal	33 (-0.039)	15 (-0.052)	6 (0.002)	45 (-0.028)	73 (-0.02)	100 (-0.023)
agent modifier	39 (-0.084)	15 (-0.031)	9 (0.023)	61 (-0.059)	88 (-0.039)	100 (-0.03)
punctuation	33 (-0.04)	12 (-0.023)	3 (0.044)	45 (-0.021)	61 (-0.008)	100 (-0.019)
root	39 (-0.077)	15 (-0.031)	9 (0.023)	61 (-0.059)	88 (-0.039)	100 (-0.03)
open clausal complement	39 (-0.094)	15 (-0.031)	9 (0.023)	67 (-0.064)	88 (-0.044)	100 (-0.033)
global and local parsed tree structure						
dependency link: avg length	33 (-0.038)	12 (-0.024)	3 (0.04)	45 (-0.015)	55 (-0.002)	100 (-0.016)
parsed tree: avg max depth	33 (-0.039)	12 (-0.05)	3 (0.039)	45 (-0.03)	55 (-0.002)	100 (-0.016)
dependency link: avg max length	33 (-0.04)	12 (-0.05)	3 (0.04)	45 (-0.021)	55 (-0.0)	100 (-0.015)
prepositional chain: avg length	45 (-0.066)	30 (-0.03)	15 (0.017)	70 (-0.054)	91 (-0.036)	100 (-0.022)
dependency links: max length	33 (-0.052)	12 (-0.058)	3 (0.04)	45 (-0.032)	55 (-0.01)	100 (-0.02)
prepositional chain: number	45 (-0.036)	27 (0.009)	12 (0.01)	67 (-0.043)	91 (-0.026)	100 (-0.017)
post-verbal object	39 (-0.039)	27 (0.014)	12 (0.01)	67 (-0.038)	91 (-0.014)	100 (-0.012)

continued on next page

continued from previous page

	speech features					
linguistic features	**F0**	**F1**	**F2**	**F3**	**mean duration**	**signal duration**
pre-verbal object	42 (-0.038)	24 (0.005)	12 (0.017)	64 (-0.033)	85 (-0.023)	100 (-0.021)
prepositional chain: 1 element	48 (-0.036)	24 (0.002)	12 (-0.057)	70 (-0.047)	94 (-0.021)	100 (-0.017)
prepositional chain: 2 elements	45 (-0.027)	27 (-0.042)	12 (-0.057)	70 (-0.047)	91 (-0.028)	100 (-0.018)
prepositional chain: 4 elements	48 (-0.044)	24 (-0.024)	12 (-0.056)	70 (-0.049)	94 (-0.029)	100 (-0.022)
post-verbal subject	42 (-0.09)	24 (-0.066)	9 (0.023)	64 (-0.066)	85 (-0.036)	100 (-0.033)
pre-verbal subject	42 (-0.075)	21 (-0.084)	9 (0.023)	64 (-0.059)	85 (-0.028)	100 (-0.026)
use of subordination						
subordinate chains: avg length	48 (-0.039)	27 (-0.031)	15 (-0.007)	70 (-0.039)	97 (-0.029)	100 (-0.019)
principal proposition distribution	48 (-0.046)	27 (-0.057)	15 (-0.052)	70 (-0.049)	94 (-0.029)	100 (-0.026)
subordinate: embedded 1	48 (-0.031)	24 (0.002)	12 (-0.057)	70 (-0.042)	94 (-0.022)	100 (-0.014)
subordinate: embedded 2	48 (-0.04)	27 (-0.032)	15 (-0.053)	70 (-0.048)	94 (-0.027)	100 (-0.02)
subordinate: embedded 3	52 (-0.033)	24 (0.002)	12 (-0.057)	70 (-0.042)	94 (-0.021)	100 (-0.014)
subordinate: embedded 4	52 (-0.046)	24 (-0.023)	12 (0.002)	70 (-0.038)	94 (-0.021)	100 (-0.014)
subordinate: embedded 5	52 (-0.046)	24 (-0.044)	12 (-0.057)	70 (-0.056)	94 (-0.034)	100 (-0.027)
post-verbal subordinate	55 (-0.04)	30 (-0.021)	18 (-0.033)	70 (-0.043)	97 (-0.029)	100 (-0.016)
pre-verbal subordinate	48 (-0.032)	30 (-0.032)	15 (-0.052)	70 (-0.043)	97 (-0.029)	100 (-0.019)
subordinate proposition distribution	48 (-0.032)	27 (-0.032)	15 (-0.052)	70 (-0.047)	97 (-0.032)	100 (-0.019)

