CHALMERS
UNIVERSITY OF TECHNOLOGY

Dynamics of Affective Information in English Novels

Using Information Theory to measure interaction patterns between fictional characters

Kevin Loris Jaquier

Department of Space, Earth and Environment
CHALMERS UNIVERSITY OF TECHNOLOGY

Master's thesis 2021

Dynamics of Affective Information in English Novels

Using Information Theory to measure interaction patterns between fictional characters

Kevin Loris Jaquier

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Space, Earth and Environment
Physical Resource Theory
Chalmers University of Technology
Gothenburg, Sweden 2021

Dynamics of Affective Information in English Novels
Using Information Theory to measure interaction patterns between fictional characters
KEVIN JAQUIER

Supervisor: Dr. Markus Lukzak-Roesch, School of Information Management, Victoria University of Wellington
Examiner: Prof. Kristian Lindgren, Department of Space, Earth and Environment

Master's Thesis 2021
Department of Space, Earth and Environment
Division of Physical Resource Theory
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Photograph by Nong Vang (`https://unsplash.com/photos/9pw4TKvT3po`)

Typeset in LaTeX
Gothenburg, Sweden 2021

Dynamics of Affective Information in English Novels
Using Information Theory to measure interaction patterns between fictional characters

Kevin Jaquier
Department of Space, Earth and Environment
Division of Physical Resource Theory
Chalmers University of Technology

Abstract

A central hypothesis in psychology is that every human language, and therefore every human culture, has an implicit theory of personality behind the meaning of words. As the language evolves over time, so do these ideas about similarities and differences between people. The increasing availability of digitised books creates opportunities for studying this evolution, and perhaps gaining new perspectives on how humans describe their peers across cultures.

Unfortunately, conventional approaches from contemporary personality theory rely on assumptions that may not apply to this situation. Following previous work inspired by dynamical systems theory, we experimented with a model of personality based on a feedback loop mechanism. In order to estimate how appropriate this model would be in such settings, we automatically extracted sequences of emotions about the characters of 150 classic English novels, and then used Information Theory to measure characteristics of temporal information patterns in those sequences – namely entropy rate and transfer entropy.

We faced a number of challenges related to the extraction of good quality semantic information, resulting in insufficient data to draw any solid conclusions from our entropy estimates. As a guide for future work, we discuss how our purpose-built natural language processing (NLP) program should be improved, in order to obtain the desired data quality and reliable estimates of entropy rates. We also provide suggestions for how more recent advances in NLP may be exploited while minimising the sources of biases which can be problematic in this context.

Keywords: information theory, natural language processing, psychology, entropy, emotion detection

Acknowledgements

Many thanks to Markus for having me in his lab, for his time and support, and for the many captivating ideas and conversations.
Thank you so much Ron for giving me the opportunity to do this work, and for the feedback.
Thanks to Johannes for his insights, and to Markus' students for their advice and company.
Thanks to Kristian for his time and guidance.
Thanks to my family for their significant support and encouragements.
And finally, special thanks to Anne-Laure without whom this project would not have been possible, and for her patience and understanding.

Kevin Jaquier, Lausanne, January 2021

Contents

List of Figures

List of Tables

1

Introduction

Similarities and differences in how people think, feel and act, and the consistence of these traits over time, is a subject of interest for social sciences. Personality Theory seeks to make sense of this complexity and its implications. The emergence of new data sources and computational methods has inspired researchers to compare known personality dimensions across different cultural and historical contexts. However, concerns have been raised about the appropriateness of some of the structures used, as cultural and situational factors are often not properly addressed [1, 2].

Rather than reducing this complexity to single attributes (e.g. open, extroverted) describing an assumed *average* personality across time and contexts, we explore in this work a possible alternative based on Information Theory, where we attempt to quantify the information produced by occurrences of certain emotional attributes (e.g. joy, anger, valence) over time.

This approach requires fewer assumptions about how personality manifests over time and situational contexts, while also reducing the dependence on preexisting ethnocentric personality theories. The resulting information structure could provide opportunities for more general and comprehensive personality models. The challenge, however, lies in the extraction of data of appropriate quality and size for such qualitative analysis.

1.1 Basic definitions

Let us first clarify the meaning of some key psychology constructs that will be discussed:

- *Affect* is the unconscious experience underlying feelings and emotions [3].
- *Behaviour* is the range of actions and mannerisms made by an individual [4].
- *Cognition* refers to mental information processing mechanisms such as thought, attention or memory.

A theory of personality is "concerned with describing and explaining the observed complexity of individual differences in the patterning of affect, behaviour, cognition and desires over time and space" [5].

A situation or situational context refers to the sequence of events preceding the manifestation of an affect, behaviour or cognition, and more specifically to the affective content of those events, e.g. threat, reward, surprise etc.

1.2 Personality across history and cultures

Much of the results from personality theory come from the *lexical hypothesis*, which postulates that (1) important characteristics of individual differences eventually become encoded in language, and (2) the most significant ones are likely to be expressed as single words. Based on this assumption, *psycholexical approaches* consists in deriving the main dimensions of personality variation from language, usually based on the meaning and usage of a large selection of words – a lexicon.

This line of research resulted in the identification of five universal factors underlying most of the observed variations [6], interpreted as *openness*, *conscientiousness*, *extroversion*, *agreeableness* and *neuroticism*. Referred to as the Five Factor Model (FFM) or Big Five, this is the most widely used personality model in the literature.

However, the lexicons used to derive the FFM originate from a contemporary Western context [7]. Some form of translation is required to reproduce such structure to other corpora, which is prone to bias since word meanings differ across languages and also change over time [8]. Although the FFM was successfully reproduced in cross-cultural studies [9], the validity of such generalisation has recently been challenged [1] due to interpretation issues with commonly used personality questions, both by the respondents and the survey authors. This confirms the high sensitivity of psycholexical models to how the personality lexicon is constructed and used, and motivates us to explore different approaches.

Being able to compare personality models across languages is critical for studying how implicit theories of personality evolve over time, because of how the use and meaning of words change. An ideal model would have to be generic enough to stay relevant for any language or culture, yet retain a meaningful enough structure so different languages would only vary in the parameters without requiring a complete redesign of the model.

1.3 A new computational approach

The availability of large literary corpora and the recent progresses in Natural Language Processing (NLP) allow for new experimentation with computational approaches, with two potential advantages over lexical approaches.

Rather than merely compiling personality-related terms, it is now possible to extract high-level high-resolution semantic meaning from the text, such as "who does what to whom and how?". This requires moving from a lexicon to a *language model*, an abstract representation of semantic meaning derived from a quantitative analysis of language rather than manual annotations or surveys. Language models also have the potential advantage of being reproducible to other languages without manual translation [10, 11].

The idea suggested here is to use these high-level semantic meanings to capture personality-relevant information in a principled way. More concretely, it means that instead of using some pre-existing lexicon, an algorithm identifies relevant words and semantic meanings in a text using a mix of syntactic and semantic features.

Not only would such approach allow to extract emotions, behaviour or any relevant characteristic of personality, but also the context in which they occur in the text. Sample sequences of situations and behaviours may be used to experiment with dynamical models of personality which include these aspects but lack validation from empirical data, such as work from Mischel and Shoda [12] and Read et al. [13].

1.4 Dynamical systems and personality

The definition given in 1.1 makes it clear that any observation of affect, behaviour or cognition should be contextualised with the current situation and the individual's past interactions with his environment. Some alternatives to psycholexical approaches acknowledge the relevance of the sequence of situations and treat personality as a behavioural trend in relation to it, moving the formalism from static latent variables to dynamical systems.

The shift to a dynamical perspective can shine new light on apparent inconsistencies observed with current models [12], and phenomena of psychological change such as age trends [14], cycles of inter-personal behaviour [15], bipolar depression [16] or non-linear life transitions [17]. The effect of the situational context and environment on personality over time has also been recognised as an underdeveloped aspect of the current state-of-the-art [18, 2, 19]. The NLP-based method described here could eventually help characterise the range of environmental (or external) "states" and their effect on subsequent behaviour, affect and cognition (internal states), thereby depicting the "state space" of an individual-environment system and its long-term dynamics. Such conceptualisation has already been formulated in previous work [12, 13, 20, 5], but empirical data is lacking for using such models in practical applications.

Rather than assuming a specific model, we focus on the general structure of this individual-environment system. We investigate the dynamical properties of sequential data from the text, such as temporal correlations, in order to validate the relevance of this data source and inform the choice of an appropriate class of model. Hence, data needs to be extracted as sequences of either numerical values or symbols, representing dimensions of psychological meaning, such as emotions (e.g. joy, anger), relevant theoretical constructs (e.g. valence, agreeableness) or possibly any data-driven abstract representation of semantic meaning (e.g. vector embedding from a language model).

1.5 Project purpose and scope

This project is an early step towards finding an appropriate abstraction level to model personality as a dynamical system, in such a way as to provide statistical methods for capturing personality characteristics from text, without depending on language- or corpus-specific features, or on pre-existing personality dimension, for the reasons previously described.

We attempt to translate the dynamical system perspective in terms of information-theoretic concept, and to apply the corresponding statistical measures on a corpus of English novels in order to investigate their potential for capturing personality-relevant properties.

We proceed by developing a text mining method to extract psychological information about the characters and their environment (other characters and entities), in the form of symbolic sequences of emotion-related features. We then estimate some long-term dynamical properties of these sequences, to determine the extent to which the sequences representing a character's state are (1) predictable (i.e. correlated over time), and (2) affected by those representing the environment's state (i.e. correlation from/to specific features and entities). Information theory is particularly suited to this kind of investigation, for reasons explained later in the Background chapter (Section 2.5).

It must be emphasised that the subject of study are fictional characters, and thus reflect cultural representations of personality rather than real people. However, the methodology could also be applied to non-fictional sources such as diaries or biographies. This work focuses on fictional characters as a first step since large datasets are easily accessible, and because insights on the information structure of novels are also relevant to the study of cultural aspects of personality [9] and for literary analysis [21, 22].

2

Theoretical Background

The reader may skip sections of this chapter depending on his background. Section 2.1 gives an basic overview of contemporary Natural Language Processing, sufficient for understanding the methods used here and the recent state-of-the-art. Section 2.2 elaborates on the various concepts and models from psychology relevant to this work. Sections 2.3 and 2.4 picture the specific research space in which this project takes place. Finally, Section 2.5 provides the definitions of all concepts and measures used in the analysis. It also serves as an introduction to Information Theory for unfamiliar readers, and should be comprehensible with elementary notions of probabilities and asymptotic limits.

2.1 Natural Language Processing

Natural Language Processing (NLP) and Understanding (NLU) is a research area concerned with the automated extraction and analysis of syntactic and semantic information from text written in human (natural) language. Earlier work relied on the use of formal grammars and heuristics for syntax parsing ("rule-based approach"), with semantic information usually compiled in manually constructed lexical databases, such as WordNet [23].

Modern NLP, however, is characterised by the *statistical approach*. It is based on distributional semantics [24], namely the hypothesis that two words have similar meaning if the context in which they are used (i.e. the distribution of surrounding words) is similar. This structure is encoded in an abstract vector space, or embedding, used to formulate many NLP tasks as typical inference tasks in Machine Learning (ML). The range of NLP tasks varies from low-level syntactic parsing (e.g. Part-of-Speech tagging) to high-level semantic understanding (e.g. sentiment analysis).

Recent breakthroughs followed the adoption of more advanced ML techniques and architectures, such as notably Long Short-Term Memory (LSTM), Attention, Transfer learning and Multi-Task Learning. While powerful, these techniques often require large amounts of training data and computational resources, which is why simpler rule-based models and lexical databases still remain useful.

The rest of this section introduces in more details the concepts and tasks relevant to this project.

2.1.1 Word embedding and language models

Word vector embeddings are generated by language models such as Word2Vec [25]. It is a well known neural network language model, illustrated in Figure 2.1, that predicts words from their immediate context. It became very popular because of its availability, efficiency and ability to capture both syntactic and semantic regularities while preserving linear relations between word meanings. As an example from the paper, $vector(\text{"}King\text{"}) - vector(\text{"}Man\text{"}) + vector(\text{"}Woman\text{"})$ is very close to $vector(\text{"}Queen\text{"})$. This allows many NLP tasks to be formulated in terms of mapping in vector space.
So-called "bag-of-word" approaches like Word2Vec average word frequencies without considering their sequential order. As a result, they fail to capture sentence-specific contextual information and distinguish between multiple word senses (polysemy). Another limitation is the inability to handle out-of-vocabulary words. More sophisticated architectures such as flair [26] or BERT [10] overcome these limitations by capturing temporal correlations between words or word parts.

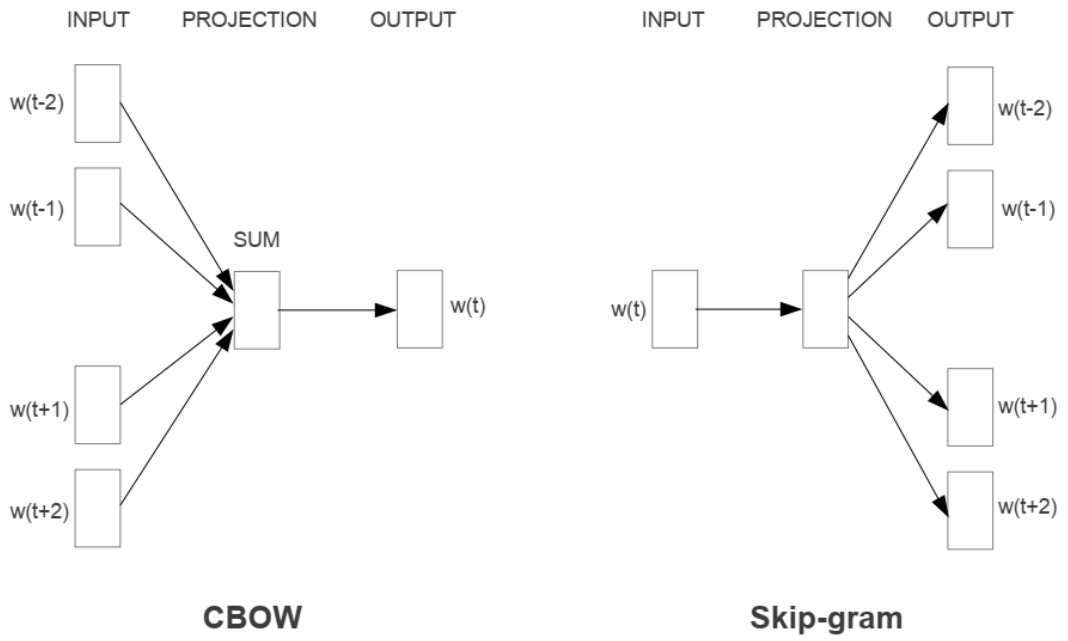

Figure 2.1: Word2Vec model architecture (from the original paper [25]). The CBOW (Continuous Bag-of-Word) architecture predicts the current word from its context (surrounding words), whereas the Skip-Gram architecture does the opposite: given the current word, it predicts the context.

2.1.2 Low level tasks

The first task in any NLP pipeline is *tokenisation*, which consists of segmenting a text into *tokens* (e.g. words, punctuation), and sometimes also into *spans* (e.g. sentences, compound names, noun phrases). Many tasks can be implemented as classification tasks on single tokens, such as the ones below. These consist in extracting syntactic-level features from documents, and are typically used as part of a more complex NLP pipeline to achieve some purpose.

- **Part-of-Speech (POS) tagging**: determining the grammatical role of a given token, e.g. noun, adjective, verb etc.
- **Dependency Parsing (DP)**: representing sentences as a syntactic dependency tree, where each token points to its parent (e.g. subject and object point to the verb), and the sentence parent is the main verb. See example in Figure 2.2.
- **Name Entity Recognition (NER)**: detecting entity names and classifying them into pre-defined categories (person, location, organisation etc.).

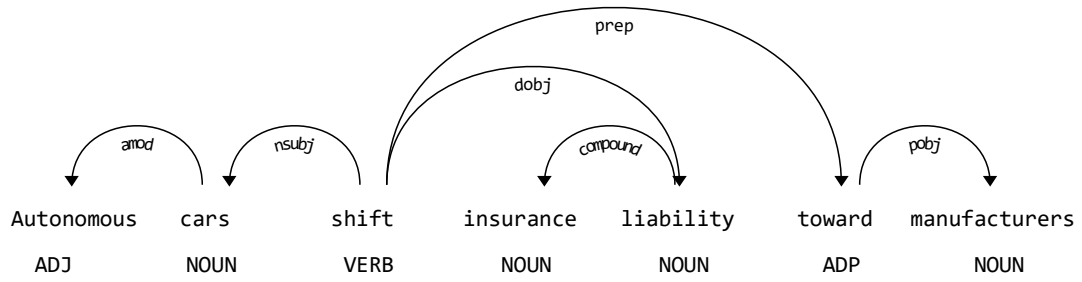

Figure 2.2: Example of Dependency Parsing (DP) and Part-of-Speech (POS) tags. Arrows illustrate syntactic dependency relations from parent to child with the corresponding relation type. Since every token must have a parent (the root is its own parent), the task is implemented as a token classification task. POS tags are also shown under the corresponding words. From spaCy's documentation[1].

