Master's Degree in Data Science
Final dissertation

On the Symbolic Universes of Language: from the Economy of Words to the Semantic Embeddings, a Study of the Units of Culture

Supervisor
Elena Pavan

Candidate
Andrea Leone

Co-Supervisor
Aurelie Georgette Geraldine Herbelot

Academic Year 2021/2022

Acknowledgments

I would like to thank my supervisor, prof. Elena Pavan, for her guidance throughout this project. Since high school, I have been fascinated by language and the way it enables information to flow through society bringing about substantial changes. This document is a compilation of my greatest academic passions.

To the provost of my college Marco and his deputy Mauro. We have known each other for ten years and counting, thank you for bearing with me and for all the thoughtful (and sometimes foolish) conversations we had along the way.

To the comrades of the *Café de la Paix*. You have been my daily spark of joy in these last months, thank you for the chats and the time spent together. I genuinely admire what you do.

To my colleagues Alessandro, Alessio, e Andrea. I am so thankful to call you my dearest friends. We had heaps of good times, and some tough ones too. If friendship is shown in times of need, you did way more than that. Do not forget: *choose life*.

Lastly, my family deserves immense gratitude. I don't know what I would have done without the love and support of my parents. You taught me the value of perseverance and have always pushed me to invest in my ideas whilst striving to learn and achieve more. A tender thought then goes to my grandma and the fond memory of my grandpa. I am so proud and grateful of what you have passed on to me.

... Cose appunto quanto mai esilaranti.
Pier Paolo Paolini, *Petrolio*, *Appunto 84*

Life is not a problem to be solved,
but a reality to be experienced.
Søren A. Kierkegaard

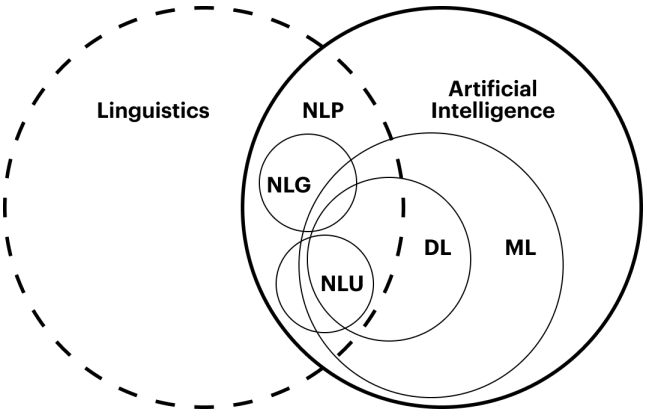
Abstract

The purpose of this dissertation is to present my multidisciplinary investigation into how personality traits and units of culture can be procured from the latent semantic knowledge that is embedded in a task-agnostic, pre-trained neural architecture for natural language processing. A systematic study that begins by delving into the concept of culture and unravelling several of its interpretations to assess an actionable definition that encompasses interpretations from sociology, anthropology, semiotics, ecology, and cybernetics. The focus of the examination then shifts to language and its remarkable characteristics that enabled our evolution as a sense-making species. As language offers a window on the cognitive processes through which we conceptualise the world and arrange interactions and social roles, I shaped my assumption that the words we use reflect and reveal much of our culture in a given situation. To exhibit the hyper-generalised systems of meaning representative of a cultural milieu, I then analysed the symbolic universes displayed by semantic networks aroused from a single topic. These mappings were computed by leveraging word embeddings, scilicet, vector representations of a word's meaning obtained by learning its statistical occurrences in an extensive corpus of text. The rest of the dissertation is dedicated to the experimental appraising of such embeddings, benchmarking their results, and evaluating the performances of baseline statistical machine learning models along with state-of-the-art transformer architectures.

Eventually, I unravel the setting for zero-shot learning and showcase my results. To derive the indicators for personality traits and units of culture, I exploited a transformer pipeline that overturns the conventional supervised approach of semantic categorisation. Instead of training a classifier to recognise a set of labels, I designed an algorithm that takes advantage of a massive pre-trained language model fine-tuned for information entailment estimation. The model relies on its internal meta-semantic understandings of language to infer how much each label is descriptive of a specific document. The intermediate scores are ultimately aggregated into idiosyncratic insights.

Table of contents

Project Repository • Data Repository



Related Works

Every section of this dissertation has a dedicated literature review. The main themes discussed throughout my investigation are clustered below.

Defining culture: Sewell [266], Keesing [139], Levinson and Ember [164], Ember [88], Chick [39], Roberts [233][234], Rambo [229], Lynch [179], Pagel [208], Damen [66];

Units of culture and memes: Chick [38][39][40], Durham and Weingart [81], Durham [79][80] , Blackmore [20], Dawkins [68][69][70];

Personality Traits: Allport [316][317][319][320];

Language evolution: Everett [92][93][94];

Cybernetics. Bateson [17][18][19], Ashby [11], Wiener [305][306];

Language and Semiotics: Peirce [212], Pinker [216][217][218], Wittgenstein [312];

Semiotic Cultural Psychological Theory: Salvatore et al. [251][252][253], Ciavolino et al. [51], Veltri et al. [293], Valsiner [290][291], Venuleo et al. [294][295];

Symbolic universes: Blumer [31], Salvatore et al. [249][254][252];

Economy of words. Zipf [320];

Semantic embeddings: Bengio et al. [24], Jurafsky and James [134], Globerson [146], Navigli [200], Mahesh et al. [181], *Word2vec*: Mikolov et al. [189][190][192][193][194], *GloVe*: Pennington et al. [213], Turian et al. [288];