Correlations between Electrodermal Activity Features and Linguistic Features

Table C.3: Results of correlation analysis between Electrodermal Activity 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 (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 and different from zero.

	electrodermal activity features								
		phasic component					**tonic component**		
linguistic features	**eda** **symp**	**max** **pks**	**no** **pks**	**sum** **pks**	**mean** **ph**	**std** **ph**	**max** **ton**	**mean** **ton**	**std** **ton**
raw text properties									
average clause length	3 (0.047)	6 (0.278)	21 (0.015)	27 (0.086)	3 (-0.062)	39 (0.017)	3 (0.062)	3 (0.062)	39 (-0.025)
sentence length	3 (-0.508)	12 (0.166)	39 (0.062)	52 (0.051)	3 (-0.504)	64 (-0.006)	·	·	52 (0.0)
lexical density									
lexical density	·	·	·	·	·	3 (0.064)	·	·	3 (0.066)
morpho-syntactic information									
auxiliary form: finite	·	6 (0.058)	21 (0.017)	30 (0.029)	3 (-0.064)	42 (-0.019)	3 (0.065)	3 (0.064)	36 (-0.027)
auxiliary form: infinite	·	6 (0.05)	18 (0.026)	30 (0.024)	3 (-0.063)	39 (-0.028)	3 (0.064)	3 (0.063)	36 (-0.018)
auxiliary mood: indicative	·	6 (0.049)	15 (0.022)	24 (0.001)	3 (-0.063)	36 (-0.012)	3 (0.064)	3 (0.063)	33 (-0.013)
auxiliary mood: subjunctive	·	6 (0.056)	18 (0.028)	21 (0.029)	3 (-0.061)	30 (-0.004)	3 (0.063)	3 (0.062)	33 (-0.011)
auxiliary person: 3rd plural	3 (-0.057)	3 (0.052)	21 (-0.0)	30 (0.013)	3 (-0.066)	42 (-0.037)	3 (0.066)	3 (0.065)	33 (-0.035)
auxiliary person: 2nd singular	3 (-0.056)	6 (0.049)	21 (0.015)	30 (0.013)	3 (-0.064)	42 (-0.03)	3 (0.064)	3 (0.063)	33 (-0.035)
auxiliary person: 3rd singular	3 (-0.056)	6 (0.055)	21 (0.037)	30 (0.013)	3 (-0.064)	42 (-0.019)	3 (0.065)	3 (0.063)	33 (-0.021)
auxiliary tense: future	·	3 (0.063)	12 (0.012)	21 (0.029)	3 (-0.06)	24 (-0.004)	3 (0.061)	3 (0.059)	33 (-0.02)
auxiliary tense: imperative	·	6 (0.053)	12 (0.014)	21 (0.009)	3 (-0.06)	24 (0.009)	3 (0.061)	3 (0.06)	30 (-0.006)
auxiliary tense: present	·	6 (0.054)	18 (0.01)	27 (-0.002)	3 (-0.063)	36 (-0.003)	3 (0.064)	3 (0.063)	36 (-0.006)
verb from: finite	·	3 (0.058)	12 (0.011)	12 (0.028)	·	15 (-0.001)	·	·	21 (-0.016)
verb form: gerundive	·	·	6 (0.064)	3 (0.068)	·	9 (0.016)	·	·	18 (0.006)
verb form: infinite	·	3 (0.063)	12 (0.038)	18 (0.023)	3 (-0.061)	27 (-0.002)	3 (0.064)	3 (0.063)	27 (0.014)