The advantage of such low-level tasks is that pre-trained models with decent performance are widely available, at least for the English language. Since these features are typically combined into more complex algorithms, NLP applications are usually structured as a *pipeline*. A pipeline is a sequence of single-purpose tasks that use the accumulated outputs (e.g. lower-level features) of the previous ones to generate new output (e.g. higher-level features).

[1] `https://spacy.io/usage/linguistic-features#dependency-parse`

2.1.3 High level understanding tasks

The recent successes in applying advanced ML architectures to NLP has led to breakthoughs in addressing non-trivial higher-level tasks. Those typically require a certain level of understanding and interpretation at the semantic level.

Coreference resolution consists in binding words or expressions that refer to the same thing or person (referent). For example: "My sister has a dog. She loves him". "She" refers to "My sister" and "him" to "a dog". The problem is more complex than merely resolving pronouns. Consider the following example (indices represent distinct referents): "Although he$_1$ was playing with it$_2$, Alice and Bob$_3$ said they$_3$ could not hear his$_1$ guitar$_2$". The referent can occur before (1 and 3) or after (2) the reference, or include another referent (2), or the reference can have multiple referents (3), or even be ambiguous (1). Some of the difficulties related to coreference resolution are currently unsolved problems. This includes collapsing the different names of a person or character (e.g. "John Doe", "John", "M. Doe"), or dealing with plural pronouns or group membership (e.g. "they", "their") [27].

Semantic Role Labelling (SRL) refers to the identification and classification of predicate-argument structures. For example: "The keys$_{1,2}$, which were needed$_1$ to access the building$_1$, were locked$_2$ in the car$_2$". This sentence contains two predicates, or verbs: "needed" and "locked", respectively with the arguments ("the keys", "to access the building") and ("the keys", "in the car"), corresponding in both cases here to the agent (*who*) and a modifier (*why*, *where*).
Semantic frame induction, *slots-filling* and *event extraction* are variants where typically only a set of predefined predicate and related arguments (frames) are considered, based on the needs of the tasks or on databases such as FrameNet[2].

Sentiment analysis consists in classifying the general sentiment expressed in a sentence or document, typically positive or negative. As illustrated on Figure 2.3, sentiment is distinct from emotion as it reflects an acquired predisposition rather than a momentary affect. The difficulty of the classification task lies in ambiguities such as negations or irony. Emotion detection is typically associated with sentiment analysis [3]. Psycholexical approaches are commonly used in this context [28] by simply matching the words with a lexicon.

2.1.4 Other Machine Learning concepts

Supervised learning describes the process of fitting "a function that maps an input to an output based on example input-output pairs." [29]. In classification tasks, very common in NLP, the *classifier* function maps sample tokens to discrete *labels* or *annotations*. The set of examples, or gold standard, represents the contextual knowledge required to accurately reproduce the mapping, and is often built through a tedious manual process.

POS tagging, for instance, is a typical supervised learning task: labels (e.g. noun, adjective) are derived from a theory of grammar and provided explicitly in the gold standard ; whereas language modelling is unsupervised: by trying to predict e.g. the

[2]Available online on `https://framenet.icsi.berkeley.edu/fndrupal/`

next word given the previous ones, all the contextual knowledge contained in the words, which also includes POS information, gets encoded into the representation model.

Unsupervised learning does not rely on previous knowledge, but aims instead at capturing patterns. For example, by learning faithful representations of the data, or by automatically extracting the most relevant features from the inputs. It can be mixed with supervised models to provide better input features (representation learning), or to improve the generalisation of a model trained on a restricted labelled dataset to a larger non-labelled dataset (semi-supervised learning). Learned representations are helpful for implementing complex high-level supervised learning tasks with relatively few annotated samples [30], by using the compressed representation as input. This is the essence of **transfer learning**, which allows large generic models like BERT [10] to be re-used by changing only the output layer. When introduced, BERT outperformed multiple state-of-the-art task-specific classifiers with the same model, while using much less labelled data. BERT also uses the **attention** mechanism to dynamically adapt the representation of a specific token occurrence based on the surrounding tokens.

Multi-task learning architectures are designed to optimise the representation (i.e. language model in NLP) across multiple tasks simultaneously. This approach improves generalisability, which is particularly important for transfer learning.

Knowledge distillation is a technique for reducing model size by training a smaller model to reproduce the output of a larger one [31]. It is useful for compressing a large composite model resulting from transfer learning, or for expanding a smaller set of annotations into a bigger one.

2.1.5 State-of-the-art overview

Language modelling is an exceptionally active field of research at the time of writing, and is quickly pushing progress on many NLP tasks. The state-of-the-art in many tasks is characterised by the combination of transfer learning with an attention-based language model [10], further improved with multi-task learning [32]. An interesting property of such language models is that all common lower or higher level NLP features seem to be integrated and optimised across the different levels of abstraction [33], despite never being explicitly encoded in either the model structure or the data. These breakthroughs enabled the tackling of increasingly complex higher-level tasks such as text summarising, question answering or common sense reasoning [34, 35]. However, despite the many promising applications, important limitations apply, as discussed in 2.1.6.

Probabilistic Graphical Models (PGM) are a common alternative to deep neural networks in NLP. A Bayesian PGM is a directed graph of inter-dependent random variables encoding knowledge and uncertainty as a probability distribution, with parameters fit to empirical data using Bayesian inference. A typical approach in NLU tasks is to represents words as the output of a generative model that explicitly includes the relevant latent variables, such as sentiment or semantic role. This

makes interpretation straightforward compared to black-box neural network models. Learning is unsupervised: annotations are not required, although they can be added as additional features.
There are other kinds of PGM but Bayesian models are commonly used in NLP. PGMs have been applied with varying levels of success to event detection and frame induction [36, 37, 38], sentiment analysis [36], language modelling [39], Semantic Role Labelling and character persona modelling [40].

2.1.6 Current issues in Natural Language Processing

As in other areas of Machine Learning, the use of deep neural networks comes with non-negligible issues, essentially due to two important limitations. Firstly, their black-box structure makes them hard to interpret. This recently led to concerns about their effectiveness for practical NLU applications, as they are prone to learn shallow heuristics that perform well on benchmarks but systematically fail in more subtle cases [41, 42, 27].

Another practical limitation is the large model size, requiring considerable computational resources for both training and execution, as well as large amounts of annotated data. Although this can be drastically reduced with transfer learning [30] and distillation [31], the linguistic resources required for training the initial model are still out of reach for many languages other than contemporary English.

These limitations complicate the adoption of modern NLP models outside the domain of common large corpora, composed mostly of English news, social media and web content. Despite recent efforts [43, 44, 45], good quality resources and models are still lacking at the moment to address the specific challenges of non-contemporary literature [27].

2.2 Psychological meaning and models

2.2.1 Factor-based models of personality

Factor Analysis (FA) is a dimension reduction technique often used in psychology research, which decomposes the sample variance into a fixed number of uncorrelated and normally distributed latent variables or factors. FA has been used in psycholexical studies to determine the main dimensions of inter-individual variation in personality, based on the lexical hypothesis, as mentioned in Section 1.2.

This line of research originated in seminal work from Allport and Odbert [46]. They built a lexicon of English personality *trait names*, classified in different categories then mapped onto a smaller set of scales for self-report studies [7]. These studies led to a consensus about the Five Factor Model (FFM) [6], also known as the Big Five. Each factor represents a spectrum along a specific, relatively stable personality trait, the meaning of which is interpreted from the words most/least associated with it.

The five factors are defined as follows:

1. *Openness*: interest for new ideas or experiences
2. *Conscientiousness*: self-discipline, orderliness
3. *Extroversion*: enthusiasm, social engagement
4. *Agreeableness*: compassion, politeness, cooperativeness
5. *Neuroticism*: emotional instability, tendency towards negative emotions

Other well-known models based on a similar methodology include MBTI [47] and HEXACO [48].

As previously discussed, one limit of such purely lexical model is the assumption that the manifestations of these traits are not affected over time by the situational circumstances. One would expect a more comprehensive model to make the relation between one's past and present environment and one's current personality more explicit. Indeed, the effects of an individual's environment over time on his/her personality has been stressed in multiple empirical studies. Twin studies suggest that 40% of individual differences are explained by genetic factors, leaving 60% to environmental influences [49]. While the importance of the situational context in which the traits manifest is recognised, the structure and long-term effects of situation characteristics is not well understood [2, 19]. Traits tend to stabilise over time [14], but major life events can also induce permanent change [50, 19].

As mentioned in 1.2, in addition to the complexity left unaddressed by the factor model, the validity of FFM outside of Western populations has recently been questioned [1]. This further motivates a more behavioural approach focused on correlations across time and situations, rather than on specific dimensions that depend on linguistic interpretations, which this work aims at exploring.

2.2.2 Dynamical Systems Theory in Psychology

Dynamical Systems Theory has been used in many areas of psychology. Here we are interested in formalising a system of quantifiable interactions between the state of the individual and that of his external environment. Personality would then be conceptualised as the average trend of the former.

Mischel and Shoda [12] proposed the mostly theoretical Cognitive-Affective Personality System (CAPS), a feedback loop structure between situation "inputs" and behaviour "outputs", the latter being determined by an internal (mental) network of *mediating processes*. The specifics of these internal processed are left open, but the model illustrates the recurrent nature of the interaction between situations and behaviours.

Modelling the internal cognitive processes that determine behaviour has been an objective in multiple research areas of psychology and cognitive science. The most empirically relevant approach to personality theory is perhaps Reinforcement Sensitivity Theory (RST), based on Jeffrey Gray's seminal work on the biology and evolution of brain-behavioural systems [5]. This behaviourist approach models cognitive processes as computation from stimulus to behaviour based on three interacting processes that mediate responses to punishments and rewards. These processed are described in Table 2.1.

System's name	Reacts to	Controls
Behavioural Activation (BAS)	Rewards	Approach / Arousal
Flight-Flight-Freeze (FFFS)	Punishment	Fight / Flight / Freeze
Behavioural Inhibition (BIS)	Goal Conflicts	Inhibition / Arousal

Table 2.1: Mediating processes in Reinforcement Sensitivity Theory, one of the notable dynamical models of personality. Each system reacts to a certain type of stimulus and controls a certain type of behaviour. The response is "computed" from the activation level of these behaviour types.

RST predicts that observed personality traits are primarily determined by the sensitivity and output of those systems. While it has the advantage of allowing formal implementations and simulations, its relevance is questioned as those central sensitivity characteristics are not available for introspection. Also, it is uncertain whether this neural-level processing structure modelled after animal brains can appropriately explain the complexity of psychological phenomena [5].

Read et al. [13] synthesised CAPS, RST, FFM and other work on neurobiology, goal-based models and evolutionary approaches, into a neural network model with a similar structure. Although more comprehensive than its predecessors, this model was only evaluated qualitatively with manual parameter tuning, and thus remain theoretical.

We may also mention Friston's Free Energy Principle [51], stating that the internal (mental / hidden) states of an organism encode a model of its environment, updated through a Bayesian inference process. The feedback loop between internal and external states previously discussed is extended here with active inference. That is, the system (here the individual) changes its configuration to affect both how it samples (action) and encodes (perception) the environment. Hence, perception is not passive but has its own feedback loop. Again, this *principle*, while elegantly connecting cognition, thermodynamics, Bayesian inference and Information Theory, is difficult to evaluate empirically.

The takeaway is that these theoretical models of long-term affect and/or behaviour, despite their lack of empirical relevancy, seem to agree on an environment-cognition feedback loop. Importantly, feedback may go in any direction, including internal feedback. But we expect most relevant information, as personality theory is concerned, to come from the perception-cognition-action loop, to the extent that it is measurable. Which, of course, is the core challenge.

2.2.3 Emotional Brain Theory as information processing

Characterising the range of possible behaviours is a difficult and potentially intractable problem in itself. However, it can be argued that behaviour reflects present and past affect associated to environmental stimuli. Hence, we can expect correlations over time between affect and behaviour. This is supported by the emotional brain theory [52], which suggests that affect (the pre-conscious experience of emotions) is the primary brain mechanism for guiding behaviour towards homeostasis, in humans and many other living organisms. In other words, affect determines the behaviour most likely to ensure survival.
This does not mean that behaviour is deterministic. But through mechanisms of adaptation and reinforcement, the situation-affect-behaviour feedback loop can be seen as a circular information processing system, where affect information is at least partly preserved within behaviour information.

This is why in this project we focus on affect. Because in this view, behaviour and cognition, the two other basic dimensions of personality, tend to manifest consequences of affect. It simplifies the problem of capturing behaviour and cognition, but assumes that these are predicted by affect. This is important to keep in mind when interpreting the results in terms of personality structure.

To measure affect from text in this work we use an emotion detection method. This adds the assumption that affect and emotion are also sufficiently correlated, since they represent different concepts and should not be confused. Figure 2.3 illustrates the definition and relations of these constructs.

From now on we may refer to affect or emotion interchangeably, since we focus on emotion analysis.

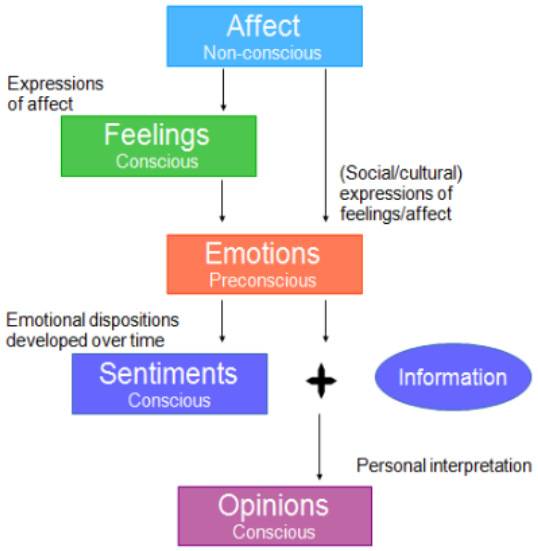

Figure 2.3: The relation between affect, feeling, emotions, sentiments and opinion (taken from [3]). Affect is the umbrella term for the underlying physiological phenomena, preceding conscious identification. The conscious experience is a feeling, and once recognised and identified it becomes an emotion. It may then be expressed as such and/or internalised over time as sentiments or opinions. Affects and feelings are universal, whereas emotions are cultural: their description and manifestation is influenced by personal experiences, culture and social norms. In language, emotional words can then describe sentiments, opinions or feelings.

2.2.4 Situational context

One way to integrate the environmental (or situational) context in psychological models is to identify general situation characteristics, such as positive or negative valence, adversity, typicality or importance [53, 18]. Those are determined either from the physical environment, or from the interaction with other individuals. It should be noted that such characteristics are evaluative in nature, aiming to describe a subjective psychological situation rather than an objective one [53].
Taxonomies of psychological situation are a relatively recent development, with two notable psycholexical models. DIAMONDS (2014) [18] provides 8 dimensions: Duty (D), Intellect (I), Adversity (A), Mating (M), Positivity (O), Negativity (N), Deception (D) and Sociality (S). Parrigon (2017) [53] identified 7 somewhat similar dimensions across different methodologies, referred to as the CAPTION model: Complexity (C), Adversity (A), Positive Valence (P), Typicality (T), Importance (I), Humour (H) and Negative Valence (N). Note the limitations discussed about psycholexical approaches to personality also apply.

We intended to use the CAPTION model in this work, but unfortunately could not get access to the full lexicon. However, this illustrate what the relevant dimensions might be and how they could be measured in future work. Another model was eventually used here as a replacement, as explained in Section 3.2.1.

2.3 Challenges of text mining for psychological meaning

The most widespread text analysis tool used in personality research for extracting psychological meaning is Linguistic Inquiry Word Count (LIWC)[3], a proprietary software consisting of counting word frequencies from various dictionaries of relevant psychological constructs, such as affect, cognitive processes, needs etc., as well as other grammatical features. Such tools are typically used to analyse the psychology of the text's author automatically. However, this approach is prone to bias due to the selection and interpretation of words in the dictionaries used, and to classification errors due to polysemy (words with multiple possible meanings), negations and other qualifiers, irony and other artefacts of natural language.