Statistical NLP: Johnson [130], *Ensemble learners*: Ho [119], Quinlan [225], *XGBoost*: Chen et al. [37];

Neural NLP: Rumelhart et al. [240], Goldberg et al. [104][106], Bengio et al. [25][26][27], Qiu et al. [223];

Neural architectures: Goodfellow et al. [107], *FFNN*: Glorot et al. [102], *CNN*: Albawi et al. [1], Fukushima [99], Ioffe et al. [127];

Transformer architectures: Vaswani [292], *BERT*: Devlin et al. [76], *RoBERTa*: Liu et al. [174], *SqueezeBERT*: Iandola et al. [126], *DistilBERT*: Sanh et al. [255], *BART*: Lewis et al. [168];

BERTology: Rogers et al. [236], Clark et al. [52], Michel et al. [210], Tenney et al. [284][285], *Attention*: Vig [297], Lin et al. [171];

Zero-shot learning: Davison [67], Yin et al. [314], Radford et al. [227], Larochelle et al. [154], Romera-Paredes et al. [237], Edunov et al. [84]

Datasets and Tools

Since the very inception of this project, it was my determination to work on a large collection of talks from TED, the multinational nonprofit media organisation that promotes «*ideas worth sharing*», and arranges conferences and independent events around the world. A talk essentially consists in a captivating monologue encompassing a wide variety of topics and lasting about ten minutes. The main reason of my interest is convenience: because of their format, the transcripts of the talk provide clean, original utterances that can be used at ease to estimate personal traits and cultural descriptors. They are also annotated with the topics of the discussion, making them really handy for the analysis of the symbolic universes through semantic networks in section Three.

Hence, I assembled a database of 5400 talks from TED's online archive with a custom web scraper. Among the general information of the talk, each record features the full English transcript and a ranging number of tags. As tag labels are manifold and their distribution across the entire dataset is highly uneven, I designed a scoring system to sort every record in one macro category out of three: science and innovation (34%), culture and society (39.4%), economy and environment (26.6%). The result is a rather balanced and diversified dataset, suited for the classification task in sections Six and Seven.

Tools and Libraries

- PostgreSQL: the world's most advanced open-source relational database.
- Jupyter: web-based interactive development environment or notebooks, code, and data.
- NumPy: fundamental package for data structures and scientific computing.
- Pandas: fast, powerful, and flexible tool to analyse and manipulate data frames.
- NetworkX: popular package for the creation, manipulation, and study of the structure, dynamics, and functions of complex network graphs.
- ExplosionAI's SpaCy: industry standard library for Natural Language Processing operations.
- Scikit-Learn: simple and efficient suite of methods for predictive data analysis.
- XGBoost: scalable, flexible, and configurable gradient boosting algorithm.
- PyTorch: production-ready ML framework.
- Huggingface's Transformers: comprehensive repository of transformer-based architectures for NLP, NLU, and NLG.
- Matplotlib: popular and versatile plotting library.
- BertViz: interactive tool for visualising attention in transformer language models.

What Is Culture?

Not many words in our dictionary are as complex to explain as *culture*. There is actually no standard, commonly accepted definition. Asking around, I always hear a profusion of elaborate, enthralling interpretations about collective and personal expression and its indispensability to enrich life. Few other terms are as persistent and pervasive in the modern discourse, and even less feature so many conceivable contrasting meanings.
In this section, I am going to revolve the different meanings of the word *culture* and propose a working concept for the exploration I want to cover in the dissertation.

Four Main Descriptions

The English word *culture* descends from German as a sheer agricultural metaphor. As we cultivate barren fields to bear fruit, we nurture our minds with fine arts and literature. We refer to it as *high culture*:[10,85] the serious, intellectually-rigorous, involved art forms celebrated in established institutions like museums, galleries, and academies. High culture has permanent value and provides some standards of excellence. Precisely like religion, it is a restricted cult with its ministers: historians, theorists, critics, and writers. Elitism is therefore inevitable, as these art forms take on their most crucial meanings only after intentional study and prolonged exposure. Without investing time and effort, high cultural artefacts do little more than serve as background.[307,308]

Studying human evolution, anthropologists extended the original elitist definition of culture to the customs and practices of pre-modern tribal societies.[164,32] This provided a second application: *communal culture*,[309] an unconscious repository of behaviours, rules, and norms that lead the individual lives of people in a shared ecosystem. We tend to be blind to our own communal culture until we face different customs from an outsider. As English philosopher Roger V. Scruton wrote, culture is «*the defining essence of a nation, a shared spiritual force manifest in all the customs, beliefs, and practices of a people*».[263] We are always part of a group of individuals involved in a context of persistent social interaction.

Still, when referring to our contemporary culture, we are unlikely to think about clerkly intellectualism, elite art, or folk reminiscences. For the last century, we have observed the emergence of *pop culture* and its diffusion to every aspect of our lives:[309] movies, TV shows, hit songs, fashion trends, food, sport, tech devices, mobile apps, slang — our enthusiasm is mainly driven by commercial products and entertainment content that is featured on a highlight outlet. The primary driving force is mass appeal, which permeates people's lives and influences their attitudes.[139,280]

Ultimately, in recent decades, a new meaning has emerged in the workplace: the *organisational culture*, a collection of goals, values, and practices that better align coworkers in achieving the company's objectives whilst raising productivity.[258]

Components and Commodities

Perhaps, what makes terms like *Economics* and *Psychology* much easier to grasp is that both are sectors of human life organised around specific aims, principles, rules, and actions. *Culture*, by contrast, feels like a vague concept that is somehow connected to society. Not everything is culture, though. It unerringly entails our lives and comprises discrete components: material objects, concepts, attitudes, behaviours, meanings, and values. Culture is what we eat, the news we read, the people we meet or follow on social media, the music we like, our habits and traditions, leisure activities, job perspectives, careers, political views, the choices we make for the future, our urgencies, inspirations, priorities, and how we feel about all of these things at a certain moment of our life, in a definite socio-economic context.