continued on next page

continued from previous page

	electrodermal activity features								
		phasic component					**tonic component**		
linguistic features	**eda symp**	**max pks**	**no pks**	**sum pks**	**mean ph**	**std ph**	**max ton**	**mean ton**	**std ton**
verb form: participe	·	·	9 (0.029)	6 (0.075)	·	9 (0.067)	·	·	18 (0.002)
verb mood: imperative	·	·	·	·	·	3 (0.064)	·	·	3 (0.066)
verb mood: indicative	·	·	·	·	·	6 (0.067)	·	·	12 (0.007)
verb mood: subjunctive	·	·	·	·	·	6 (0.057)	·	·	·
verb person: 2nd plural	·	6 (0.059)	18 (0.011)	21 (0.028)	3 (-0.061)	30 (-0.015)	3 (0.064)	3 (0.063)	33 (-0.011)
verb person: 3rd plural	·	6 (0.06)	24 (0.026)	21 (0.024)	6 (-0.285)	30 (-0.014)	3 (0.064)	3 (0.063)	33 (-0.021)
verb person: 2nd singular	·	3 (0.061)	12 (0.004)	18 (0.04)	3 (-0.061)	27 (0.015)	3 (0.063)	3 (0.062)	27 (-0.015)
verb person: 3rd singular	·	6 (0.054)	21 (0.035)	21 (0.027)	6 (-0.285)	30 (-0.017)	3 (0.065)	3 (0.064)	30 (-0.018)
verb tense: future	·	·	·	·	·	6 (0.067)	·	·	12 (0.007)
verb tense: imperative	·	·	3 (0.064)	·	·	12 (0.001)	·	·	18 (-0.01)
verb tense: past	·	·	3 (0.062)	3 (0.063)	·	6 (0.068)	·	·	15 (0.115)
verb tense: present	·	·	3 (0.067)	·	·	15 (0.017)	·	·	21 (-0.021)
adjective	·	·	·	·	·	·	·	·	·
adjective (possessive)	·	·	·	·	·	·	·	·	·
adverb	·	·	·	·	·	·	·	·	·
adverb (negation)	·	·	·	·	·	·	·	·	·
conjunction (coordinative)	·	·	·	·	·	·	·	·	·
conjunction (subordinative)	·	·	·	·	·	·	·	·	·
determiner (demonstrative)	·	·	·	·	·	·	·	·	·
determiner (indefinite)	·	·	·	·	·	·	·	·	·
preposition	·	·	·	·	·	·	·	·	·
punctuation (balanced)	·	·	·	·	·	·	·	·	·
punctuation (clause boundary)	·	·	·	·	·	·	·	·	·
punctuation (comma, hyphen)	·	·	·	·	·	6 (0.066)	·	·	3 (0.069)
punctuation (sentence boundary)	·	·	·	·	·	·	·	·	·
number (cardinal)	·	·	·	·	·	·	·	·	·
number (ordinal)	·	·	·	·	·	·	·	·	·
pronoun (clitic)	·	·	·	·	·	·	·	·	·
pronoun (demonstrative)	·	·	·	·	·	6 (0.066)	·	·	3 (0.069)
pronoun (personal)	·	·	·	·	·	·	·	·	·