Supervised learning methods have also been used for automated personality recognition from text, or *author profiling* [54, 55], using diaries or social media status in combination with personality surveys as training data. In addition to the shortcomings of the underlying Five Factor Model, these models also tend to be biased due to a lack of diversity in the training data, which can be difficult to notice at first given their black-box nature [41, 27].

Lexical databases (e.g. WordNet [23]) and Latent Semantic Analysis are also commonly used for personality and emotion detection [56]. The former, requiring manual construction, is subject to the same limitations as personality lexicons, and the latter also lacks the resolution to overcome bias from language artefacts.

[3] `https://liwc.wpengine.com/`

The core problem of these approaches comes down to finding how much contextual knowledge needs to be attached to the text. This includes dictionaries, symbolic or numerical annotations of lexicons, the training data used to fit a statistical model, and the annotations in the case of supervised learning. A general methodology for identifying the appropriate level of context is to use a common representation for lower and higher levels of abstraction, and identify information structures (e.g. bursts, recurrent patterns) that only occur above a certain level [57, 58]. This is why we chose in this work to investigate information-theoretic properties from minimal semantic features.

2.4 Similar work

The purpose of this project being highly interdisciplinary, we found previous work on similar problems scattered across different disciplines and using a wide range of different methodologies. The trend, if there is any, goes towards task-specific models, with a similarly purpose-built data extraction pipeline when sophisticated NLP features are involved. So we chose to use elements of these previous work as inspiration and to develop a custom pipeline, in order to avoid introducing unfamiliar or inappropriate models.

2.4.1 Personality modelling

Much work has focused on predicting personality traits from language use, using word count [59], various supervised learning approaches [60, 54, 61, 62], or unsupervised approaches like FA [63]. Such approaches are popular in psychology research and online user targeting. However, in addition to the methodological issues already discussed in Sections 1.2 and 1.3, they are also subject to bias due to domain-specific characteristics of the training data [64, 65] or the natural drift of word use towards popular words [66].

There has also been attempts at measuring the personality of fictional characters. Flekova et al. [67] used available MBTI-based polls on the personality of popular book characters to evaluate text classifiers for extraversion (which correlates well with the corresponding scale in FFM), applying various lexical and semantic features. They reported that action and appearance was more predictive than direct speech, consistent with other studies.

An unsupervised approach was implemented by Bamman et al. [40] using a hierarchical Bayesian model applied on discretised word embeddings, with author, prior word distribution and character persona as latent variables. Evaluation was performed by comparison with known relations of similarity (e.g. “Character X is more similar to Y than X or Y is to Z”). They found no alignment of the resulting personas with character types known in the literary analysis literature, but did find correlations within literary genres, social categories and gender.

Johnson et al. [65] rated the personality of fictional characters from Victorian novels using the FFM scales, in order to investigate the “implicit theory of personality and human nature” embedded in literature. They concluded to a good agreement with

modern personality psychology, but found an over-representation of trait agreeableness, suggesting a bias towards cooperative behaviour and language.

More similar to dynamical approaches previously described, Narang et al. [68] modelled user activity profiles as Gaussian Hidden Markov Model, with data from two domains (academic publications and activity on a large online platform). They identified meaningful "archetypes" of behaviour with specific progression stages over time.

2.4.2 Text understanding

Kim and Klinger [45] endeavoured to create a corpus of annotated novels, including emotions and events linked to characters. They reported difficulties, in line with previous work, due to the subjectivity of emotion interpretations leading to poor agreement between annotators.

Zhang et al. [69] developed a complete NLP pipeline for generating animations from screenplays, featuring the main entities, their environment, their actions and additional information such as speed, distance or emotions. They first apply a text simplification model so each sentence describes only one action. Then entities and dialogues are resolved using screenplay annotations and coreference resolution. Words describing actions are mapped to a list of pre-defined animations, using their similarity through WordNet [23]. Finally, Semantic Role Labelling is used to extract all contextual information into a key-value store. The main limitation of this work is its scope, as it is meant more as a productivity tool for artists than a rigorous data mining method.

2.5 Information Theory

Many standard statistical methods rely on assumptions that often do not hold in complex systems like human behaviour or natural language. Commonly assumed properties include linearly correlated variables, thin-tailed probability distribution with well-defined mean or memoryless processes. In contexts where such methods fail, Information Theory provides measures to quantify similarities and interactions between variables requiring minimal assumptions.

Shannon's Theory of Information is a probabilistic theory of uncertainty, where information is understood as surprise. An event, such as observing a system in a certain state, or reading a certain message from a symbol string, is informative if it is unexpected: there is nothing new to learn about an event that was already predicted. Since information is a property of a probabilistic representation, it is distinct from meaning, which is the context-specific interpretation of the message or state itself [70].

2.5.1 Basic notions

Given an observable X with a probability distribution $P(X = x_i) = p_i$, $i = 1, 2, ..., n$, the *Shannon information content* of the symbol or state x_i is[4]

$$I(x_i) = \log_b \frac{1}{p_i} \tag{2.1}$$

With a base $b = 2$, information is given in *bits* and can be interpreted as the minimum number of yes/no questions to ask to determine the state of the system [71]. Any other base can be also used.

The central concept of information theory is *entropy*, which quantifies the uncertainty about the state of a system by measuring the expected information gain from an observation:

$$H_X = \mathbb{E}\left[I(x_i)\right] = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} \tag{2.2}$$

A fully deterministic model $p_i = 1, p_j = 0 \, \forall \, j \neq i$ has no uncertainty and therefore produces no information: $H_X = 0$. On the other hand, entropy is maximal when all outcomes are equally likely: $p_i = \frac{1}{n} \Rightarrow H_X = \log n$. Complete disorder or randomness thus imply maximal information gain on any observation.
Entropy is additive between independent information sources. This property is useful for detecting correlations.

$$X \perp Y \implies H_{X \cup Y} = H_X + H_Y \tag{2.3}$$

Many other information-theoretical measures are defined as the information gained, or reduction of uncertainty, between a previous model $P^{(0)}$ and a new model P. This quantity is called the *Kullback-Leibler divergence*, or *relative entropy*:

$$D_{KL}(P^{(0)}||P) = \sum_{i=1}^{n} p_i \log \frac{p_i}{p_i^{(0)}} \geq 0 \tag{2.4}$$

2.5.2 Information Dynamics

Recent research in complex dynamical systems has focused on mechanisms by which information flows over time within and between individual components, as in the firing of neurons in the brain [72] or the diffusion of information in social media [73]. Because of the dynamical nature of such patterns, information structures that do not otherwise appear in static data can be revealed, such as new unexpected patterns (information production), pattern reproduction and recurrence (information storage) or interactions between components (information transfer).

[4] To be absolutely rigorous, we should write $I(P(X = x_i))$ as it is a property of the probability, not of the state itself. But this shorter notation is commonly used for simplicity.

Information dynamics refers to information-theoretic concepts that focus on such temporal information structures, in systems composed of parts that "interact to create non-trivial computation where the whole is greater than the sum of the parts" [74].

Consider an infinite sequence of states $\{x_0, x_1, x_2, ...\}$, approximated by a Markov process of order m, $X^{(m)}$. This means we assume that the next state x_t depends only on the m previous states (Markov property) [71]:

$$p(x_t|x_{t-1}, ..., x_{t-m}, ...) = p(x_t|x_{t-1}, ..., x_{t-m}) \tag{2.5}$$

In such symbol sequences, we need to consider the temporal structure of information, as it is spread across different time scales. This is captured by measuring the probability distribution over blocks of length m. Using the shorthand notation $x_t^{(m)} = (x_t, ..., x_{t-m+1})$, we define *block entropy* as follows:

$$H(m) = \sum p(x_t^{(m)}) \log \frac{1}{p(x_t^{(m)})} \tag{2.6}$$

We can fully characterise the information structure of the underlying process only in the limit of infinite block length. Given the temporal dimension, this structure now has two aspects: *entropy* and *complexity*. Here entropy refers to the average rate of information *production*: how frequently we observe new unexpected patterns, which is associated with randomness ; and complexity can be seen as the extent to which such patterns are preserved and re-occur, which can be described as information *storage* [75].

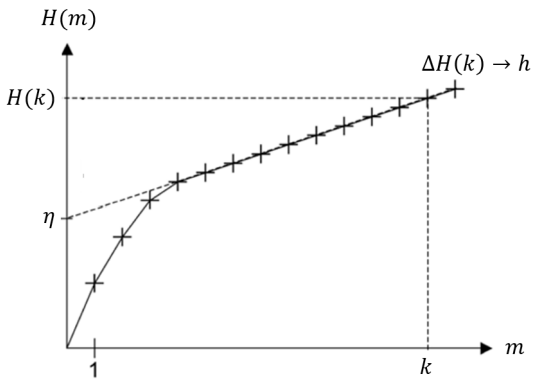

Figure 2.4: Block entropy $H(k)$ approaches an asymptotic line with slope h representing the entropy of the sequence or *entropy rate*, and an intercept η representing *excess entropy*, a measure of complexity. Adapted from [70].

These concepts are formalised in the asymptotic properties of block entropy as a function of the length m. As shown in Figure 2.4, $H(m)$ increases monotonically with m, approaching a line with slope h called the *entropy rate* or simply entropy:

$$h(m) = H(m) - H(m-1) \tag{2.7}$$

$$h = \lim_{m\to\infty} h(m) \tag{2.8}$$

$$= \lim_{m\to\infty} \frac{H(m)}{m} \tag{2.9}$$

Complexity is characterised by the intercept of this asymptotic line, called the *excess entropy* η. It can also be understood as the rate of convergence of entropy rate:

$$\eta(m) = H(m) - m\,h(m) \tag{2.10}$$

$$\eta = \lim_{m\to\infty} \eta(m) \tag{2.11}$$

These two complementary measures can be visualised on a complexity-entropy diagram to reveal the 2-dimensional structure representing the intrinsic information processing embedded in different processes. This allows the comparison of different kinds of system, and may provide clues for finding an appropriate class of model given empirical data [75].

So far we focused on properties of single random processes, but we are also interested in the interactions between processes. In information-theoretic terms, these manifest as information transmitted from one to another, and it requires accounting for past states of both processes.

Consider two processes X and Y and assume current and past states of Y have no influence on next state of X [76]:

$$p(x_{t+1}|x_t^{(k)}) = p(x_{t+1}|x_t^{(k)}, y_t^{(l)}) \tag{2.12}$$

We can use Kullback-Leibler divergence to measure the deviation from this assumption. The resulting quantity represents the information flow from Y to X, as illustrated on Figure 2.5. It is called *transfer entropy* [76]:

$$T_{Y\to X}^{(k,l)} = \sum p(x_{t+1}, x_t^{(k)}, y_t^{(l)}) \log \frac{p(x_{t+1}|x_t^{(k)}, y_t^{(l)})}{p(x_{t+1}|x_t^{(k)})} \tag{2.13}$$

In the limit of infinite k and l, transfer entropy generalises entropy rate to multiple processes. It is worth pointing out that it measures an observed correlation and should not be treated as evidence of causality [74]. It is better interpreted as the extent to which observing a process helps predict another, even in the presence of complex non-linear phenomena. Unlike other measures of correlation and mutual dependence, transfer entropy is asymmetric: it also captures which of two variables has more influence on the other.

Because reliably estimating $T_{Y\to X}^{(k,l)}$ from data requires a lot of samples (significantly more than for $H_X(k)$ because of the condition on Y), it is common practice in the literature to perform a test of significance in support of the null hypothesis $\mathcal{H}_0$ of independence (2.12) [71]. A surrogate distribution under $\mathcal{H}_0$ is obtained either analytically or using some sub-sampling technique, while preserving the statistical properties of the original data. We can then obtain a p-value of the probability that our measurements are sampled from this surrogate distribution.

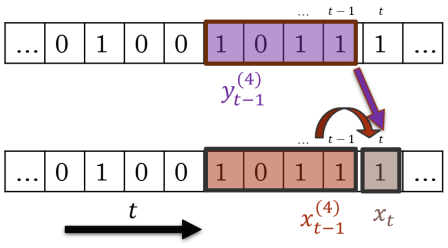

Figure 2.5: Illustration of transfer entropy (2.13).

3
Methods

The information extraction procedure generates sequences of psychological states from the text of a book, for every character identified as such, using universal low-level features such as POS and DP. This was done by matching emotional words and using dependency parsing to find the subject or object they refer to. The resulting entity-predicate tuples were associated to either *character* or *environment*. These sequences were then transformed into a symbolic form representing the occurrences of emotional states throughout the book, for every character and its environment, and every emotion feature. Finally, the information-dynamical formalism presented in Section 2.5 was used to investigate the temporal structures and interactions of this character-environment system. Because no similar analysis was found in the domain literature, the validity of the approach was assessed by comparing the results to those obtained with baselines representing only parts of the data extraction pipeline.

The method was applied on a corpus of 150 English novels published in txtLab's *Novel450* dataset [77], comprising books from 98 authors written between 1770 and 1910.

This chapter is structured as follows: Section 3.1 presents the general methodology ; Section 3.2.5 describes the important features used ; Section 3.3 elaborates on how the data extracted is transformed into binary strings ; Section 3.5 details the methods and challenges of entropy estimation from these strings ; finally, Section 3.6 explains the baselines used for validating the significance of the results. Limitations and improvements are discussed in the next chapter, in Section 4.5.

3.1 Procedure overview

Given the concerns discussed in Section 2.1.6 about high-level black-box models in NLP, and the limited scope of the project, we decided to perform the information extraction with a custom heuristic algorithm using readily available lower-level NLP models and psychological lexicons, rather than to train and evaluate a new task-specific supervised model.

The information extraction pipeline is the following:

1. **Tokenisation**: each document is split into tokens (words and punctuation).
2. **Extraction of low-level linguistic features**: Part-of-Speech (POS), Dependency Parsing (DP), Name Entity Recognition (NER), using existing pre-trained models from the spaCy library[1]. Hypernyms (word meaning categories) are also introduced from WordNet.
3. **Coreference resolution**: finding and resolving all mentions of the entities, using a pre-trained model from the NeuralCoref library[2].
4. **Lexicon matching**: lemma-based matching of tokens from psychological lexicons (*NRC VAD* and *NRC Emotion*).
5. **Semantic parsing**: heuristic based on DP and resolved entities to determine semantic roles and generate semantic frames as tuples: (position, frame, frame arguments).
6. **Entity type classification**: a heuristic to determine whether the entity is classified as *environment*, *person* or *unknown*, based on NER, POS and hypernyms.
7. **Mapping semantic frame to discrete symbols**: based on semantic roles and psychological features in the frame arguments, resulting in sequences of *Stimuli* (as analogy for information flow from environment to individual) and *Responses* (as analogy for information flow in the opposite direction).
8. **Binary representation**: sequences are mapped to binary strings for every character and feature.

The outcome of steps 1–5, exemplified in Table 3.1, was a sequence of predicate-entities records (semantic frames) containing the position in the text, the predicate, the name of the associated entities and their semantic role in the sentence (agent or patient), and a discrete vector representation of the features (psycholexical emotion dimensions) associated with the predicate. Additional processing (steps 6–8) was done to obtain binary strings as shown in Table 3.2.

The relevant information-theoretic measures were then estimated from these binary strings to investigate the questions formulated in the introduction (Section 1.5). The analysis was done according to the following steps:

1. **Single features analysis**: estimation of block entropies (per sequence, character and document) and investigation of the complexity-entropy information structure of each sequence.
2. **Feature interaction analysis**: estimation of transfer entropies between pairs of feature categories (per pair of sequences, character and document) and their statistical significance, group pairs by feature categories and find out which have information transfer and under what conditions.
3. **Validation**: repeating the analysis with different baseline methods, i.e. the same processing with some of the features or steps removed. Differences in the resulting information structure would indicate information structure added by these features or processing steps.

[1]Model `en_core_web_sm`, library version 2.1.3. URL: `https://spacy.io/`

[2]Model `en-coref-md v3.0.0`, library version 4.0. URL: `https://github.com/huggingface/neuralcoref`

Time	Predicate	Agent	Patient	VAD	Emotion
4	fellow	Joe	-	Valence, Dominance	Trust
4	fellow	Me	-	Valence, Dominance	Trust
4	sufferer	Joe	-	Arousal	Fear, Sadness
4	sufferer	Me	-	Arousal	Fear, Sadness
9	confidence	Joe	-	Valence, Dominance	Trust, Joy, Fear
9	confidence	Me	-	Valence, Dominance	Trust, Joy, Fear
15	confidence	Joe	Me	Valence, Dominance	Trust, Joy, Fear
...	...	...	...	...	...