Sociologists view culture as consisting primarily of the symbolic, conceptual, and intangible aspects of human societies.[31,307,308] The focus is indeed not on the material elements themselves but on how the group members interpret, use, and perceive them.[16] Culture is mankind's primary adaptive mechanism:[66] a collection of distributed models for living that pervade all aspects of human social interaction, providing the individual with meaning and purposes. These patterns generally fall into one of the following categories.

- **Customs**: unconscious habits and behaviours distinctive to a community.
- **Traditions**: conscious rituals performed to celebrate and retain communal values.
- **Lifestyle**: the way according to which modern individuals choose to live and spend their time.
- **Leisure**: voluntary activities for entertainment, amusement, and restoration.
- **Art**: intentional creation and aesthetic contemplation of themes and experiences.

These categories highlight some relevant contradictions in the notion of culture.

- *Is culture unconsciously inherited as customs, or consciously created as art?*
- *Is culture directed by elites or is it shaped by the occurring behaviour of the masses?*
- *Is culture meant to preserve values over generations in the form of traditions, or challenge them through innovation?*

As a matter of fact, culture is not a coherent concept, since more internal contradictions are revealed as we deepen into it. Some believe that these intrinsic discrepancies are what make culture such a powerful concept. According to American historian Joshua D. Rothman, «*culture is more than the sum of its definitions; if anything, its value as a word depends on the tension between them*».[239] This same inconsistency reflects the intricate relationship between the individual and the social group they are embedded in.

German sociologist Georg Simmel stated that culture is «*the development of human nature beyond its natural state*».[267] The environment does considerably influence the parameters of culture, however, within such constraints, humans show incredible resourcefulness in devising a wide range of material and conceptual solutions to meet their social needs. This leads to the definition of Russian sociologist Pitirim A. Sorokin, who, focusing on this aspect of interaction, writes that culture is «*everything which is created or modified by the conscious or unconscious activity of two or more individuals interacting with one another or conditioning one another's behaviour*».[277]

Information

All definitions of culture I encountered in my research start from shared ideas and knowledge within a group – including beliefs, attitudes, values, and ideals – organised in systems of meaning that underlie how people live[3]. Social habits are the manifestation of these systems, exposing the individual's thoughts, emotions, and feelings to the outer world. Patterns of behaviour usually reflect learned values of a particular society or population,[164,88] as do material objects.

From an alternative perspective, culture can be defined as the *information* shared by a social group and exchanged in a sort of marketplace.[233,234] This notion is perfectly compatible with the previous one, yet more general, inclusive, and apt for evaluating the concept of units of culture.

When viewed as information, culture comprises shared knowledge, behavioural patterns, and material artefacts that are distinctive of either small groups of individuals or large aggregates of people. This information can be archived and recorded in devices specifically designed to store it, like stone tablets, books, or computers. From this standpoint, cultural interchange and evolution may be viewed as additions to, deletions from, or alterations to a distributed information repository.[38,39] Cultures can sometimes be lost when societies lose their members, collapse on their own, or are conquered by outsiders, as in the case of the Mayans or the Roman Empire.

Broadly defining culture as shared information removes the need to restrict it to specific categories. Information theory suggests that information can be defined as discrete units known as *bits*: if culture is information and it can be treated as discrete units, it would therefore seem reasonable for culture to be considered in terms of discrete units, or at least somehow measurable and quantitatively describable.

Arbitrariness

Despite the seemingly contradictory definitions I provided in the previous paragraphs, it is reasonable to think that culture relates to the social part of life beyond biological instincts, economic activity, and technical requirements. This is manifested in lifestyle, traditions, customs, leisure, and art — none of which is static. Looking at culture in action, we observe dynamics of comparison and competition that sometimes lead to disagreement and conflict. In linguistics, the term *arbitrary* is used to describe the fact that there is no natural relationship between the articulation of a word and its meaning, as another sound can theoretically signify the same concept. Whilst culture is more complicated than natural language, the same idea can be extended to explain cultural practices.

There are always specific circumstances that explain why our culture is *ours*. Most of our personal choices are more or less arbitrary, as we tend to adopt a cultural trait depending on our sense of belonging to a group and our willingness to take part in its dynamics. Arbitrary choices are fundamental to culture, as humans have to frequently engage in what game theoreticians call *coordination problems*. Choosing to dress glamorous and voguish for a party, for example. It means complying with the rules of a specific social

group and accepting a set of implied conventions and meanings. These are called *problems* because there are countless solutions and different ways to coordinate among peers that lead to different outcomes.