continued on next page

continued from previous page

	electrodermal activity features								
		phasic component					**tonic component**		
linguistic features	**eda symp**	**max pks**	**no pks**	**sum pks**	**mean ph**	**std ph**	**max ton**	**mean ton**	**std ton**
pronoun (indefinite)	·	·	·	·	·	·	·	·	·
pronoun (possessive)	·	·	·	·	·	·	·	·	·
pronoun (interrogative)	·	·	·	·	·	·	·	·	·
pronoun (relative)	·	·	·	·	·	·	·	·	·
article (determinative)	·	·	·	·	·	6 (0.065)	·	·	·
article (indeterminative)	·	·	·	·	·	6 (0.058)	·	·	9 (0.028)
noun (common)	·	·	·	·	·	·	·	·	·
noun (proper)	·	·	·	·	·	·	·	·	·
predeterminer	·	·	·	·	·	·	·	·	·
verb (main)	·	·	·	·	·	6 (0.064)	·	·	6 (0.066)
verb (auxiliary)	·	·	·	·	·	·	·	·	·
verb (modal)	·	·	·	·	·	·	·	·	·
verbal predicate structure									
verb arity (average)	18 (0.002)	48 (0.0)	70 (-0.01)	64 (-0.011)	27 (-0.017)	82 (-0.014)	18 (-0.041)	18 (-0.061)	88 (-0.012)
verb arity (0 dependency links)	18 (-0.096)	48 (-0.008)	67 (-0.043)	61 (-0.01)	27 (-0.093)	82 (-0.036)	18 (-0.039)	15 (-0.062)	88 (-0.034)
verb arity (1 dependency link)	15 (0.012)	48 (0.0)	70 (-0.013)	58 (-0.008)	24 (-0.0)	82 (-0.015)	18 (-0.041)	15 (-0.064)	88 (-0.014)
verb arity (2 dependency links)	18 (-0.097)	48 (-0.007)	67 (-0.047)	61 (-0.015)	27 (-0.094)	82 (-0.04)	21 (-0.125)	18 (-0.136)	88 (-0.037)
verb arity (3 dependency links)	15 (-0.107)	45 (0.014)	67 (-0.031)	61 (-0.005)	24 (-0.071)	79 (-0.029)	18 (-0.038)	15 (-0.059)	88 (-0.024)
verb arity (4 dependency links)	15 (-0.109)	48 (0.007)	73 (-0.033)	64 (-0.007)	27 (-0.084)	82 (-0.037)	18 (-0.062)	15 (-0.061)	88 (-0.03)
verb arity (5 dependency links)	15 (0.009)	45 (0.005)	67 (-0.015)	58 (-0.008)	21 (0.01)	82 (-0.014)	18 (-0.017)	15 (-0.034)	88 (-0.009)
verb arity (6 dependency links)	15 (-0.084)	48 (0.0)	70 (-0.033)	58 (-0.002)	24 (-0.071)	79 (-0.036)	15 (-0.039)	15 (-0.04)	88 (-0.026)
verbal head per sentence	3 (-0.056)	6 (0.048)	21 (-0.018)	30 (0.016)	3 (-0.065)	42 (-0.021)	3 (0.065)	3 (0.064)	36 (-0.043)
verbal roots percentage	3 (0.047)	6 (0.05)	21 (-0.003)	30 (0.016)	3 (-0.063)	39 (-0.009)	3 (0.064)	3 (0.063)	36 (-0.025)
syntactic relations (distributions)									
clausal modifier of noun	12 (-0.006)	21 (0.009)	39 (-0.006)	42 (0.001)	12 (-0.024)	55 (-0.043)	9 (-0.025)	9 (-0.026)	64 (-0.021)
relative clause modifier	12 (-0.007)	21 (-0.006)	42 (-0.024)	45 (0.0)	12 (0.004)	58 (-0.035)	9 (-0.028)	6 (-0.007)	70 (-0.018)
adverbial clause modifier	9 (0.018)	18 (0.039)	36 (0.008)	39 (0.026)	9 (0.03)	52 (-0.032)	6 (-0.005)	3 (0.066)	58 (-0.023)
adverbial modifier	12 (0.0)	21 (-0.007)	39 (-0.026)	45 (-0.002)	12 (0.004)	58 (-0.031)	6 (-0.006)	6 (-0.008)	67 (-0.018)
adjectival modifier	12 (0.001)	21 (-0.006)	48 (0.004)	45 (0.009)	12 (0.004)	58 (-0.022)	9 (-0.029)	6 (-0.007)	64 (-0.016)
appositional modifier	6 (-0.003)	12 (-0.0)	24 (-0.021)	33 (0.022)	6 (0.008)	48 (-0.03)	3 (0.064)	3 (0.062)	48 (-0.024)