Table 3.1: Example of extraction of semantic frames from the sentence "Joe and I being fellow-sufferers, and having confidences as such, Joe imparted a confidence to me, the moment I raised the latch of the door and peeped in at him opposite to it, sitting in the chimney corner." (from *Great Expectations* by C. Dickens). Each row represents a semantic frame based on the occurrence of a predicate at a certain position in the text (*Time*), associated to an *agent* and/or *patient* entity in the sentence, and to binarised psycholexical features in the lexicons. Some predicates are duplicated here because "Joe and I" is the subject hence both "Joe" and "I" (resolved to "Me") are agents, except in the last row (from "Joe imparted a confidence to me") where "Me" is *patient* with regard to the *agent* "Joe" and fills the second argument of this predicate.

Time	Subject	Valence (Stimulus)	Valence (Response)	Joy (Stimulus)	Joy (Response)	...
4	`NARRATOR`	1	1	0	0	...
4	`NARRATOR`	0	0	0	0	...
9	`NARRATOR`	1	1	1	1	...
15	`NARRATOR`	1	1	1	1	...
...	...	...	...	...	...	...
4	`Joe`	1	1	0	0	...
...	...	...	...	...	...	...

Table 3.2: Example of symbolic representation corresponding to the example semantic frames from Table 3.1. In this case the sequences for *Stimulus* feature and *Response* features are the same because the predicates refer to both entities. Entropies are then estimated from each columns, with every new subject marking the beginning of a new sequence (thus a new sample).

3.2 Feature extraction

The main reason for the sophisticated NLP pipeline described here was to provide a distinction between the state of a character and that of its environment. The environment comprises other characters as well as locations, objects or anything else mentioned in the text.
Doing this distinction reliably requires a high-level semantic understanding of the text, beyond coreference resolution. This was implemented with a SRL heuristic that associates each *predicate* (i.e. word matched in the lexicon) to its *agent*, and *patient* if any. This is conceptually similar to previous work using typed dependency relations [40]. Following the SRL, another heuristic algorithm was used to classify the agents and patients found into either character or environment.

3.2.1 Lexicon matching (psycholexical features)

As discussed in 2.2.3, affect has a predominant role in determining short- and long-term behavioural tendencies. This is convenient as the state space of affect is much more restricted than that of behaviour or cognition. The open-ended complexity of the task, namely determining the space of psychological states and how to infer those states from text, was then reduced to an affect detection task.
Emotion (or affect) detection is far from a trivial endeavour and is subject to many subtleties and open questions [45, 78, 79]. A psycholexical approach was used in this work, consisting of matching words from the text with a lexicon. This is a common approach so we considered it a good starting point. But of course, the limitations of psycholexical approaches mentioned in Sections 1.2 and 2.2 also apply here, and other approaches from the literature may be considered in future work.

Affect is expressed in language as emotions, which are not culturally universal. There is no consensus on a universal classification of emotions, which is a reason why two different lexicons were considered for this work.
The first is the NRC Emotion Lexicon [28], which comprises 141,820 words annotated with 8 basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and their polarity (positive, negative), which is arguably the most comprehensive English lexicon available for common emotions.
The second is the NRC VAD Lexicon, from the same author [80], containing 20,007 words classified along a three-dimensional spatial model of emotions and general word meaning: valence (positive-negative dimension), arousal (active-passive dimension) and dominance (dominant-submissive dimension).
Both datasets were obtained through large-scale surveys using crowdfunded platforms, allowing for a larger lexicon and wider demographic sample than commonly used in other psycholexical studies.

Another reason for using the VAD lexicon specifically is the general word meaning representation it provides, which is useful for characterising situations more comprehensively than with emotions which may not always apply well. See Section 2.2.4 for a more detailed discussion of the relevant dimensions of situation description.

In the implementation, the words from the lexicon were matched to the document's words in a case-insensitive way, by comparing either the token's raw text or its lemma. The lemma is the base form of a word, for example: *look* (*looking*, *looked*, ...), *produc* (*production*, *product*, ...).

3.2.2 Entity merging

Because coreference resolution is not always sufficient for properly merging all mentions of an entity into the most relevant name, tokens corresponding to the same entity were merged based on NER tags (e.g. "Mr. Smith", "John Doe"), using a function provided by spaCy[3].

3.2.3 Semantic role labelling

This processing step was formalised as a semantic frame inference problem. Every predicate, that is, every word matched in the lexicon, was associated to an entity with the semantic role of *agent*, and possibly another with the role of *patient*.
For example, the sentence "Alice pets the dog. The dog is happy." has two predicates "pet" (agent: "Alice", patient: "the dog") and "happy" (agent: "the dog", patient: none). It's also possible to have a patient but no agent, as in "The dog is being pet". Finally, it is also possible to have multiple agents or patients, e.g. "Alice and her brother pet the dog" (agents: "Alice", "her brother").

Our implementation is based on as a custom two-pass heuristic algorithm using the DP tree and the coreference graph. The pseudo-code is given in Appendix A.1. The general idea is that certain DP relations are indicative of an agent role (e.g. noun-subject), and that the parent-child relation in the DP tree is similar to an agent-patient relation[4]. Hence, tokens identified as agent with regard to a root token can be propagated downwards to the descendents.
A first "upward" pass assigns the appropriate agent/patient relationships, for every coreference in the document with regard to the token's ancestors in the DP tree. Then the "downward" pass iterate through all the tokens and inherits agents from its ancestors.
Additional checks were added so that whenever possible, only the closest and most relevant agent (or patient) gets assigned, rather than all agents and patients from the sentences.

This method assumes that the grammatical structure of the text can be framed as a sequence of propositions about entities, and that the grammatical subject reflects the actual agency relationship between an entity with regard to a proposition. This is a rough approximation, and a fairly bold assumption. This limitation is one of the aspects of the method discussed in Section 4.5 that would need to be investigated in more detail.

[3] `https://spacy.io/api/pipeline-functions#merge_entities`

[4] For some DP relations this direction is reversed. For example, *nsubj* (noun-subject) points to the root (i.e. the verb), rather than the opposite.

3.2.4 Time resolution and sub-sentences

There are different ways to put the semantic frames extracted from the text into a sequence that properly reflects its overall chronology and thus the temporal structure. Of course, not all novels follow a linear chronological structure, but recovering the actual chronological structure of the narration would be beyond the scope of this project, so we had to assume they generally do.

Another challenge, possibly more specific to literary artworks, is related to the often long sentences, composed of *sub-sentences* separated by punctuations (“,”, “;”) or conjonction words (“and”, “then”...), or dialogues (e.g. “’The weather is nice’, he said”). Dialogue detection, and the proper handling of different narration types (i.e. first-person or third-person perspective) are additional non-trivial NLP tasks [27] and were not addressed in this work.

The problem then, if we intend to stay close to the semantic sequence and use the sentence position as a unit of time, is that all the predicates and related semantic frames of the sentence become merged together. So we decided instead to split sentences into their sub-sentences and use sub-sentence position as a timestamp.

The Sub-sentence detection mechanism was implemented as part of the SRL algorithm. When iterating the dependency tree upwards to get the ancestors of a given token, the iteration stops if the root of the sub-sentence is reached, that is, a token with a DP tag which is absent from a white-list of tags that represent relations within a sub-sentence. The list is shown in table 3.3.

`acl`, `advcl`, `amod`, `appos`, `attr`, `aux`, `auxpass`, `compound`, `conj`, `csubjpass`, `dislocated`, `dobj`, `iobj`, `neg`, `nmod`, `nsubjpass`, `obj`, `obl`, `orphan`, `pobj`, `poss`, `prep`

Table 3.3: White-list of DP tags for relations within a sub-sentence. See reference in Appendix A.3

3.2.5 Entity classification

Entities were classified into either character or environment. The environment of a book character comprises non-person entities such as groups, locations, objects etc., as well as all other book characters. This means that the environment differs for every character: each character is represented by one *subject* state sequence and one *environment* state sequence (also for every lexical feature).

This classification was implemented as a rule-based heuristic using POS, NER and WordNet [23]. Because the models used do not attempt to classify pronouns (e.g. *he*, *herself*) and common nouns (e.g. *lord*, *girl*), these were resolved using a manual list of exceptions given in Table 3.4. The rules are illustrated as a flow-chart in Appendix A.2. In some cases the entity a pronoun refers to may be ambiguous (e.g *they*), and this may still not be resolved after coreference resolution. Such cases are classified as *Unknown*.

Category	Words
Environment	*it*, *this*, *that*, *its*, *itself*, *something*
Narrator (person)	*I*, *me*, *myself*, *my*, *mine*, *we*, *us*, *ourselves*, *our*, *ours*
Reader (person)	*you*, *yourself*, *your*, *yours*
Person	*he*, *him*, *himself*, *his*, *she*, *her*, *herself*, *hers*, *man*, *boy*, *sir*, *woman*, *girl*, *madam*, *miss*, *lord*, *lady*
Unknown	*they*, *them*, *themselves*, *their*, *these*, *those*

Table 3.4: Words on the right were classified as the category on the left. Since this is done on mentions of entities in the text, which are spans (one or more subsequent tokens), the span's root in the dependency tree was used for matching, in a case-insensitive way.

3.3 Sequence processing

The feature extraction pipeline described above produced, for every book in the corpus, a sequence of time-stamped semantic frames with a single word predicate, an agent entity, and an optional patient entity. This was the first objective of the work as formulated in Section 1.5. To get to the second part, which was the analysis of the dynamical properties of character-environment interactions, the sequences of frames were mapped onto binary strings on which information-theoretical measures can be computed, as follows:

- **Entity filtering**: only entities occurring at least 100 times in the book were selected. Others were replaced by a generic name corresponding to their category (person, environment or other).
- **Semantic to Cognitive Roles**: semantic roles, i.e. *agent* and *patient*, indicate active and passive forms which we use to denote mental state and perception of the environment, respectively. Using a neurological analogy, we refer to the latter as *stimulus* features, and the former as *response* features. Since this depends on a given subject, semantic frames were reclassified that way for every selected entity. Each semantic frame were labelled as *response* when the entity matches the given subject and its role is *agent*, or as *stimulus* otherwise.
- **Binarization**: numerical values for psycholexical features associated to the words in the corresponding lexicons were mapped to 0 or 1 using the average of each feature as a cutting threshold.
- **Time resolution**: time in the sequence of semantic frames was represented as the position within the document of the root token of the sub-sentence.
- **Time alignment**: the time-steps were used for ordering the sequence and for aligning the different sequences from a same book. At a given time-step, the symbol depends on the entity. When the entity was the subject, the symbol for the corresponding feature category (*stimulus* or *response*) and feature dimension (e.g. *valence*, *joy*) was set to the feature's binarised value, i.e. 0 or 1 ; otherwise, when the predicate occurring at this time-step was not related to the subject, the symbol was set to 0, representing the absence of expression of the feature dimensions.

The resulting strings represent the occurrences over time of one of the psycholexical features (*valence*, *arousal* and *dominance* for VAD ; *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust* for Emotion), associated either to a given character of a book, or to the environment of this specific character.

3.4 Block entropy analysis

From the previously described character-emotion strings, block entropy (Section 2.5.2) was estimated to measure the extent to which they produce recurring or novel patterns over time.

Note: we mention the notion of time or temporal properties for the sake of conciseness, to refer to relative positions along the sequences, that is, along the book.

We were interested in the two dimensions of (univariate) information structure: entropy rate h (2.7–2.9) and excess entropy η (2.11). The former represents the average rate of information production, that is, novel patterns in the sequence ; while the latter represents complexity in the sense of pattern recurrence or information storage. See Section 2.5.2 for definitions.
Since these are asymptotic properties of block entropy $H(k)$ (2.6), we looked at the convergence of their finite-sample estimates (2.7, 2.10).
For this to happen however, sample sequences must be long enough.

3.4.1 Block entropy estimation

Estimating block entropy can also be difficult in practice, as it is defined for infinite sequences. Here, every sample represents the occurrences of a single feature (e.g. *fear*, *joy*) for a single character throughout a given book. The probability distribution P_k of blocks of length k was estimated as the sample frequencies $\hat{P}_k$:

$$\hat{P}_k = \left\{ \frac{n(x_i^{(k)})}{c_k} \right\}_{i=1}^{c_k} \qquad \hat{H}(k) = H(\hat{P}_k) \tag{3.1}$$

Where $n(x_i^{(k)})$ is the number of occurrences of a given block $x_i^{(k)}$ in the book and c_k is the number of distinct blocks of length k occurring. Blocks that never occur do not contribute to block entropy.

Entropy rate can be formulated asymptotically as either the difference (2.8) or the time-average (2.9) of block entropy. The difference formulation was chosen in order to preserve aggregate statistics over a given k:

$$\hat{h}(k) = \frac{\hat{H}(k) - \hat{H}(k - \Delta k)}{\Delta k} \qquad \hat{\eta}(k) = \hat{H}(k) - k\hat{h}(k) \tag{3.2}$$

3.4.2 Finite sample effects

Since samples are finite, they may be too short to capture the whole range of possible sub-sequences, a situation we refer to as *undersampling*. In such situation, there is a critical length k^* above which each blocks only occur once:

$$P_k^* = \left\{ \frac{1}{c_k} \right\}_{i=1}^{c_k} \qquad \hat{H}^*(k) = \log c_k \tag{3.3}$$

Then block entropy $\hat{H}^*(k)$ only depends on the number c_k of distinct blocks, which either remain constant or decreases. This violates the property of monotonically increasing block entropy, used to calculate entropy rate and excess. Therefore, the sample estimate of such asymptotic properties are only reliable if they converge within some length $k < k^*$.

3.5 Transfer entropy analysis

After investigating the information structure of single features, we looked for possible interactions between them in the form of information transfer. In particular, we were interested in interactions between *environment* (*stimuli* features) and *subject* (*response* features).
The strength of these interactions represents the extent to which the environment-subject feedback loop described in Section 2.2.2 is measurably represented in the corpus, and whether this is a relevant model for extracting meaningful psychological information from a corpus of novels.

Transfer entropy $T_{Y\to X}^{(k)}$ (2.13) was used to measure how much information from the current state of X comes from the past states of Y. As for block entropy, it is defined for some finite history length k, and l respectively for Y which here will be $l = k$.
The additional condition on Y significantly increases the cost of its computation in terms of computational time and required sample size. Thus, empirical measurements from small finite samples are sensitive to noise, and can then give positive values even for variables with no direct relationship [71]. For this reason a statistical significance test was also performed over these estimates, as described in Section 2.5.2.

We used the implementations from the JIDT toolkit [81], which provides estimators for transfer entropy and the surrogate distribution, the latter being inferred either analytically or using permutation resampling. Transfer entropy was estimated by applying its definition (2.13) on the frequency of configurations directly, as was done for block entropy in Section 3.4, since this is the standard method for discrete variables [81].

3.6 Comparison with baselines

In order to ensure that the data processing and feature extraction does capture meaningful information structure, the analysis was repeated with baselines representing a lower level of additional context added to the raw text. As discussed in Section 2.3, this was used to determine what additional features (or "context") were needed to capture informational patterns.

The first baseline was *no semantic role*, consisting in the full application of the method described in Section 3.1, but without the use of semantic roles (*agent*, *patient*) for classifying predicates into *environment* or *subject*. It is equivalent to simply using the co-occurrence of an entity and a predicate within a sub-sentence as the criterion to decide which entity the predicate relates to.

The second baseline, *randomised subjects*, replaces the entity in every frame (regardless of the semantic role) by a randomly selected one from the same document with equal probability for every unique entity (including the one to be replaced). This breaks the *subject*/*environment* classification, as well as any subject-specific temporal correlations at the level of semantic features. Therefore, any information structure captured by the method that is also present using this baseline is a syntactic-level feature of language, rather than of any higher-level semantic meaning.

4

Results and discussion

Sections 4.1–4.4 present and discuss the results. Then Section 4.5 provides a more in-depth discussion of the limitations of the results and the implications for future work.

4.1 Data extraction

4.1.1 Semantic role labelling

Looking at the extracted records revealed that the method did not cope well with the complex language of novels. Below are the most common problems identified:

1. Missing entity (agent or patient or both)
2. Entity unresolved, e.g. “myself”, “her”
3. Irrelevant predicate (mostly matched by VAD), e.g. “is”, “have”, “nutmeg”, “soap”, “tall”
4. Misleading entity, e.g. “his arms”, “Mr. Wopsle’s great-aunt”
5. Parsing issues, e.g. “her left hand” parsed as predicate “hand” with agent “her left”
6. Wrong entity

In this qualitative evaluation of the samples, we saw that in most cases, the sentences do not map well onto a meaningful sequence of predicates about entities. Usually, either the predicate or the entity is directly relevant to a book character, but not both. An accurate estimation of how many samples are concerned would require an annotated dataset which we did not have, but the effects listed here clearly occur frequently enough to conclude that this will have a significant impact on the subsequent analysis.