When talking about culture, we are indeed describing the arbitrary aspects of human behaviour. To survive, humans generally need food, shelter, clothing, and a means to communicate. In the words of American anthropologist Marshall D. Sahlins: «*men do not merely survive; they survive in a definite way*».[245] Again, there is a nearly infinite number of these "definite ways" for eating, drinking, dressing, speaking, thinking, and enjoying our time alive; the same is just not true for agricultural practices, medicine, or mathematics. For the cultural aspects of life, many alternatives can serve the same purpose, which makes the final choice so arbitrary. This becomes evident when we think about customs. English economist Adam Smith wondered why our customs «*though no doubt extremely agreeable, should be the only forms which can suit those proportions, or that there should not be five hundred others, which antecedent to established custom, would have fitten them equally well*».[274] Meeting a stranger, for example. It does not have to be a handshake, as members from another culture would instead bow. Social groups enforce norms like this and, over time, develop a complete sense of aesthetics to pursue and defend a sense of identity. We detest seeing our culture as arbitrary and replaceable. Italians, for instance, love to remark that they are masters of good coffee, and pineapple on pizza is nothing but outraging blasphemy for them.
On the matter, Norwegian economist Jon Elster commented that «*human beings have a very strong desire to have reasons for what they do and find indeterminacy hard to accept*».[87] Even when we make arbitrary choices, our brains often provide post-facto rationalisations. Habits become heuristics that silently govern our ordinary expected behaviour, freeing our minds from solving these problems every time over when not driven by biological instincts and economic rationality. A question arises spontaneously: *why do we make the same choices when other options can serve the same purpose?* Theoretically, we have countless alternatives, but in practice, we feel no choice at all as the environment and our previous actions already set our path.

Conventions

If culture orbits around the arbitrary aspects of social life happening in concert, conventions then give us an explanation of why social groups repeatedly stick to the same arbitrary choices over so many alternatives. The word *convention* is slightly aloof and mainly used to discuss stylistic and artistic waves. To puzzle out culture as a macro-phenomenon, we have to understand the mechanism that pushes humans into such customs. For American philosopher David K. Lewis, conventions are regular, well-known, and socially accepted behaviours that individuals follow and expect others to follow. They can be natural in relation to a given context, completely arbitrary, or wholly made up.[244,245]

Conventions elucidate many of the components of culture. Common practices are plain long-standing conventions that people notice only upon contact with a possible alternative. The chances are that the sole actionable way to measure and understand two cultures is to directly compare them, dispassionately. We all believe that no culture is superior to another, and with the same objectivity, we can compare and detect the patterns that can describe human behaviour.

We tend to be more cognisant of conventions when they are *manners,* like how to set the silverware for a formal dinner, because they take effort to follow. Traditions are nothing more than conventions anchored in historical precedence and serve as explicit symbols for the community. Eating a traditional dish may not just be a conventional dietary staple in a region but a way to feel part of the local culture. Superstitions are conventional beliefs, like the number thirteen being unlucky in the US and the opposite in Italy. Modern life is full of short-term conventions called fads, and fashions are another kind of convention that regularly changes when it comes to style and expression. All artistic trends are conventions: to paint like a Romantic or a Cubist means something that is specifically contextualised in space and time and can be defined with a conventional explanation. When it comes to comparing artists, things get nuanced. We start from the elemental definition of the movement, and through collation, we outline a profile of the artist, acknowledging their similarity and uniqueness compared to their fellows. Because artists are individuals – each one with their own experience and perspective of reality – their contribution to the movement is not

tied to *how much* Romantic they are, but determined by the resonance of their work in the present and future cultures. People's perspectives change over time, and so do artistic influences.

Arbitrariness and conventions are strictly connected: humans do not need them to breathe, yet they are essential in solving coordination.[307] Conventions draw their power from our emotional responses to expectations.[139] Cognitively, we need conventions because our brain avoids expending extra mental energy on thinking through a wide range of alternatives. When our expectations get disappointed, we become frustrated, even in times when the underlying behaviour has no substantial impact on us.[32] We then convert the emotional responses, either positive or negative, into outward expressions. Meeting expectations elicits smiles and cheers while failing to meet them causes hostility and acrimony. For instance, we deal with this in the social discourse about integration. Whether it is a foreign student in a class or an insulated ethnicity in a town, when people refuse or find it hard to assume the traits of mainstream culture, they often get mocked or set aside by the less open-minded. Individuals who cannot find refuge in a smaller social group eventually switch to the dominant convention to win the favour of the larger community.

Conventions spring up to solve coordination in a group and quickly become social norms. As people follow them, they refine the behavioural patterns and the sense of aesthetics we associate with that group. If appropriately measured, these patterns have the potential to become units of culture.

Over time, we internalise the conventions of our society. We follow them without even considering the alternatives, forgetting that there is a latent choice we are neglecting. Conventions become habits, and more importantly, they form the very perceptual framework in which we understand and represent the world. Colour, for example. Even though we are able to distinguish between 7.5 to 10 million different hues, our linguistic conventions determine how we derive that spectrum into specific chromatic units. Italians, for instance, perceive blue in two distinct shades: lighter (*azzurro*) and darker (*blu*).

As culture describes arbitrary behaviours and cultural patterns involve conforming to conventions, we can reckon that all cultural behaviours can be explained through conventions. Conventions explain how culture can be both *conscious* (the intentional following of conventions for social coordination) and *unconscious* (the habitual following of internalised conventions). The function of art is also clear at this point: artists and creatives propose new ways of perceiving the world that may have the potential to become conventional, to shift people's perspectives, and ultimately to change the socio-cultural narrative.

Social influencers are generally the agents that poke and incite changes in culture. They present a creative, innovative idea to their public, and if it is appealing enough, their impact grows and expands. This process is essential for an idea to reach a broader audience and be heard. Even in times when certain behaviours offer clear, practical advantages over alternatives, conventions are constantly reshaped by public discourse and social events. Habits and expectations that stem from conventions are also very fluid. What is not so liquid is high culture: from the classics of art and literature to the most iconic articles of clothing or guitar riffs, there are symbols that are hard to replace or outshine.