continued on next page

continued from previous page

	electrodermal activity features								
		phasic component					**tonic component**		
linguistic features	**eda symp**	**max pks**	**no pks**	**sum pks**	**mean ph**	**std ph**	**max ton**	**mean ton**	**std ton**
auxiliary	6 (-0.002)	6 (0.05)	21 (-0.033)	33 (0.021)	3 (-0.061)	42 (-0.012)	3 (0.061)	3 (0.06)	42 (-0.028)
passive auxiliary	9 (-0.025)	24 (0.016)	42 (-0.021)	45 (0.01)	9 (-0.012)	58 (-0.033)	6 (-0.005)	6 (-0.007)	70 (-0.021)
case marking	12 (-0.007)	15 (0.013)	39 (-0.033)	42 (-0.007)	12 (0.004)	55 (-0.043)	9 (-0.024)	6 (-0.007)	67 (-0.017)
coordinating conjunction	9 (-0.026)	15 (0.059)	45 (-0.04)	39 (0.011)	12 (-0.143)	52 (-0.038)	9 (-0.03)	6 (-0.007)	58 (-0.022)
clausal complement	12 (0.0)	18 (-0.015)	45 (-0.011)	45 (-0.002)	12 (-0.021)	55 (-0.035)	9 (-0.021)	9 (-0.022)	61 (-0.02)
compound	6 (-0.003)	12 (-0.002)	24 (-0.035)	33 (0.02)	6 (0.008)	48 (-0.038)	3 (0.063)	3 (0.061)	48 (-0.032)
conjunct	12 (-0.023)	18 (0.08)	45 (-0.004)	45 (0.01)	12 (-0.014)	55 (-0.004)	9 (0.185)	9 (0.184)	61 (-0.015)
copula	15 (0.01)	18 (-0.015)	42 (-0.025)	45 (0.0)	15 (-0.01)	55 (-0.041)	9 (-0.026)	9 (-0.027)	64 (-0.021)
clausal subject	9 (-0.026)	21 (-0.008)	39 (-0.033)	42 (-0.007)	12 (0.004)	55 (-0.042)	9 (-0.025)	6 (-0.007)	67 (-0.017)
determiner	3 (0.049)	6 (0.05)	18 (-0.031)	33 (0.019)	3 (-0.061)	39 (-0.017)	3 (0.06)	3 (0.059)	42 (-0.037)
possessive determiner	6 (-0.003)	18 (0.037)	36 (-0.01)	33 (0.009)	12 (0.034)	48 (-0.038)	3 (0.063)	3 (0.061)	55 (-0.041)
predeterminer	9 (-0.025)	21 (0.029)	45 (-0.023)	48 (-0.002)	9 (-0.012)	58 (-0.04)	6 (-0.005)	6 (-0.007)	67 (-0.026)
expletive	3 (0.046)	6 (0.051)	18 (-0.032)	33 (0.02)	3 (-0.059)	42 (-0.021)	3 (0.06)	3 (0.059)	39 (-0.035)
reflexive pronoun in reflexive passive	6 (-0.003)	9 (-0.018)	24 (-0.021)	33 (0.02)	6 (0.008)	45 (-0.022)	3 (0.063)	3 (0.061)	48 (-0.024)