These issues revealed our underestimation of the complexity of Semantic Role Labelling. Even though we only focused on two semantic roles (*agent* and *patient*) and two entity categories (*entity* and *environment*), we had to include numerous features from pre-trained supervised models (DP, NER, POS, hypernyms, coreference resolution, entity merging, lemma) as well as sophisticated heuristics (two-pass sentence and sub-sentence parsing, rule-based entity classification, manually coded exceptions). Given the unconvincing results in term of data quality, this undermines the initial argument that this approach would require less contextual knowledge and be easier to interpret compared to training a custom state-of-the-art attentional model as depicted in Section 2.1.5.

As a consequence of the significant amount of noise and inaccuracies in the resulting dataset, the results discussed in the following analyses may not properly reflect information from higher levels of semantic meaning, which the method was designed to capture. This also holds for the baselines methods.

4.1.2 Characteristic time scale

In order to interpret the number of words represented by a given history length k, we attempt to estimate the expected number of tokens between every predicate.
An unintended consequence of splitting the text into sub-sentences, as described in Section 3.2.4, is that it breaks the order in which predicates occur. Indeed, when following the dependency tree of a composed sentence, tokens may occur in a different order than in the original text. As a result it produces artefacts on the distribution of token positions, shown in Figure 4.1.

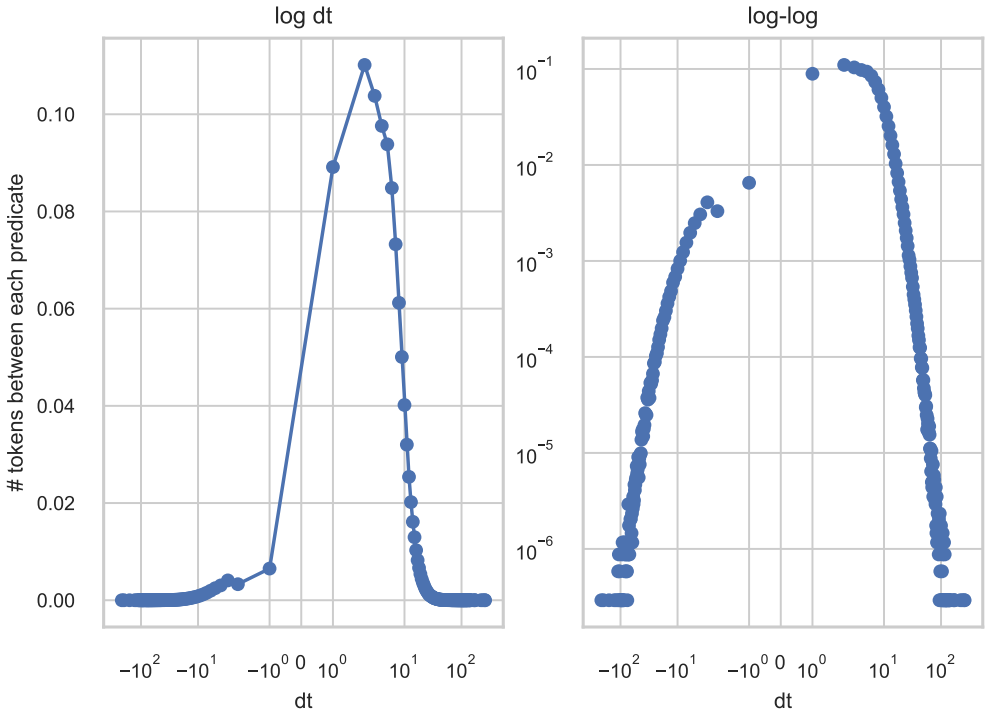

Figure 4.1: Distribution of the number of tokens between every predicates (log distribution left, log-log scale right), written dt as it is the difference in postition t within the document. The disruption of the order of tokens leads to 3.11% of samples with $dt < 0$. The overall distribution looks similar to a log-normal distribution, but with a discontinuity at $dt \in [-1, 1]$ and a left skew (visible in the tails on the left plot). On a log-log scale, we also notice an asymmetry due to negative values having lower frequency, suggesting a qualitatively different distribution for negative values.

These artefacts break any attempt at fitting a statistical model on the whole distribution naively. The best approximation was obtained by fitting a log-normal model[1] on the positive values (96.889% of samples), giving an overestimated mean of $\langle dt \rangle \approx 7.049$ (Figure 4.2). This model is still not a good fit, but serves as a very rough approximation.

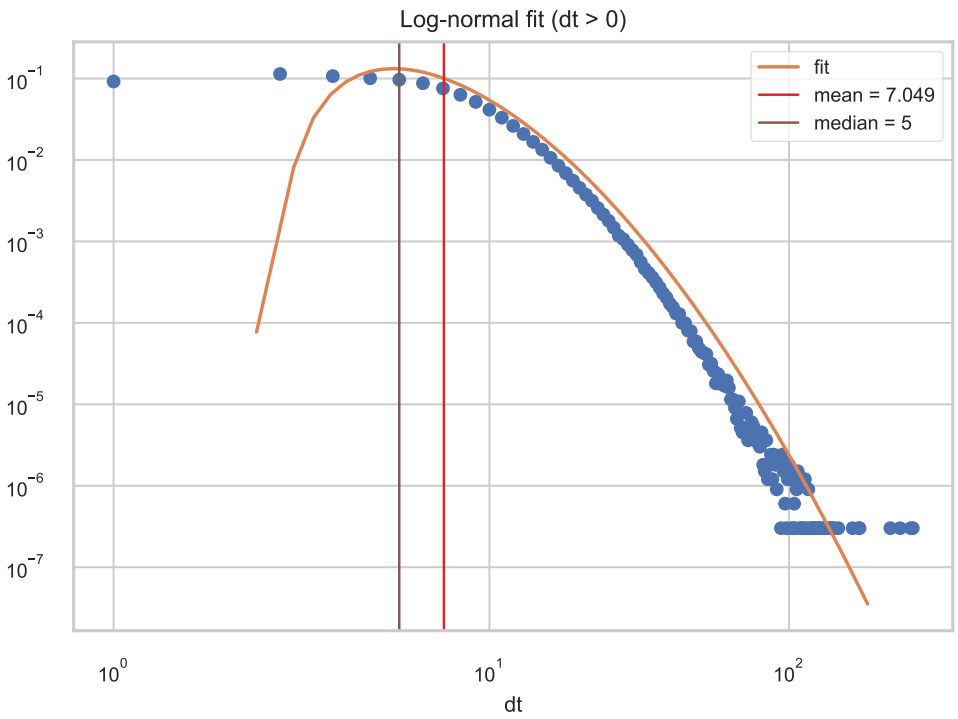

Figure 4.2: Maximum-likelihood estimation of a log-normal distribution for the positive values from Figure 4.1 (which represents 96.889% of samples). The model fits the data poorly (notice the log scale), especially on the head of the distribution where the discontinuity occurs.

[1] Maximum-likelihood estimator provided by the `SciPy` package, see `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.lognorm.html`.

4.2 Corpus

The distribution of characters identified (subjects) across books is shown on Figure 4.3. The number of subjects per book ranged from 1 (*The Frontiersmen*) to 26 (*Camilla*, *Vanity Fair*).

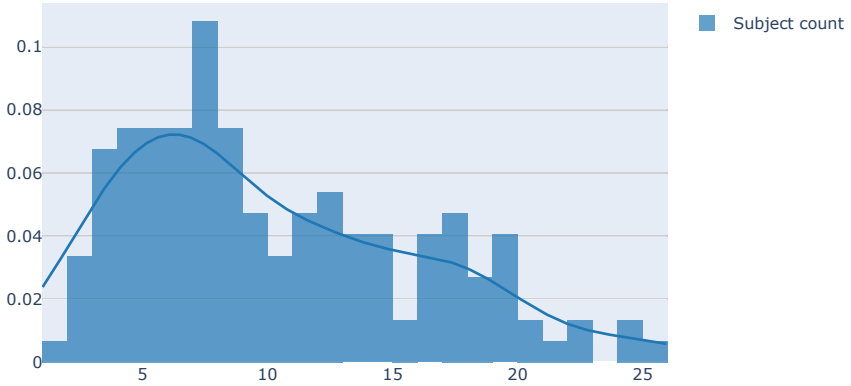

Figure 4.3: The number of distinct subjects matched per book.

The sequences extracted from the books, on which entropies have been estimated, have lengths ranging from 4213 to 64,536 symbols, with distribution shown on Figure 4.4.

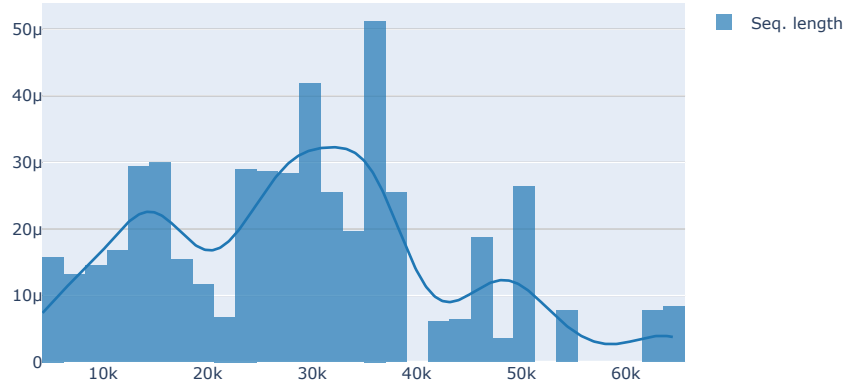

Figure 4.4: The length of symbol sequences extracted from the books, that is, the number of token matched, per subject.

4.3 Block entropy

4.3.1 Initial results

Because aggregate statistics do not properly reflect individual block entropy curves (illustrated in Figure 2.4), we focus on estimates of asymptotical properties, as defined in Equation (3.2). We look for convergence to the "true" estimate.
Measurements for entropy rate $\hat{h}(k)$ are shown on Figure 4.5. We clearly see here that entropy rate is converging to 0 as k increases, in all cases. The only exception is with `NoSemRole` where negative entropy rate is obtained at $k > 40$. Entropy rate being zero or negative is evidence for undersampling (see 3.4.2). Excess entropy is not shown, since in this case it simply converges by definition to block entropy (which remain constant after convergence).

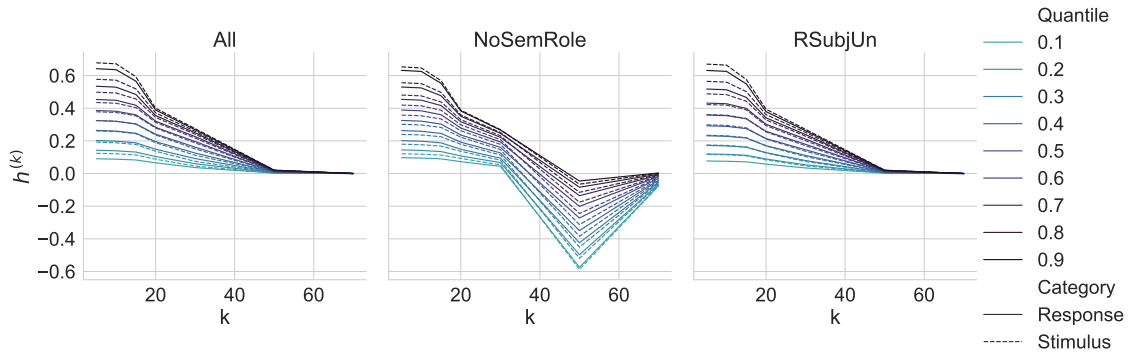

Figure 4.5: Entropy rate $\hat{h}(k)$ estimates distribution for increasing values of k, compared by method (rows) and feature category (*stimulus* and *response*). All entropy rate estimates converge towards 0, for all methods and feature category. This behaviour is consistent across the whole distribution in all of the plots.

There is no sign here for an intermediate convergent value before entering the undersampled regime. This implies that $k^* = 0$, and therefore, no conclusion can be made about these results because all data points unreliable.
Moreover, as evidenced by the lack of qualitative difference from the baselines, the method does not seem to capture any meaningful pattern at the semantic level. This is further confirmed by the absence of any distinct trend for *response* with regards to *stimulus*.

4.3.2 Alternative processing scheme

We investigated a possible remedy to the data processing which would allow to properly capture temporal correlations. It sacrifices time-scale consistency between sequences, which is only required for the computation of transfer entropy. This alternative consists in the following modifications:

1. One major issue is that 0s in the sequence can encode either of two different things: a low feature weight discretised as 0 ; or the matching of a word associated with another entity, set to 0 for alignment (see Section 3.3). The latter leads to long sequences of meaningless 0s. In this alternative, those are simply removed from the sequence, leading to sequence of any size and not aligned anymore.
2. Another issue is the encoding of the feature vector spaces (8-dimensional for the *Emotion* lexicon, and 3-dimensional for *VAD*) as separate binary strings for each dimension, rather than encoding the whole space with an alphabet appropriate to the spatial distribution. In this alternative, we encode all combinations of discretised features occurring in the lexicon as distinct symbols. The alphabet then has up to 2^8 or 2^3 symbols for *Emotion* and *VAD* respectively.

We look at the longest sample as the best-case in undersampling behaviour. Note that with this new processing samples are considerably shorter: the longest sample has a length of 6000, instead of 64,536 previously.
Convergence is faster but entropy rate (Figure 4.7) still goes to zero. To ensure this is due to undersampling, we compare the measured block entropy $H(k)$ to that in the undersampled regime $H^*(k)$ (Figure 4.6). We see that with *VAD* we reach a peak around $k = 3$, where we we have $\hat{h}(3) = 1.996894$ bits, before the effects of undersampling start becoming visible. But such a short length is irrelevant as an estimate of the asymptotic entropy rate. For *Emotion*, undersampling is visible from the start. Indeed, since it uses more symbols, it requires longer sequence lengths for reliable estimations.

These additional results confirm that undersampling occurs way before we see any convergence of the entropy rate estimates to the asymptotic value. Therefore, we cannot investigate the information structure from these measurements.

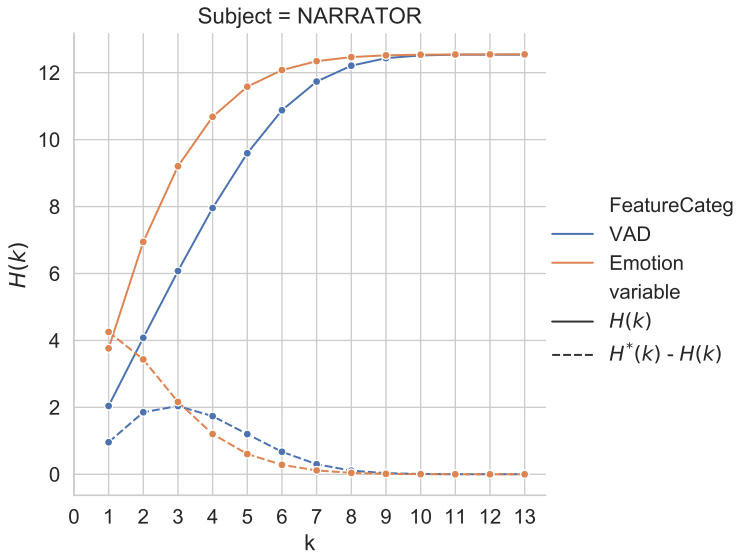

Figure 4.6: Block entropy with alternative processing for the longest sample (subject `NARRATOR` from book *Lorna Doone* by *R. D. Blackmore*, with length 6000). The same behaviour as in 4.5 is observed, that is, block entropy converges to a constant and therefore the entropy rate is 0. But the convergence now occurs earlier ($k \approx 12$ instead of $k \approx 50$). Dashed line represent the difference from the undersampled baseline: it increases with "true" block entropy and decreases with the effects of undersampling. This quantity is equal to the Kullback-Leibner divergence of block entropy from the undersampled regime.

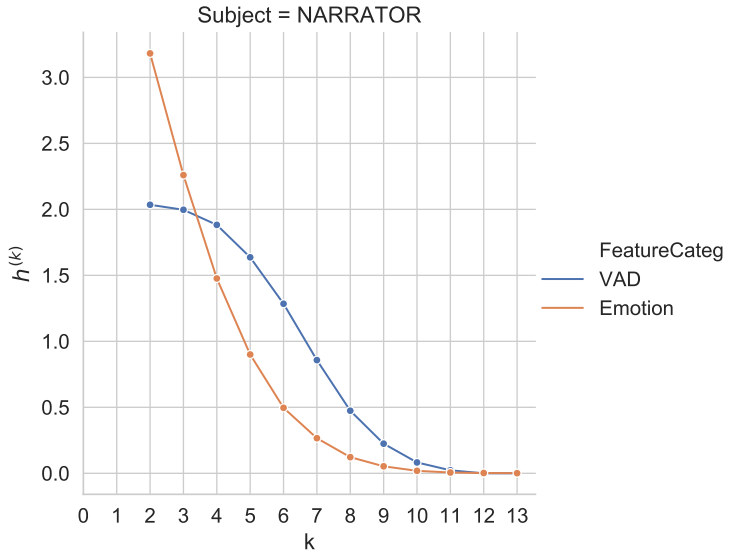

Figure 4.7: Entropy rate with alternative processing for the longest sample (as in Fig. 4.6).