Meanings and Values

Conventions clarify why we follow the same regular behaviours, and they also describe how groups agree on values and meanings. Conventionally, baby boys wear blue and baby girls wear pink. Like any other, this is an arbitrary assignment. Most people conform to this rule when buying clothes for a newborn, and manufacturers make the convention even easier by selling products that adhere to this arbitrary rule, reinforcing the convention. Cultural norms demonstrate how conventions set our behaviour and how they provide us with meaning. A blue or pink ribbon hanging on the front door clearly symbolises the recent birth of a boy or a girl. In this context, pink is not just a flattering colour but a distinguishable and indisputable social signifier.

Conventions transcend being mere recipes for behaviour and end up being communicative signs that are rich in meaning and connotations. They allow us to dive into a deep rabbit hole of associations through a single object. Culture is never just a linear list of conventions but a convoluted and dynamic nexus with new connections being created and old ones being discarded every day. «*The sociocultural world,*»

writes Russian-American sociologist Pitirim A. Sorokin, «*consists of endless millions of individual objects, events, processes, fragments, having an infinite number of forms, properties, and relationships*».[278] Conventions quickly become social norms. They are not just the regular way of behaving, but *the* proper way. In the process of sharing conventions within a community, values are formed. Whilst individuals share certain practices, conventions help form social bonds through shared meanings.

I started this section with the tangled intention of clarifying one of the most ambiguous words in the English language, and now I can draw an actionable definition of it. Culture denotes the conventions of a community, which guide the individuals into regular patterns of behaviour and provide communal meanings and values we can use to describe it. Much of it is about conventions, the atomic units of cultural behaviour.
They explain:

- why culture encompasses our conduct, habits, language, and art;
- why culture manifests as customs, traditions, styles, fashions, and fads;
- why it can be both conscious and unconscious;
- why individuals are so devoted to their customs and wary of equally-valid alternatives.

To reason about culture is to locate specific conventions in a social group and deconstruct their origin, practice, outcomes, and connectedness to others. At their core, conventions arise around arbitrary behaviours, and I think it is important to highlight three crucial aspects.

- *Arbitrary does not mean cultural activity is random*. Cultural possibility overrides narrow biological determinism: some traits are inherently biological, yet the specifics of our conventions are never accidental. They arrive through and are a continuous product of specific historical circumstances.

- *Arbitrary does not mean equal outcomes for all behaviours*. Different conventions have different consequences. Some are productive, others are harmful and biased. We look at our arbitrary practices through the eyes of modernity and maybe think about whether a different arbitrary practice would be more beneficial to our community. Cigarette smoking, for example, was a well-established convention for most of the 20th century in Western countries. Today, it is forbidden in almost every public venue. Once its negative effects become clear, a behaviour becomes niche and segregated, or even some sort of taboo. It may be arbitrary whether an individual smokes or not, yet its effects on health are not. Either way, it is an interesting cultural trait to observe when studying the dynamics of a social group.

- *Controlling conventions is a form of power*. Fashion is the best example of cultural arbitrariness. At every single point in time, there is a *right* and a *wrong* way to wear jeans, and those who conform to the proper convention are rewarded with social status. Conventions give arbitrary practices a differential social value.

The Units of Culture

Reducing large, complex entities into smaller, simpler units has been one of the most successful strategies in Western science for more than two millennia. Social science also attempted to study human culture considering this principle, and despite its clear line of successes in the West, it has not been universally accepted. In 19th century Europe, Belgian statistician Adolphe J. Quételet and French philosopher Auguste Comte proposed a scientific, reductionist approach to the study of culture. This new discipline, *social physics*, was never truly adopted at the time but laid the foundations of what Comte will later on call *Sociology*.[128] Yet now, the thriving field of data science is reviving the analysis of social phenomena, looking for regularities or behavioural patterns to describe socio-economic events. We witnessed the emergence of computational social science, modelling human interaction through its plentiful digital footprints in the cybersphere. Such ample abundance of data sources marks a radical new era of empirical research: whilst researchers in the past had to design end-to-end experiments to gather the data they needed, today, they can find answers to their questions by digging and scraping online platforms or sampling massive open-source datasets. This wealth of digital human records brought, for instance, to the consolidation of *Culturomics* — a computational lexicology that studies patterns of human behaviour and cultural trends through the quantitative analysis of digitalised texts.[161]

Conversely, Ancient Chinese tradition always emphasised the holistic and continuous character of the universe, focusing on nature's harmonious and hierarchical properties.[78] According to this worldview, nature cannot be subdivided into discrete and constant elements. A reductionism distrust that was also widespread among many Western social scientists of the 20th century and I believe it is noteworthy to elucidate why. Such scepticism is mainly due, on one side, to the idea that efforts to reduce social phenomena into smaller, more elemental units somehow deprive people of their humanity. On the other, reductionism might lead to an oversimplification of the social dynamics it aims to study, as cultural behaviour is too complex and integrated for the analysis of smaller, simpler components to be of any value. This conviction has been bolstered by the study of *complexity*, where higher levels of organisation exhibit emergent properties that neither exist at lower levels nor could be predicted from them. In other words, the whole is more than the sum of its components.

The concern over cultural change, evolution, and transmission has, however, involved a debate on the possibility and necessity of the units of culture. Two quotations frame the discussion.