fixed multiword expression	12 (-0.0)	18 (0.005)	45 (0.005)	42 (0.013)	12 (-0.022)	55 (-0.042)	9 (-0.021)	9 (-0.022)	61 (-0.026)
names	9 (-0.026)	24 (-0.002)	42 (-0.033)	48 (-0.006)	9 (-0.014)	58 (-0.049)	6 (-0.006)	6 (-0.008)	70 (-0.029)
indirect object	9 (-0.049)	18 (-0.031)	36 (0.005)	36 (0.003)	9 (0.03)	52 (-0.048)	3 (0.066)	3 (0.064)	58 (-0.024)
marker	6 (-0.003)	9 (0.061)	24 (-0.005)	33 (0.032)	6 (0.006)	45 (-0.044)	3 (0.063)	3 (0.062)	45 (-0.038)
nominal modifier	6 (-0.003)	18 (0.037)	36 (-0.001)	36 (0.012)	6 (0.006)	48 (-0.031)	3 (0.064)	3 (0.062)	58 (-0.028)
nominal subject	3 (0.047)	6 (-0.006)	18 (0.026)	30 (0.004)	3 (-0.057)	39 (-0.034)	3 (0.058)	3 (0.057)	39 (-0.017)
passive nominal subject	9 (0.019)	18 (0.019)	42 (-0.016)	36 (0.001)	9 (0.031)	52 (-0.016)	6 (-0.006)	3 (0.066)	55 (-0.012)
numeric modifier	6 (-0.002)	18 (0.06)	36 (-0.047)	33 (0.02)	6 (-0.285)	45 (-0.026)	3 (0.061)	3 (0.059)	55 (-0.019)
object	6 (-0.001)	6 (0.051)	21 (-0.033)	33 (0.021)	3 (-0.062)	39 (-0.009)	3 (0.062)	3 (0.061)	42 (-0.03)
oblique nominal	6 (-0.002)	12 (-0.002)	33 (-0.031)	33 (0.006)	6 (0.008)	48 (-0.036)	3 (0.064)	3 (0.062)	48 (-0.031)
agent modifier	12 (-0.006)	15 (0.06)	39 (-0.014)	42 (0.014)	12 (-0.143)	55 (-0.033)	9 (-0.026)	6 (-0.007)	61 (-0.016)
punctuation	3 (0.05)	6 (0.056)	21 (-0.033)	33 (0.021)	3 (-0.06)	42 (-0.012)	3 (0.061)	3 (0.059)	42 (-0.03)
root	12 (-0.006)	18 (0.024)	39 (-0.007)	42 (0.02)	12 (-0.143)	55 (-0.028)	9 (-0.026)	6 (-0.007)	61 (-0.012)
open clausal complement	12 (-0.005)	18 (0.026)	39 (-0.013)	42 (-0.004)	12 (-0.174)	55 (-0.04)	9 (-0.029)	9 (-0.03)	67 (-0.022)
global and local parsed tree structure									
dependency link: avg length	3 (0.047)	6 (0.049)	18 (-0.013)	27 (0.009)	3 (-0.062)	39 (-0.009)	3 (0.062)	3 (0.061)	39 (-0.018)