4.4 Transfer entropy

Transfer entropy being more combinatorially expensive to measure than block entropy, the limitations due to the relatively small size of books and computational resource constraints are even more significant in this analysis. We saw already that the symbolic representation used here did not afford the reliable estimation of asymptotic entropy rates due to undersampling. Therefore, any apparent information structure resulting from these measurements likely reflects noisy artefacts, and this is even more the case for transfer entropy. We show some of the results anyways as baseline for future work.

The goal here was to identify significant correlations between feature categories (*stimulus* and *response*). First, we looked at the range and distribution of estimated transfer entropy values $T^{(k)}$ depending on history length k and feature categories. Based on the test of significance described in Section 3.5, we excluded here all samples with $p > 0.05$, and also those with $p = 0$ as they seem suspect.
The results shown on Figure 4.8 suggest that transfer entropy estimates are very low, close to 0 for most samples. Differences are noticed mostly from $k = 12$, with *stimulus* having higher extreme values. We also noted slightly higher values between *stimulus* and unclassified predicates (*other*). This is consistent with our expectation since *stimulus* is the aggregate of all entities but the subject, and hence should contain more information and complexity. Temporal correlations within same-feature pairs are less interesting since they likely reflect the same correlations already studied in the block entropy analysis.
For other feature pairs, we could not identify any general trend or convergence, due to the high variance between samples. Extrapolating from Figure 4.8 would be unrigorous.

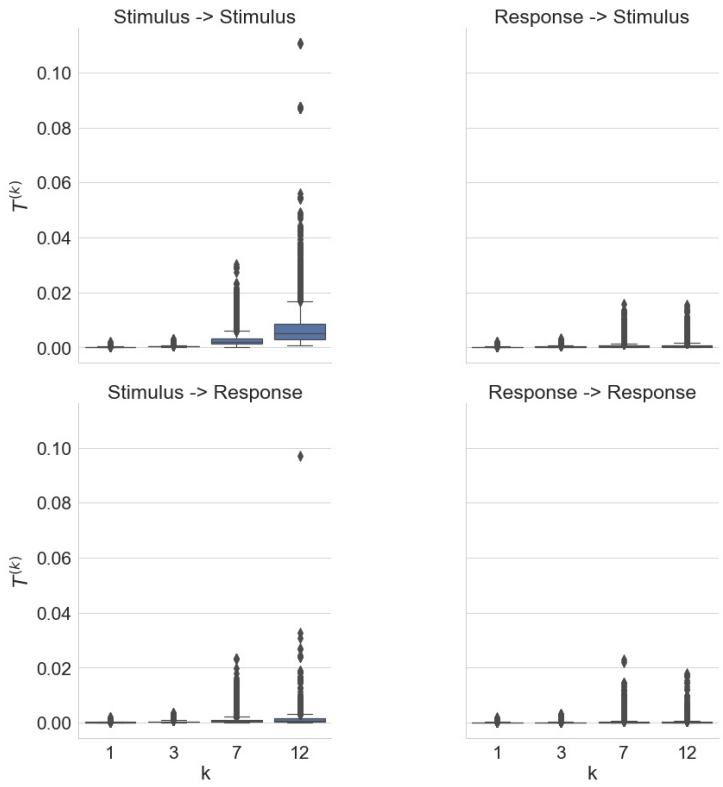

Figure 4.8: Box-plot of all transfer entropy estimates with $0 < p \leq 0.05$, per feature pair and history lengths k. This graph illustrates the distribution of the transfer entropy estimates. It is not meant to be compared to Figure 2.4. Features classified as *other* (unclassified predicates / semantic roles) follow a similar distribution (not shown here), with estimates reaching $T^{(k)} \approx 0.02$ and a higher value with *stimulus* compared to *response*.

Given the inconclusiveness of this limited analysis, we investigate how the significance of the estimates, rather than the values, is affected by the features and methodology used.

For p-value calculation, only the surrogate based on permutation resampling was used, because the analytical surrogate distribution for discrete variables is defined asymptotically for a large enough number of samples, and convergence is much slower with a skewed distribution [81], which is the case here.

Figure 4.9 shows the p-value distribution for various history lengths k. The cumulative distribution is used to highlight the differences at the edges of $p = 0$ and $p = 1$. Estimates for $k < 7$ appear overconfident (many $p = 0$), and undersampling is obvious at $k \geq 12$. $k = 7$ gives a more balanced distribution, but as discussed previously, at such history length the probability distribution of blocks has not yet stabilised, even without the additional condition needed for transfer entropy.
Hence, these p-values are no evidence for any significant transfer entropy in the asymptotic limit, only for correlations at length $k \leq 7$ which may disappear when considering longer ranges. This also means that the apparently higher estimates at $k = 12$ shown in Figure 4.8 are unreliable.

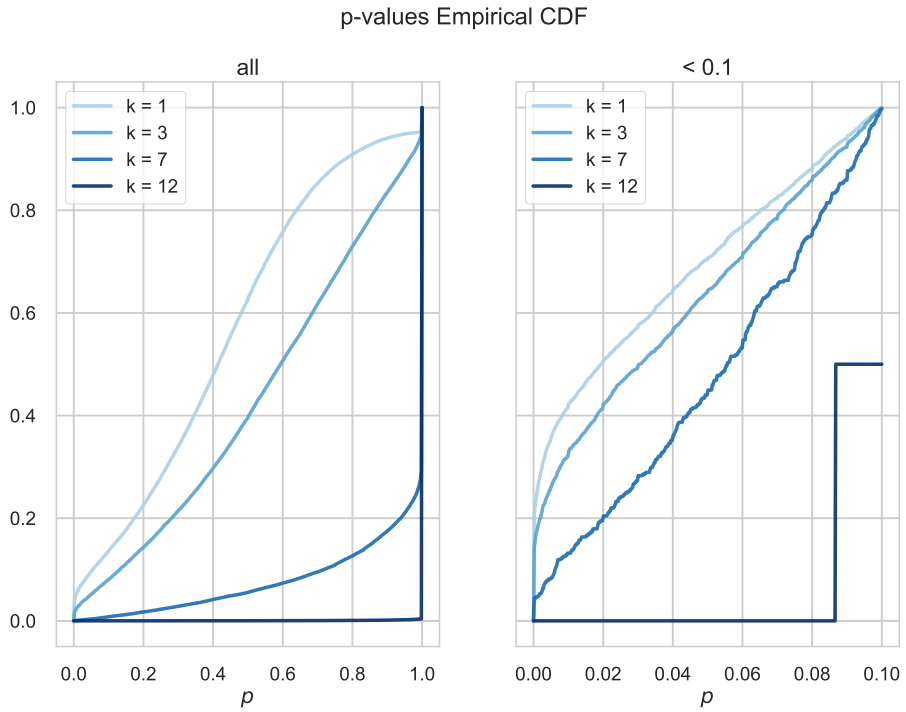

Figure 4.9: Cumulative distribution of p-values resulting from significance test for transfer entropy between all pairs of feature sequences. Left: whole empirical cumulative distribution ; right: only the interval $0 \leq p < 0.1$. Notice the impossibly high frequency of $p = 0$ and the suspicious smoothness of the curve for $k < 7$. Also note the absence of significance at $k = 12$. From there, only $k = 7$ seems like an appropriate length.

To compare the effect of features and methodologies on p-values, we measured the relative frequency of samples with $0 < p \leq 0.05$, as a measure of the probability of a relationship between a source feature and a target feature. Results are shown on Figure 4.10. Overall, the probabilities are quite low, and particularly so in the relations between *stimulus* and *response*, for all features and methods. This is hardly surprising given the problems of the data extraction methodology.

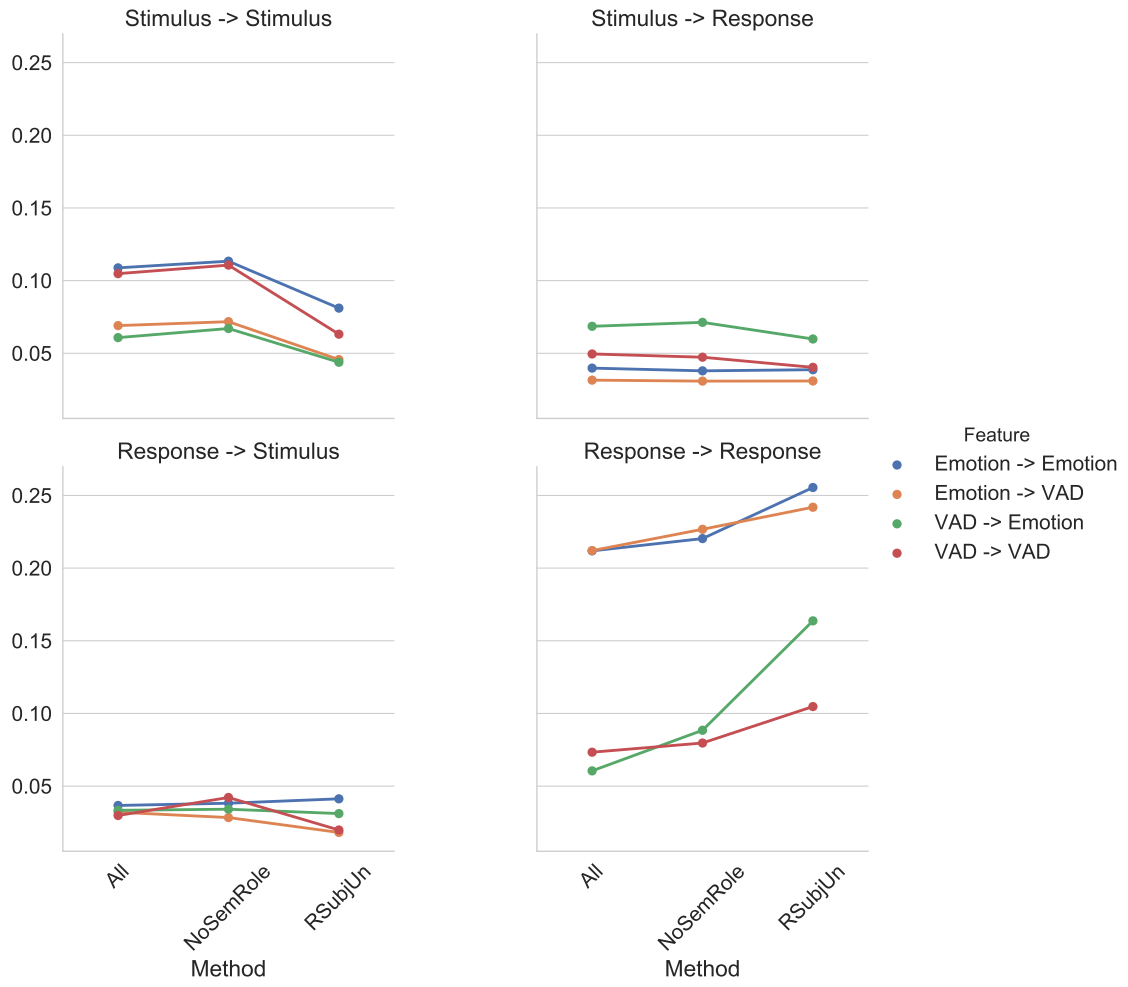

Figure 4.10: Proportion of samples with $0 \leq p < 0.05$ per method and feature categories (*stimulus* / *response*), at $k = 7$. Feature category *other* representing classification errors and ambiguities is not shown here, but doesn't exceed a proportion of 0.05 (with the exception of the pair *other* → *response* with *VAD* and *All*/*NoSemRole* which slightly exceeds 0.10).

We see a small but consistent difference (∼2–10%) when subjects are randomised, but only when the source and target features are of the same category. Because randomising subject also randomises the agent and subject of a predicate, it affects the *stimulus* / *response* classification and can change their relative distribution, which explains this apparent shift from *stimulus* samples to *response* samples.
A small effect is also visible with the removal of semantic roles, although it is tiny (less than 2%). While these observations may indicate small differences between

methods to some extent, the size of these effects is not significant enough to provide evidence that the processing done in this work produces any information relevant to personality.

The important takeaway here is that our doubts about the ability of the methodology used to capture higher-level meaning are only reinforced. One should be careful not to extrapolate from the very small effect sizes obtained in these results. At best, it could suggest that future work may use much simpler text processing methods, such as word-entity co-occurrence detection, but with a greater focus on the features used (psycholexical dimensions or other).

4.5 Limitations and improvements

4.5.1 Feature selection

In addition to the limitations and issues already mentioned, it must also be noted that the baselines used to validate whether the method extracts any semantic-level information do not address its ability to extract the lower-level information.

For instance, the benefit of parsing and iterating sub-sentences is not clear, given how it obscured the interpretation of time scales. It seemed necessary to avoid mixing together all entities and predicates within long sentences. But verifying this formally by implementing a baseline first would have allowed us to evaluate whether it was worth the additional complexity.
The same can be said for all lower-level features introduced here. Each feature, such as word lemma or POS and DP labels, requires a dedicated model, which adds complexity to the processing pipeline. As the results obtained made clear, such complexity may not be necessary and increases the difficulty of interpretation of the results and the likelihood of erroneous artefacts.

Hence, we stress the importance of performing the analysis on the simplest possible baseline first, e.g. raw text, and then add features progressively and evaluate how each affects the relevant properties being measured, e.g. entropy rate. This prevents premature optimisation and minimises the complexity of the processing steps.

4.5.2 Evaluation of custom NLP pipeline

The insufficient investigation of feature relevance would be less problematic if the semantic role labelling performance had been formally evaluated. This is another important shortcoming of the methodology used here, as evidenced by the amount of erroneous outputs and the lack of qualitative difference with the baselines.

We tried to get around the complexity of black-box semantic-level models by focusing on a small subset of SRL. While the methods used are typical of processing pipelines before the advent of deep attention-based language models [33], achieving successful extraction of semantic information is still far from straightforward even with recent models. Natural Language Understanding entails many open research problems, so developing tailor-made solutions requires more careful choices and evaluation than what was possible within the scope of this project.

For this reason, it is critical to leverage existing models whenever possible, regardless of their size. This suggestion complements the previous one in that it gives us a chance to identify the most useful features. Further optimisations for simplicity and performance can be added later. The task of labelling agent/patient roles or stimulus/response may be framed as a typical NLP classification task, allowing us to use transfer learning to adapt an available state-of-the-art model to this task (see Section 2.1.5). Decent performance may be achieved with relatively little annotations [30], and further improved by fine-tuning the language model on the corpus, or re-using the language model or even the classification layer of some state-of-the-art classifier for a similar task. Apart from SRL and other high-level NLU task, existing datasets may also be leveraged for emotion classification [45].

The advantage of deep neural network approaches is the vast literature available for improving the effectiveness of classification, training, representation, interpretation and generalisability. In particular, knowledge distillation [31, 82], semi-supervised learning and multilingual word embedding [10] are worth considering for reducing the need for corpus-specific annotations, and hence facilitating cross-corpus comparisons.

Regardless of the type of NLP models used, instead of using their primary output features for information extraction, we may also investigate information structures emerging from their internal features. This is especially relevant for unsupervised language models as it removes the need for annotated corpora.

4.5.3 Choice of corpus

As highlighted by the results presented in Section 4.1.1 and the challenges discussed in Section 3.2.4, a corpus of literary English does not make for an easy benchmark for a first attempt at the kind of data extraction we attempted here. The language can be particularly complex, and using models trained on contemporary English without fine-tuning is likely to produce more erroneous artefacts, in addition to the language bias issues mentioned in Section 1.2. Implementing a working pipeline first on contemporary texts may facilitate its formal evaluation, giving a solid baseline upon which to improve for a literary English application.

4.5.4 Potential of probabilistic modelling

As mentioned in Section 2.1.5, the possible solutions are not limited to deep neural networks, but also include various kinds of generative probabilistic models, e.g. PGM. There is a broad literature of applications of such models to tasks relevant to computational literary analysis [38, 83, 36, 37].

Considering the initial motivations formulated in Section 1.2, PGM would seem an appropriate approach as it affords an explicit representation of the expected invariant structure of personality models. The undersampling issues we faced in the estimation of entropies could also be addressed by a careful choice of prior, such as the distribution all words and predicates in the text. The possibility to add explanatory variables from manual annotations, as done with supervised models, also provides a welcomed flexibility.