«*Our definition of culture is not at all specific about the nature of the information that affects phenotypes. In particular, we do not assume that culture is coded as discrete "particles". Moreover, it is possible to construct a cogent plausible theory of cultural evolution without assuming particulate inheritance*». — Boyd & Richerson, 1985[32]

«*Formulation of a theoretically plausible, and empirically usable, characterisation of a unit for storage and transmission of cultural information is, in my view, the major task that must be completed before further advance can occur in the study of cultural evolutionary processes*». — Rambo, 1991[229]

Reflecting on T. A. Rambo's angle, a question immediately emerges: *which, if any, unit of culture has both theoretical and empirical merit?* Whether culture can be encoded in discrete units is also pertinent to another, more general case in cross-cultural research and culture theory: *What is the nature of this culture-bearing unit? How can we distinguish cultures from one another? What should be the basis of difference be, and when two cultures can be regarded as distinct?*

Questions regarding the unit of culture and the culture-bearing unit are intricately related. Let us suppose some theoretically plausible and empirically sound unit of culture is discovered: it could be used to outline and single out cultures as these would feature different scores. It would still be necessary to know – or decide – what degree of dissimilarity is sufficient between two culture-bearing units to make them actually *different*. Numerous labels have been applied to the components of culture: some of these (such as themes, configurations, complexes, and patterns) appear to be at high levels of cultural organisation; other ones (such as ideas, beliefs, values, rules, principles, symbols, and concepts) instead seem to be operationalised at lower, more fundamental levels. Thus, the higher-level labels give the impression of being somehow particular arrangements of the lower-level units.

A substantial part of the concern over the units of culture derives from analogies made between biological and cultural evolutions. Some researchers saw the need to adopt a particulate unit of cultural transmission analogous to the gene. The most notable include C. Lumsden and E. O. Wilson's *culturgen*[178] and Richard Dawkins' *meme*[13]. Thereupon, American evolutionary biologist William H. Durham adopted the meme as the unit of cultural transmission as part of his coevolution theory[81], while several individuals outside of anthropology have embraced the notion with few apparent misgivings.[30,72,73,74] Some even proposed a new field called *memetics*.[30,179] After a decade, Wilson abandoned his construct and also adopted the notion of meme, applying some variations to the original. Howsoever, indifference or scepticism seem to be the most common positions about the meme. Towards the end of 20th century, Rambo suggested that there is no credible unit of cultural selection — long time after Boyd and Richerson assessed that culture is not necessarily composed of discrete particles nor that such an assumption is required for a plausible theory of cultural evolution.

The investigation into the utility of the units of culture, mainly furthered by American professor Garry Chick[40,41] in the early 00s, provides us with the necessary insights to know what to look for in a unit of culture, however, no practical example has ever been proposed. Persisting with a more abstract outlook, present-day technology has the potential and the flair to detect and measure such patterns in the observable social behaviour.

The Meme

Before setting the evolutionary inquiries aside, we can glean a working definition of unit of culture from the various interpretations of a *meme*. Introduced from an anthropologist standpoint, the neologism refers to a convention that spreads by means of imitation within a social group and to the symbolic meaning embedded within.

«*Just as genes propagate themselves in the gene pool by leaving from body to body*», writes British evolutionary biologist Richard Dawkins, «*so memes propagate themselves in the meme pool by leaping from brain to brain via a process which, in the broad sense, can be called imitation*»;[69] and he continues: «*If a scientist hears or reads about a good idea, they pass it on to their colleagues and students. They mention it in their articles and lectures, and if the idea catches on, it can be said to propagate itself*».[70] Initially conceived by Dawkins as a cultural replicator, the concept of meme was later refined as a unit of information that resides in the brain, «*just as genetic information is stored in the DNA*».[68] Following this analogy, the phenotypic effects of a meme are expressed in the outside world in the form of words, fashion, gestures, and skills. These manifestations are perceived by our individual senses and can be imprinted on our brains.[68] At the same time, they can be computationally measured and analysed to model social behaviour at scale. Sociologically, there is no interest in looking for the elements that pre-constitute culture, but we can welcome an empirical reductionist approach in the attempt to decompose cultural conventions into smaller, quantifiable units. In this light, the meme offers some interesting contributions. Through data, we can look at the individual social choices that are exhibited in public and use them to give an account of a culture and its context.

For Durham, the meme represents «*actual units of socially transformed information, regardless of their form, size, and international organisation*».[79] According to him, «*whenever culture changes, some ideational unit is adopted and one or more homologous alternatives are not*»,[79,80] and provides two forms in which a meme is manifested. *Holomemes* represent the entire cultural repository of variation for a given meme, latent or unexpressed forms included. *Allomemes* instead refer to the subset of holomemes that are used as behaviour guides by (at least some of) the members of a social group.[80]

American cognitive scientist Daniel C. Dennett succinctly described the meme as «*the smallest unit that replicates itself with reliability and fecundity*»,[73] while American biologist E. O. Wilson defined it as «*a note of semantic memory [that] correlates in brain activity*».[311] Tracing the profile of an ideal unit of culture, Durham[80] suggested that it

- consists of information that actually or potentially guides behaviour;
- combines highly variable kinds, quantities, and ways of organising information (with variable amounts of hierarchy and integration);
- and singles out pieces of information that are differentially transmitted as coherent units.