continued on next page

continued from previous page									
	electrodermal activity features								
		phasic component					**tonic component**		
linguistic features	**eda symp**	**max pks**	**no pks**	**sum pks**	**mean ph**	**std ph**	**max ton**	**mean ton**	**std ton**
parsed tree: avg max depth	·	6 (0.049)	18 (0.008)	30 (0.028)	3 (-0.06)	39 (-0.027)	3 (0.061)	3 (0.06)	39 (-0.027)
dependency link: avg max length	·	6 (0.048)	18 (-0.008)	30 (0.016)	3 (-0.061)	39 (-0.006)	3 (0.061)	3 (0.06)	39 (-0.014)
prepositional chain: avg length	15 (-0.083)	30 (0.025)	52 (-0.033)	55 (0.008)	15 (-0.105)	64 (-0.018)	9 (0.006)	6 (0.052)	79 (-0.027)
dependency links: max length	·	6 (0.053)	18 (-0.013)	27 (0.011)	3 (-0.061)	39 (-0.007)	3 (0.061)	3 (0.061)	39 (-0.016)
prepositional chain: number	15 (0.008)	24 (0.002)	52 (-0.008)	52 (-0.012)	15 (-0.011)	61 (-0.035)	9 (-0.024)	6 (-0.005)	79 (-0.02)
post-verbal object	15 (0.01)	24 (0.018)	52 (0.0)	52 (0.002)	15 (-0.011)	61 (-0.023)	9 (0.014)	6 (-0.005)	79 (-0.01)
pre-verbal object	9 (-0.052)	24 (-0.053)	52 (-0.015)	45 (-0.028)	9 (0.028)	55 (-0.04)	6 (0.059)	6 (0.057)	67 (-0.014)
prepositional chain: 1 element	12 (-0.002)	33 (-0.018)	61 (-0.009)	55 (-0.0)	12 (0.007)	70 (-0.029)	9 (-0.03)	6 (-0.003)	82 (-0.007)
prepositional chain: 2 elements	12 (-0.002)	30 (0.001)	55 (-0.01)	55 (-0.007)	15 (-0.01)	70 (-0.034)	9 (0.014)	6 (-0.004)	79 (-0.009)
prepositional chain: 4 elements	12 (-0.001)	33 (-0.028)	61 (-0.025)	55 (-0.018)	12 (0.007)	70 (-0.037)	9 (-0.03)	6 (-0.002)	82 (-0.015)
post-verbal subject	9 (-0.025)	24 (0.015)	48 (-0.056)	45 (0.001)	12 (-0.141)	55 (-0.04)	6 (-0.005)	6 (-0.006)	70 (-0.048)
pre-verbal subject	9 (-0.026)	24 (0.014)	45 (-0.05)	45 (0.007)	12 (-0.142)	58 (-0.03)	6 (-0.006)	6 (-0.008)	67 (-0.037)
use of subordination									
subordinate chains: avg length	12 (-0.003)	39 (-0.013)	64 (-0.024)	58 (-0.021)	18 (-0.02)	79 (-0.031)	9 (-0.027)	12 (-0.04)	88 (-0.023)
principal proposition distribution	15 (0.011)	39 (-0.03)	64 (-0.035)	58 (-0.033)	18 (-0.019)	79 (-0.036)	12 (-0.04)	12 (-0.041)	88 (-0.03)
subordinate: embedded 1	12 (-0.002)	39 (-0.012)	61 (-0.008)	55 (0.001)	15 (-0.035)	73 (-0.03)	9 (-0.029)	9 (-0.03)	82 (-0.006)
subordinate: embedded 2	12 (-0.002)	39 (-0.026)	64 (-0.026)	55 (-0.017)	15 (-0.035)	73 (-0.037)	9 (-0.029)	9 (-0.03)	85 (-0.02)
subordinate: embedded 3	12 (-0.002)	33 (-0.016)	61 (-0.008)	55 (0.001)	12 (-0.025)	73 (-0.03)	9 (-0.029)	9 (-0.03)	82 (-0.006)
subordinate: embedded 4	12 (-0.001)	36 (-0.009)	61 (-0.013)	55 (-0.004)	12 (-0.023)	73 (-0.028)	9 (-0.026)	9 (-0.027)	82 (-0.015)
subordinate: embedded 5	12 (-0.001)	33 (-0.03)	61 (-0.034)	55 (-0.028)	12 (0.007)	73 (-0.046)	9 (-0.026)	6 (-0.002)	82 (-0.022)
post-verbal subordinate	15 (0.008)	45 (-0.003)	64 (-0.018)	58 (-0.014)	24 (-0.001)	82 (-0.024)	15 (-0.061)	15 (-0.062)	88 (-0.018)
pre-verbal subordinate	12 (-0.003)	42 (-0.008)	64 (-0.024)	58 (-0.021)	21 (-0.011)	82 (-0.028)	12 (-0.039)	15 (-0.063)	88 (-0.022)
subordinate proposition distribution	15 (0.009)	39 (-0.015)	64 (-0.02)	58 (-0.017)	21 (-0.009)	79 (-0.028)	12 (-0.036)	15 (-0.061)	88 (-0.023)