4.5.5 Numerical entropy estimation

Numerical estimation of entropy is a complex topic subject to active research [71]. There is no one-size-fits-all solution. Direct estimation by applying Equation (2.2) on probabilities from frequency measurements, as was done here, generally overestimates the true entropy [71]. The entropy rate estimator defined in Equation (3.2) is also known to consistently overestimate the true value [84].

The performance of entropy rate estimators depends not only on sequence length, history (block) length and alphabet size, but also on correlation lengths and therefore on entropy rate itself. It has been suggested to adapt the estimation approach based on a first rough estimate of entropy rate [84]. Other estimators include Lempel-Ziv complexity [84] and Bayesian methods, incorporating priors about either the probabilities or entropy itself [71].

Transfer entropy estimators are usually based on existing techniques for mutual information. It is an open research problem and it is recommended to use available software packages [71].

4.5.6 Temporal correlations under critical length

The bound k^* on history length can be found analytically, as described in [84], or by performing the analysis on artificially-generated data from a k-th order Markov chain and looking at the variation of information structure for increasing k.

In situations where longer range temporal correlations are inaccessible or cannot reliably be estimated, as was the case here, the asymptotic information-theoretic cannot be estimated either ; however, we may still investigate the contributions of shorter correlation lengths to the total information structure. This can be done by decomposing the total correlation information into contributions from all lengths [70].

4.5.7 Complete transfer entropy

Transfer entropy as used in the project is referred in the literature as the *apparent transfer entropy* $T_{X_j \to X_i}$ between two elements of a multivariate system X. It does not account for the possible influence of other elements.

A more accurate measure is the *complete transfer entropy* $T_{X_j \to X_i | X_{[ij]}}$, which measures the influence of X_j on X_i conditioned on every other variable $X_k, k \neq i, j$. We may also consider the *collective transfer entropy* $T_{\mathbf{Y} \to X_i}$ of the joint process $\mathbf{Y} = (X_1, ..., X_n) \subset X$ [71] on a single element X_i.

The downside of these more comprehensive measures of information transfer is the increased data requirements due to the additional conditions. They were not considered here since the sequence lengths were already too limited.

4.5.8 Sequence processing

With the alternative processing described in Section 4.3.2 we attempted to reduce correlation lengths by removing unnecessary 0s and recover information from the full feature space by merging the feature dimensions together. But since this increases the alphabet size, it also substantially increases the minimum sequence length for reliable estimation. One may want to try removing unnecessary 0s without merging the features in order to obtain the most reliable estimates possible, at least for individual features.

The binning of the original real-valued features into binary values is another critical part of good entropy estimation from continuous measurements. More sophisticated methods have been used in the literature, such as nearest-neighbour algorithms [71]. In order to select the appropriate binning scheme, a closer look at the distribution of feature vectors within the feature space would be helpful.

5
Conclusion

In this project we explored a possible application of information theory with literature data for a semantic-level analysis of the interactions between book characters.

We attempted to extract information from a corpus of English novels such as to obtain a symbolic representation of the sequence of emotions associated with the entities. We then looked for information structure related to these entities and their interactions, by measuring entropy rates and transfer entropies, seeking evidence for complex temporal correlations as well as psychologically-relevant interactions between certain categories of entities.

We described and implemented a methodology that would, in principle, avoid important sources of bias found in previous work at the intersection of psychology and natural language processing. But the complexity of the processing needed to make this approach viable was underestimated. As a result, the data quality obtained was insufficient for our analysis, likely due to a large number of erroneous artefacts from the processing.

We also came across challenges related to the estimation of entropy rate and transfer entropy from short symbol sequences with potentially long-range temporal correlations. Consequently, we were unable to draw any conclusions about the asymptotic information properties.

The implication for future work in this direction is that considerations about the various sources of biases in NLP and psycholexical models should be put aside for now, in order to investigate more closely which semantic-level features (1) allow the extraction of enough informative data samples and (2) capture psychologically relevant information structure. The focus should be on ensuring that appropriate statistical estimators can be applied, and on leveraging state-of-the-art language models, using Semantic Role Labelling or event/frame detection and coreference resolution only if necessary.

Glossary

affect The unconscious experience underlying feelings and emotions[3]. xi, 1, 3, 8, 12–14, 24

attention Mechanism used in Machine Learning for adaptive weighting, inspired by the homonymous cognitive mechanism and representing a kind of temporal and/or spatial memory. 5, 9, 31, 42

baseline Reference for evaluating the performance of a model, or in the case of this work, for measuring the effect of a certain feature or parameter on the outcome of an experiment. We may refer here to *the baselines* for the baseline processing methods used for this comparison. 21, 22, 30, 32, 35, 42

behaviour In psychology, the range of actions and mannerisms made by an individual. Not to be confused with the notion of behaviour used in complex systems theory, where it usually refers instead to qualitative properties of a system observed under certain conditions[4]. 1, 3, 11–13, 16, 24

BERT Bidirectional Encoder Representations from Transformers[10]: a deep neural network language model based on the Transformer architecture, with multiple attention layers encoding syntactic- and semantic-level features. Considered the first truly successful application of Transfer Learning in NLP. 6, 9

black-box In Machine Learning, a black-box model is a model for which we only know what the inputs and outputs represent. The model produces outputs based on a composition of abstract functions parametrised from the training data. Thus, understanding and interpreting what the model does is often difficult. 10, 14, 21, 42

cognition In psychology, mental information processing mechanisms such as thought, attention or memory. 1, 3, 12, 13, 24

coreference In Natural Language Processing, coreference resolution refers to the task of associating different mentions of the same entities in a text (e.g. "John Doe", "John", "M. Doe"). I, III, 8, 16, 22, 24–26, 31, 47

correlation length the length in time of a temporal correlation. 44, 45

distillation Machine Learning technique consisting of training a smaller model to reproduce the output of a larger one. 9, 10, 43

DP Dependency Parsing. xiii, I, IV, 7, 21, 22, 25, 26, 31, 42

dynamic Dynamical systems are systems which change over time according to some rules. Here the term dynamical loosely refers to systems, models or observables of such nature. We also refer to Information Dynamics which is the area of Information Theory concerned with these systems. v, 3, 4, 11, 16–18, 21, 27

embedding In Natural Language Processing, a word-vector embedding is a language model representing word meaning in an abstract vector space. 3, 5, 6, 15, 43

emotion The experience of affect as understood and expressed through a certain individual and cultural context. v, xi, 1, 3, 8, 11, 13, 14, 16, 21, 22, 24, 28, 43, 47

entity In Natural Language Processing, an entity is a person, group of persons, location, brand or anything that can have a proper name. II, III, 4, 7, 16, 21, 22, 25–27, 30, 31, 42, 47

environment Refers here to everything that surrounds an individual, used interchangeably with situation. 3, 4, 11–14, 16, 21, 22, 24, 26–31

FA Factor Analysis. 10, 15

feedback loop When the output of a system is *fed back* to itself as an input, it becomes a dynamical system: it produces some behaviour over time. This also occurs more generally when any two parts of a system are mutually affecting each other. v, 11–13, 29

FFM Five Factor Model. 2, 10–12, 14, 15

fine-tuning Additional training of a pre-trained model in order to increase its performance in a specific domain, or after an alteration such as Transfer Learning. 43

frame In Natural Language Processing, a semantic frame is a task-specific record representing semantic knowledge extracted from a certain position in the text. xiii, II, 8, 10, 22, 23, 26, 27, 30, 47

generative Generative models represent a data distribution, and can be sampled to produce new data, or used as a representation layer for other inference models. PGM are an example. 9, 43

hypernym In the WordNet database[23], represents the relation between a more general meaning (synset), e.g. *animal*, and another meaning which is a more specific instance of the former, e.g. *dog*. III, 22, 31

language model Abstract representation of semantic meaning based on distributional semantics i.e patterns of word co-occurrence. 2, 3, 6, 8–10, 42, 43, 47

lemma Base form of a word, e.g. *look* (*looking*, *looked*, ...), *produc* (*production*, *product*, ...). 22, 25, 31, 42

LIWC Linguistic Inquiry Word Count. 14

LSTM Long Short-Term Memory. 5

memoryless A memoryless process or system is one for which the current state predicts the next state, regardless of past states, and can be represented as a first-order Markov Chain. A kth-order Markov Chain with finite k represents a process which is memoryless beyond the past k states. 16

ML Machine Learning. 5, 8, 10

NER Name Entity Recognition. 7, 22, 25, 26, 31

NLP Natural Language Processing. v, 2, 3, 5–10, 15, 16, 21, 24, 26, 43, 47
NLU Natural Language Understanding. 5, 9, 10, 42, 43

PGM Probabilistic Graphical Model. 9, 10, 43
pipeline Sequence of single-purpose functions, each fed the outputs of the previous ones as input. In Natural Language Processing, these typically consist in annotating the text with features, sometimes based on previously added labels. 15, 16, 21, 42, 43
POS Part-of-Speech. xi, III, 5, 7–9, 21, 22, 26, 31, 42
predicate Here refers to a word in the text which expresses some information about an entity. xi, I, II, 8, 21–23, 25, 27, 30–32, 38, 42, 43
prior In Bayesian statistics, the prior distribution represents our current knowledge before an observation. 15, 43, 44
psycholexical An approach for assigning psychology-related meanings to words, such as emotions or personality traits, usually by the manual (or survey-based) selection and annotation of a lexicon, followed by some form of dimensionality reduction. II, 2, 3, 8, 10, 14, 22–24, 27, 28, 42, 47

response In Neuroscience, a response is a neurological signal produced by the brain as a reaction to a stimulus. Here used loosely as an analogy for any observable characteristic of an individual's mental state (affect, behaviour or cognition), implying a reaction to a situation. 11, 27, 29, 35, 38, 39, 41, 43
RST Reinforcement Sensitivity Theory. 11, 12

semantic role In Natural Language Processing, the role of a token (or token span) in a sentence, e.g. active subject (agent), passive subject (patient). xi, I, II, 9, 22, 25, 27, 30, 39
semi-supervised Semi-supervised Learning refers to the use of both a supervised model, often trained on a restricted dataset, and an unsupervised representation obtained from a much larger unlabelled dataset, to improve and generalise the performance of the supervised model on the larger dataset. 9, 43
sentiment Conscious emotional predisposition about something or someone, e.g. trust in government, fear of technology. xi, 5, 8–10, 13
situation In Personality Theory, a situation represents a feature of the surrounding environment which may affect the manifestations of an individual's personality trait. For example, a social gathering or a relationship conflict. 1, 3, 11, 13, 14, 24
SRL Semantic Role Labelling. 8, 10, 16, 24, 26, 31, 42, 43, 47
stimulus In Neuroscience, a stimulus is a physiological signal triggering a reaction (response) in the brain. Here used loosely as an analogy for any information perceivable by an individual about his/her environment. 11, 13, 27, 29, 35, 38, 39, 41, 43
sub-sentence Long sentences, common in the corpus used here, may be composed of multiple sub-sentences linked by commas or specific words (e.g. *and*, *then*). xiii, I, II, 26, 27, 42
supervised Supervised Machine Learning consists in fitting a model to reproduce given outputs (or labels) associated to a training set of input samples. 8, 9, 14, 15, 21, 31, 43

surrogate In statistical significance testing, the surrogate distribution is the distribution under the null hypothesis. 19, 40

temporal correlation In Information theory, a correlation between the current state of a random process and its state at a given time delay in the past. By temporal correlations, we usually refer here to the entire information structure in the temporal dimension, which emerges from such correlations along the entire spectrum of time delays. 3, 6, 30, 36, 38, 44, 47

token Basic unit of information in Natural Language Processing, representing words (or word parts), punctuation or task-specific control values. xi, I, 7, 9, 22, 25, 27, 32, 34

transfer learning Reusing part of a trained model as a feature extraction layer for a new model, significantly speeding up its training. 5, 9, 10, 43

unsupervised Unsupervised Learning refers to applications of Machine Learning where data labelling is not necessary for fitting the model, e.g. representation learning, generative models, cluster analysis, word-vector embedding. 8–10, 15, 43

VAD Valence Arousal Dominance. 22–24, 28, 31, 36, 41

WordNet Lexical database of English word meanings, represented as synsets (unique identifiers of word meaning) and organised in a hierarchy of hypernyms (synsets representing broader categories)[23]. III, 14, 16, 22, 26

Bibliography

[1] Rachid Laajaj, Karen Macours, Daniel Alejandro Pinzon Hernandez, Omar Arias, Samuel D. Gosling, Jeff Potter, Marta Rubio-Codina, and Renos Vakis. Challenges to capture the big five personality traits in non-WEIRD populations. *Science Advances*, 5(7), July 2019.

[2] John F. Rauthmann and Ryne A. Sherman. The description of situations: Towards replicable domains of psychological situation characteristics. *Journal of Personality and Social Psychology*, 114(3):482–488, March 2018.

[3] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing*, 5(2):101–111, April 2014.

[4] Wikipedia. Behavior, November 2019. Page Version ID: 924359888.

[5] Philip J. Corr, editor. *The Reinforcement Sensitivity Theory of Personality*. Cambridge University Press, Cambridge, 2008.

[6] L. R. Goldberg. The structure of phenotypic personality traits. *The American Psychologist*, 48(1):26–34, January 1993.

[7] Oliver P. John, Alois Angleitner, and Fritz Ostendorf. The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, 2(3):171–203, September 1988.

[8] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, 2016. Association for Computational Linguistics.

[9] Geert Hofstede and Robert R. McCrae. Personality and Culture Revisited: Linking Traits and Dimensions of Culture. *Cross-Cultural Research*, 38(1):52–88, February 2004.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL 2019*, 2018.

[11] Telmo Pires, Eva Schlinger, and Dan Garrette. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, 2019.

[12] Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2):246–268, 1995.

[13] Stephen J. Read, Brian M. Monroe, Aaron L. Brownstein, Yu Yang, Gurveen Chopra, and Lynn C. Miller. A neural network model of the structure and dynamics of human personality. *Psychological Review*, 117(1):61–92, 2010.

[14] Robert R. McCrae, Thomas A. Martin, and Paul T. Costa. Age trends and age norms for the NEO Personality Inventory-3 in adolescents and adults. *Assessment*, 12(4):363–373, December 2005.

[15] Kirk W. Brown and D. S. Moskowitz. Dynamic Stability of Behavior: The Rhythms of Our Interpersonal Lives. *Journal of Personality*, 66(1):105–134, 1998.

[16] Sheri L. Johnson and Andrzej Nowak. Dynamical Patterns in Bipolar Depression. *Personality and Social Psychology Review*, 6(4):380–387, November 2002.

[17] Adele M. Hayes, Jean-Philippe Laurenceau, Greg Feldman, Jennifer L. Strauss, and LeeAnn Cardaciotto. Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical Psychology Review*, 27(6):715–723, July 2007.

[18] John F. Rauthmann, David Gallardo-Pujol, Esther M. Guillaume, Elysia Todd, Christopher S. Nave, Ryne A. Sherman, Matthias Ziegler, Ashley Bell Jones, and David C. Funder. The Situational Eight DIAMONDS: a taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, 107(4):677–718, October 2014.

[19] Maike Luhmann, Ulrich Orth, Jule Specht, Christian Kandler, and Richard E. Lucas. Studying Changes in Life Circumstances and Personality: It's About Time. *European Journal of Personality*, 28(3):256–266, 2014.

[20] Robin R. Vallacher, Stephen J. Read, and Andrzej Nowak. The Dynamical Perspective in Personality and Social Psychology. *Personality and Social Psychology Review*, 6(4):264–273, November 2002.

[21] Dallas Liddle. Could Fiction Have an Information History? Statistical Probability and the Rise of the Novel. *Journal of Cultural Analytics*, 2019.

[22] Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1), December 2016. arXiv: 1606.07772.

[23] George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November 1995.

[24] Zellig S. Harris. Distributional Structure. *WORD*, 10(2-3):146–162, August 1954.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, January 2013. arXiv: 1301.3781.

[26] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual String Embeddings for Sequence Labeling. *Association for Computational Linguistics*, Proceedings of the 27th International Conference on Computational Linguistics:1638–1649, 2018.

[27] Niels Dekker, Tobias Kuhn, and Marieke van Erp. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5:e189, April 2019.

[28] Saif M. Mohammad and Peter D. Turney. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. event-place: Los Angeles, California.

[29] Wikipedia. Supervised learning, April 2020. Page Version ID: 949790121.

[30] Sven Buechel, João Sedoc, H. Andrew Schwartz, and Lyle Ungar. Learning Neural Emotion Analysis from 100 Observations: The Surprising Effectiveness of Pre-Trained Word Representations. *arXiv:1810.10949 [cs]*, October 2018. arXiv: 1810.10949.