On the basis of his three criteria, Durham discarded all terms that might align with the concept of unit of culture, except for two: *symbol* and *meme*. For empirical research, it is hard to directly track and estimate memes, but the idea of them determining behaviour through meaning is truly beneficial in my quest for the units of culture. After all, patterns of behaviour are (more or less arbitrary) conventions that are adopted by a social group. As elucidated by French sociologist D. E. Durkheim, the nature of the ritual is relatively insignificant: what is relevant is that people share its practices and evoke the same ideas and sentiments.[82] The reiteration of these actions lends them a sense of identity and makes them feel part of something larger than themselves. Having a family dinner every Sunday, for example. The ritual enables social cohesion, provides the individual with meaning, and denotes a set of meanings and values about its upholders. It might not be a suitable unit of culture per se, but the example is rich in socio-cultural insights, reflecting personal values and prime concerns beyond the tradition itself.

Individuals are often induced to adopt concepts and habits from the cultures they find themselves part of, but until now I have always been remiss about the aspiration of the single to break free from established conventions and take up something different or radically new. On a larger scale, such openness inherently leads to social contagion: cultures meet and, in due course, mate. When this happens, in the words of British author and journalist Matthew W. Ridley, «*human beings bring together their brains and enable their ideas to combine and recombine*».[232] Indeed, when «*ideas have sex*» is the precise moment in which actual

progress is ignited and innovation is allowed to happen. This is possible thanks to our ability to share and retain knowledge, to reason, and to socially gather and organise. Every technological improvement in our history results from cumulative ideas being reshaped, refined, challenged, or overturned when enough diversity is available. This highlights the astounding ability of the individual to assimilate new practices, but also makes us focus on the continuous, unconscious process of respective persuasion on which we all play our part. Language is the perfect scenario: we did not choose our mother tongue, yet it determines and reflects much of the cultural milieu we grew up in, its practices, and its values. If we look at it in a macro perspective, language is a custom — a collection of largely established conventions we need to understand each other. In the micro, however, language is highly nuanced. Oftentimes we use slang, vernacular, or dialects when interacting with the members of a specific community. Communication is essential to form and maintain social bonds, and the words and concepts we employ become the building blocks we use to describe our reality and experience.

Ecology and Cybernetics

Early in this section, I ventured to align cultural conventions with units of culture. It worked in my mind at first, but the idea was fragmentary and needed some refinement. In particular, it is nearly impossible to directly measure conventions, ideas, or memes. What we can do, however, is to observe behaviour in a context and reason on the cultural patterns revealed at scale. Knowing the general characteristics of a social group, it is easier to correlate features. Music genre and relative outfit, for example. If we think about punks, hippies, or trappers, our mind rushes in a stereotypical depiction of the group, first and foremost influenced by the aesthetic and perception we previously sensed, then by our socio-emotional response and the information we gather from our memory. As highly symbolic beings, we rely on social clues to access meanings in both physical and virtual environments. This adaptation process has been called *cultural ecology* by American anthropologist Julian H. Stewards, in his theorisation of *multilinear cultural advancement* analysing how societies attune to their habitat, including processes of inter-relation and modernisation. In the mid-twentieth century, Stewards proposed the historical study of micro-cultures that are representative of specific areas or regions, identifying the decisive factors that determine several development directions, such as economics and technology, but also politics, ideologies, and religion. Hence, whilst the environment influences the character of human adaptation, it does not necessarily determine it, crumbling the concept of environmental determinism over human actions. Stewards' innovative method specifically involved the systemic observation of the cultural activities associated with the environment and the assessment of how much behaviour patterns influence other aspects of culture. The term *ecology* refers to the study of the relationships between living organisms and their environment. Theoretical ecology is indeed devoted to analysing complex biological systems through models and simulations to reveal the dynamics of selected groups of individuals in a vast, ever-changing natural context. Although the biology field certainly offers ample inspiration for my research, I am keener to apply the ecological concepts to model the information exchange throughout a social group. *Information ecology* is a popular metaphor that assumes the information space to be an ecosystem of concepts interrelated among them, the result of the human semantic activity conjugated with the ontological world. What is fundamental in this perspective is the interaction between two individuals. Recollecting the notion of meme and wondering about the process that has «*the same effect in cultural evolution as sex is having in biological evolution*»,[232] Matt Ridley values exchange as a unique human feature.[231] In point of fact, it is a behaviour that has never been observed in other animals. Sure, there is some form of reciprocity, but the exchange of one object for another is exceptionally ours.[230] Culture without exchange – or «*asexual culture*» – is possible nonetheless, says Ridley. Other animals have this kind of culture: it is common for parents to pass on some practices to their offspring. Chimpanzees, for instance, teach each other how to crack nuts but their culture never grows and expands from one generation to the other. It does not become combinatorial as there is no room for diversity or innovation.[230,231] Again, the perfect example to depict humans' capacity to coordinate, despite our inherent diversities, is language. Regardless of our individual differences, social backgrounds, experiences, and values, we are able to communicate and comprehend each other.

Moreover, language is the social act that mainly allowed us to go beyond self-sufficiency, develop a sense of vertical expertise in the division of labour, and rely on a broad collectivity for exchange and support. Besides, language responds to our need for expression and reflects our personality traits: in its broad acceptation, it is the cultural protocol of organisation, reciprocity and influence.