[31] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531 [cs, stat]*, March 2015. arXiv: 1503.02531.

[32] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, 2019.

[33] 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, 2019.

[34] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, 2018. Association for Computational Linguistics.

[35] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *33rd Conference on Neural Information Processing Systems*, page 30, 2019.

[36] Haoruo Peng, Snigdha Chaturvedi, and Dan Roth. A Joint Model for Semantic Sequences: Frames, Entities, Sentiments. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 173–183, Vancouver, Canada, 2017. Association for Computational Linguistics.

[37] Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. Probabilistic Frame Induction. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, February 2013. arXiv: 1302.4813.

[38] David Bamman and Noah A. Smith. Unsupervised Discovery of Biographical Structure from Text. *Transactions of the Association for Computational Linguistics*, 2:363–376, December 2014.

[39] Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. Word2Sense: Sparse Interpretable Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5692–5705, Florence, Italy, 2019. Association for Computational Linguistics.

[40] David Bamman, Ted Underwood, and Noah A. Smith. A Bayesian Mixed Effects Model of Literary Character. In *Proceedings of the 52nd Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 370–379, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[41] Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[42] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

[43] David Bamman, Sejal Popat, and Sheng Shen. An annotated dataset of literary entities. 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 2138–2144, 2019.

[44] Albin Zehe, Martin Becker, Fotis Jannidis, and Andreas Hotho. Towards Sentiment Analysis on German Literature. In Gabriele Kern-Isberner, Johannes Fürnkranz, and Matthias Thimm, editors, *KI 2017: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 387–394. Springer International Publishing, 2017.

[45] Evgeny Kim and Roman Klinger. Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, August 2018.

[46] G. W. Allport and H.S. Odbert. Trait-names: A psycho-lexical study. *Psychological Review Company*, 1936.

[47] Isabel Briggs Myers. *The Myers-Briggs Type Indicator: Manual (1962).* The Myers-Briggs Type Indicator: Manual (1962). Consulting Psychologists Press, Palo Alto, CA, US, 1962. Pages: ii, 110.

[48] Michael C. Ashton, Kibeom Lee, Marco Perugini, Piotr Szarota, Reinout E. de Vries, Lisa Di Blas, Kathleen Boies, and Boele De Raad. A Six-Factor Structure of Personality-Descriptive Adjectives: Solutions From Psycholexical Studies in Seven Languages. *Journal of Personality and Social Psychology*, 86(2):356–366, 2004. Place: US Publisher: American Psychological Association.

[49] Tena Vukasović and Denis Bratko. Heritability of personality: A meta-analysis of behavior genetic studies. *Psychological Bulletin*, 141(4):769–785, July 2015.

[50] Bertus F. Jeronimus, Harriette Riese, and Johan Ormel. Environmental influences on neuroticism in adulthood: A systematic review. preprint, Open Science Framework, September 2018.

[51] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1):70–87, July 2006.

[52] Joseph Ledoux. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life.* Simon and Schuster, March 1998. Google-Books-ID: 7EJN5I8sk2wC.

[53] Scott Parrigon, Sang Eun Woo, Louis Tay, and Tong Wang. CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology*, 112(4):642–681, 2017.

[54] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria. Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE Intelligent Systems*, 32(2):74–79, March 2017.

[55] Farhad Bin Siddique, Dario Bertero, and Pascale Fung. GlobalTrait: Personality Alignment of Multilingual Word Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7015–7022, 2019.

[56] Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. Emotion Detection in Text: a Review. Technical report, DeepAI, June 2018. arXiv: 1806.00674.

[57] Markus Luczak-Roesch, Ramine Tinati, Max Van Kleek, and Nigel Shadbolt. From Coincidence to Purposeful Flow? Properties of Transcendental Information Cascades. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, pages 633–638, Paris, France, 2015. ACM Press.

[58] Markus Luczak-Roesch, Adam Grener, and Emma Fenton. Not-so-distant reading: A dynamic network approach to literature. *it - Information Technology*, 60(1):29–40, March 2018.

[59] Jacob B. Hirsh and Jordan B. Peterson. Personality and language use in self-narratives. *Journal of Research in Personality*, 43(3):524–527, June 2009.

[60] Jon Oberlander and Scott Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634, Sydney, Australia, 2006. Association for Computational Linguistics.

[61] Fei Liu, Julien Perez, and Scott Nowson. A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 754–764, 2017.

[62] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, April 2013.

[63] Vivek Kulkarni, Margaret L. Kern, David Stillwell, Michal Kosinski, Sandra Matz, Lyle Ungar, Steven Skiena, and H. Andrew Schwartz. Latent Human Traits in the Language of Social Media: An Open-Vocabulary Approach. *PLOS ONE*, 13(11):e0201703, November 2018. arXiv: 1705.08038.

[64] Isil Doga Yakut Kilic and Shimei Pan. Analyzing and Preventing Bias in Text-Based Personal Trait Prediction Algorithms. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1060–1067, San Jose, CA, USA, November 2016. IEEE.

[65] John A. Johnson, Joseph Carroll, Jonathan Gottschall, and Daniel Kruger. Portrayal of personality in Victorian novels reflects modern research findings but amplifies the significance of agreeableness. *Journal of Research in Personality*, 45(1):50–58, February 2011.

[66] Mark Pagel, Mark Beaumont, Andrew Meade, Annemarie Verkerk, and Andreea Calude. Dominant words rise to the top by positive frequency-dependent

selection. *Proceedings of the National Academy of Sciences*, 116(15):7397–7402, April 2019.

[67] Lucie Flekova and Iryna Gurevych. Personality Profiling of Fictional Characters using Sense-Level Links between Lexical Resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[68] Kanika Narang, Austin Chung, Hari Sundaram, and Snigdha Chaturvedi. Discovering Archetypes to Interpret Evolution of Individual Behavior. *arXiv:1902.05567 [physics]*, February 2019. arXiv: 1902.05567.

[69] Yeyao Zhang, Eleftheria Tsipidi, Sasha Schriber, Mubbasir Kapadia, Markus Gross, and Ashutosh Modi. Generating Animations from Screenplays. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 292–307, 2019.

[70] Kristian Lindgren. *Information Theory for Complex Systems*. Chalmers University of Technology, Gothenburg, Sweden, 2014.

[71] Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T. Lizier. *An Introduction to Transfer Entropy*. Springer International Publishing, Cham, 2016.

[72] Joseph T. Lizier, Jakob Heinzle, Annette Horstmann, John-Dylan Haynes, and Mikhail Prokopenko. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *Journal of Computational Neuroscience*, 30(1):85–107, February 2011.

[73] Greg Ver Steeg and Aram Galstyan. Information Transfer in Social Media. Technical report, CERN, October 2011. arXiv: 1110.2724.

[74] Joseph T. Lizier. *The Local Information Dynamics of Distributed Computation in Complex Systems*. Springer Theses. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[75] David P. Feldman, Carl S. McTague, and James P. Crutchfield. The Organization of Intrinsic Computation: Complexity-Entropy Diagrams and the Diversity of Natural Information Processing. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 18(4):043106, December 2008. arXiv: 0806.4789.

[76] Thomas Schreiber. Measuring Information Transfer. *Physical Review Letters*, 85(2):461–464, July 2000. arXiv: nlin/0001042.

[77] Andrew Piper. txtlab Multilingual Novels (NOVEL450). *txtLAB*, 2016.

[78] Carlo Strapparava. Emotions and NLP: Future Directions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 180–180, San Diego, California, 2016. Association for Computational Linguistics.

[79] R. A. Calvo and S. D'Mello. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, January 2010.

[80] Saif Mohammad. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia, 2018. Association for Computational Linguistics.

[81] Joseph T. Lizier. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1, December 2014. arXiv: 1408.3270.

[82] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Technical report, Hugging Face, October 2019. arXiv: 1910.01108.

[83] Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas, 2016. Association for Computational Linguistics.

[84] Annick Lesne, Jean-Luc Blanc, and Laurent Pezard. Entropy estimation of very short symbolic sequences. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 79(4 Pt 2):046208, April 2009.

A

Appendices

A.1 Semantic parsing heuristic

Below is the pseudo-code for the two-pass heuristic algorithm used to parse the dependency tree of sentences (given by the DP model), identify sub-sentences and determine semantic roles.

Data: Document (list of tokens) $D = \{w_0, w_1, ..., w_n\}$, entities E, resolved coreferences $C \subset E \times D$
Result: Relations $R_{agent}, R_{patient} \subset E \times D$
`// Find agent and patient relations`
foreach *coreference* $(e, w_i) \in C$ *with head* w_h **do**
  if $dep(w_h) \in \{nsubj, csubj, poss, expl\}$ **then**
    $R_{agent} := R_{agent} \cup \{(e, parent(w_h))\}$;
  end
  foreach *ancestor* w_a *of* w_h *(under sub-sentence root)* **do**
    break if $\exists\,(x, w_a) \in R_{agent}$;
    $R_{patient} := R_{patient} \cup \{(e, w_a)\}$;
  end
end
`// Associate remaining tokens to the closest agent`
foreach $w \in D, \nexists\,(x, w) \in R_{agent}$ **do**
  foreach *ancestor* w_a *of* w *(under sub-sentence root)* **do**
    $R_{agent} := R_{agent} \cup \{(e, w)\, \forall\,(e, w_a) \in R_{agent}\}$;
    break if any;
  end
end

Algorithm 1: Semantic parsing heuristic. Some optimisations are omitted for clarity. Notably, only the subset of predicate words $P \subset D$ needs to be resolved.

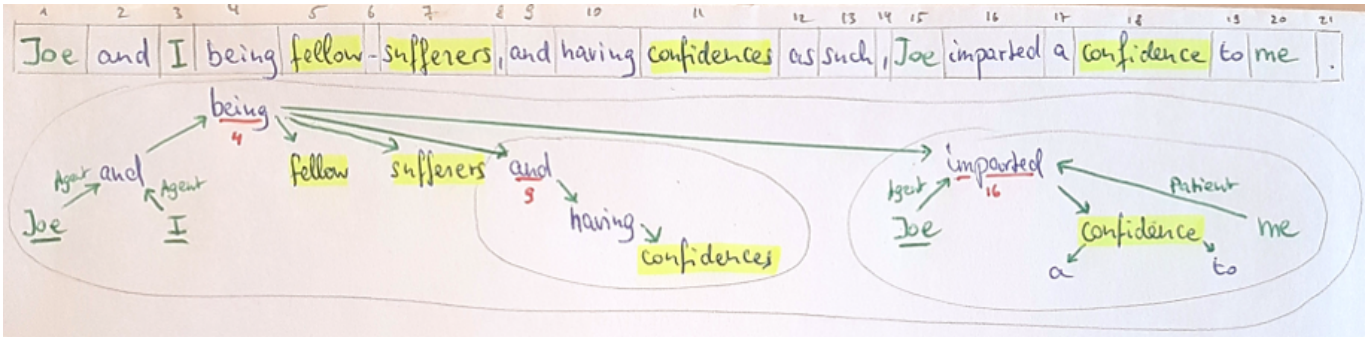

Figure A.1: Propagation of semantic roles through the dependency tree. Original sentence on top with numbers indicating positions in the text (time-steps). Upward arrows correspond to the first pass which propagates agent/patient relationships upwards, and downward arrows to the inheritance pass that associate tokens to the agent/patient entities from their ancestors. Circles represents sub-sentences, with their root in red. Predicate, i.e. words matched from the lexicon of psycholexical features, are highlighted in yellow.

The schema on Figure A.1 illustrates how the semantic roles are applied on the dependency tree. Table A.1 shows the expected semantic frames from the example sentence.

Time	Predicate	Agent	Patient
4	fellow	Joe	-
4	fellow	I	-
4	sufferer	Joe	-
9	confidences	Joe	-
9	confidences	I	-
16	confidences	-	me

Table A.1: Expected semantic frames resulting from the parsing of the example sentence on Figure A.1. Predicates are later replaced by the psycholexical features and entities classified according to rules explained in appendix A.2.

A.2 Entity classification rules

This flow chart represents the heuristic algorithm used to determine if a given entity mention (as given by the coreference resolution model) refers to a book character (`PERSON`) or another kind of entity (e.g. organisation, country, object...). In the latter case, it is categorised as `ENVIRONMENT`.

The coreference model gives a list of clusters of entity mentions, each cluster corresponding to an entity with one mention being labelled as the main one. Because sometimes the main mention is still inappropriate, the heuristic is applied on each mention and the final decision is made by selecting the most frequent decision.

Special entities *Narrator* and *Reader* are categorised as `PERSON`.

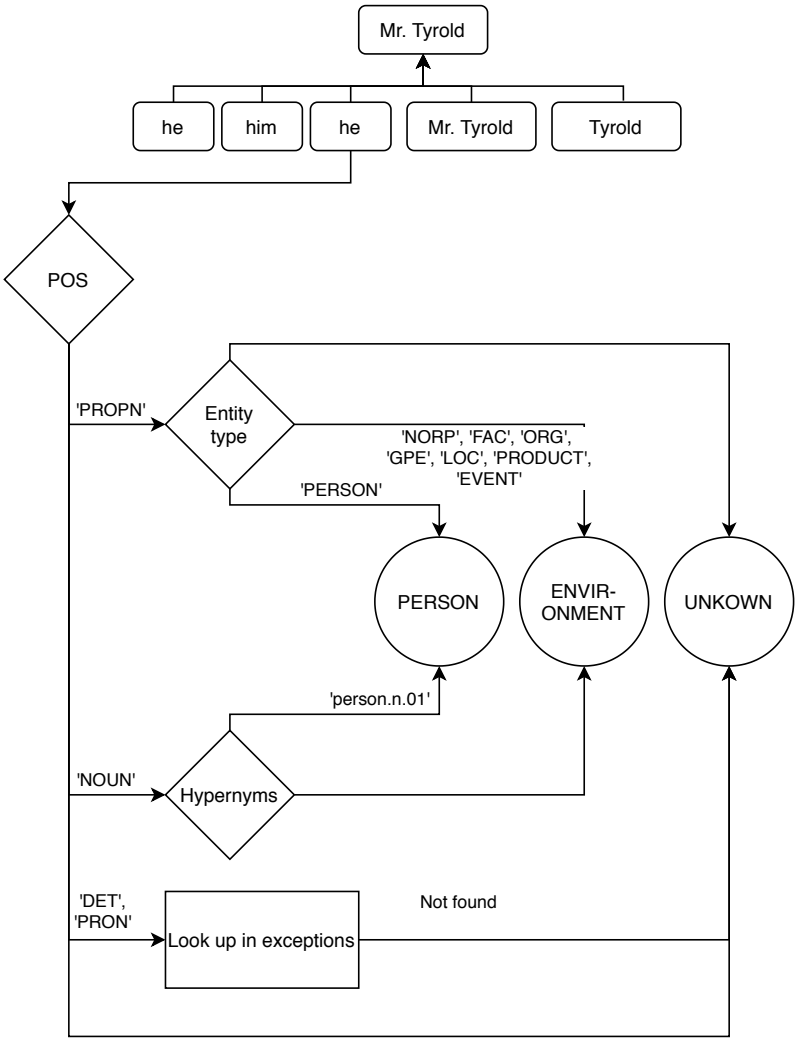

Figure A.2: Entity classification rules. On top: cluster of entity mentions. Each mention is classified based on its POS tag. Proper nouns (`PROPN`) are complemented with the POS tag, and other nouns with WordNet hypernyms (testing for *person.n.01* being one of the hypernyms). Undecided cases are set to `UNKNOWN`.

A.3 Universal Dependency Labels Reference

Label	Description
`acl`	clausal modifier of noun (adjectival clause)
`advcl`	adverbial clause modifier
`advmod`	adverbial modifier
`amod`	adjectival modifier
`appos`	appositional modifier
`aux`	auxiliary
`case`	case marking
`cc`	coordinating conjunction
`ccomp`	clausal complement
`clf`	classifier
`compound`	compound
`conj`	conjunct
`cop`	copula
`csubj`	clausal subject
`dep`	unspecified dependency
`det`	determiner
`discourse`	discourse element
`dislocated`	dislocated elements
`expl`	expletive
`fixed`	fixed multiword expression
`flat`	flat multiword expression
`goeswith`	goes with
`iobj`	indirect object
`list`	list
`mark`	marker
`nmod`	nominal modifier
`nsubj`	nominal subject
`nummod`	numeric modifier
`obj`	object
`obl`	oblique nominal
`orphan`	orphan
`parataxis`	parataxis
`punct`	punctuation
`reparandum`	overridden disfluency
`root`	root
`vocative`	vocative
`xcomp`	open clausal complement

Table A.2: Universal DP tags provided by the spaCy package, not including additional language-specific tags. From the reference documentation[1].

[1] `https://spacy.io/api/annotation#dependency-parsing-universal`