It is thus common to conceive culture as the common ground that enables *stigmergy*,[169] a behavioural mechanic of indirect coordination between individuals due to the information they gather from their environment. By reading our surroundings and observing other people's behaviour, we are able to interpret a situation and self-organise. Mostly quoted when analysing the structured behaviour of insects like bees or ants, the term has been reconsidered to model human actions at scale, especially in cyberspace, on social platforms like Wikipedia or GitHub. Originally introduced by French biologist Pierre Paul Grassé, *stigmergy* has been revamped by the research in the user-centred design of American professor Donald A. Norman, merging usability engineering and cognitive science. Delving into the social role of aesthetics and affordances for ergonomics, he emphasised how humans constantly engage in an information economy with their peers and the environment.[205,204,203] Social behaviour, however, does not necessarily lead to a process of unification. Economically, the major driver of innovation is instead competition: when different attitudes in a social group arise, the phenomenon of *schismogenesis* occurs. Literally meaning "creation of division", the concept was thought up by English anthropologist and semiotician Gregory Bateson to describe such competitive, adversary relationships. He defined it as a «*process of differentiation in the norms of individual behaviour resulting from cumulative intersection between individuals*».[17,18,19] Bateson's specific contribution was to suggest that certain conventional behaviours either inhibit or stimulate the schismogenic relationship in various ways. The consequence is a convoluted social arrangement in a continuous state of change. The analytic disentanglement of networks like this brought to the formulation of an extensive multidisciplinary research field concerned with regulatory and complex purposive systems called *cybernetics*. An essential characteristic in this panorama is the notion of *continuous feedback*: a process of circular causality where an observed outcome becomes the input for subsequent action in the ecosystem, supporting or disrupting particular conditions in order to re-establish balance. Two mathematicians provided interesting takes on cybernetics: Soviet Andrey N. Kolmogorov considered it as the study of systems of any nature, capable of receiving, storing, and processing information;[145] American Norbert Wiener instead matched it with the study of control and communication in these same systems.[305,306] As English psychiatrist Ross Ashby rigorously laid out recalling the origin of the term, cybernetics is «*the art of steersmanship*», or "governance", of a tangled structure maintaining its steadiness despite the constant adjustments of its components.[11] Looking at culture from a cybernetic perspective to study interactions and collective behaviour, we can appreciate its fresh new take on reductionism. Instead of seeking to untangle a complex system in terms of its constituents and individual interactions, cybernetic frameworks keep a system-wide perspective on behaviours and investigate how actions and relationships cause regularities to happen. Operatively, this outlook translates into more elaborate computational models that concurrently require an extensive amount of unit observations in order to be trained. With the intelligence embedded in a model we are hence able to figure out some features of a node by discovering its networks, scilicet its relations with other nodes and groups thereof.

A crucial consideration, when attempting to study the units of culture, is that we all partake in a multitude of distinct cultures at the same time, especially in today's hyper-connected, densely informatised reality that fosters individualism in a broad spectrum of areas. We are continuously influenced by the social aspects of the communities we join, and this makes it difficult to isolate the effects of a single group on the cultural milieu. Therefore, in my research I approached the problem in the opposite direction: starting from a known and predictable social setting like TED, I aimed at studying the personal and cultural nuances expressed by its umpteen speakers' utterances. The bet that language can convincingly offer the necessary material to derive the units of culture is indeed the base assumption of my thesis. After discussing and explaining culture, in the next sections I will investigate the role of natural language and the cultural insights we can computationally gather.

Language

Language is often regarded as an incredible tool of thought.[38] It shapes what we think and allows us to express feelings and share information. It is not, however, just a means to communicate and transfer intentions. It is our way of conveying what it means to be human. We indeed encode the experience of who we are in the words we use. Arguably, language is one of the most important human traits: all civilisations rest upon it. Every artefact we brought about – from the earliest rudimental hunting weapon to the latest smartphone – is an invention created thanks to language.[94] Nothing would indeed have existed without it or without benefitting from any knowledge attained by its use. Perhaps, it is also one of our best findings, the basis for all subsequent technology: the subtle device that enabled us to exchange knowledge and shape a myriad of cultures.[92] Questions about what language actually is, how it originated, how it works, and what is it for perplexed thinkers from Plato to Chomsky. Today, the focus is more on the ties binding language and our perception of reality.

Humanity's Greatest Invention

According to conventional wisdom, language originated with the *linguistic instinct* of Homo Sapiens somewhat 150.000 years ago. Conversely, American linguist Daniel Everett – drawing on evidence from a wide range of fields, including linguistics, archaeology, biology, anthropology, and neuroscience – assesses that our ancient ancestors, Homo Erectus, had already the biological and mental equipment for speech 1.500.000 years ago and that their cultural and technological achievements made it overwhelmingly likely that they had some kind of language.[93] It is proved that they knew how to sail, making them travellers and explorers. Further, the number of found colonies is more than a coincidence. Such voyages required imagination, cooperation, and planning. Homo Erectus was intelligent yet had the vocal apparatus of a gorilla.[93] They were capable of fewer sounds than we are, which was a constraint but not a limit. They also had a faster childhood development to put into place all sorts of cognitive mechanisms.[93] Nevertheless, they managed to build structured villages, corroborating the fact that they accomplished hierarchical thought, collective forethought, and decision-making.

It is outside of the scope of this dissertation to indulge in the evolutionary and ethnographic details of Everett's research, but we can take into consideration a crucial conjecture for the study of culture: since the dawn of humankind, language has been an essential device for both social and technical enhancement, encapsulating our experience, aiding mnemonic mechanisms, and enabling our evolution as a species. Like other tools, language was invented and has always been crucial in our lives. It shows significant variation across societies, is extensively diversified and can be reinvented, retained, or lost. It presents the bold and controversial notion that it is not an innate component of the brain but rather a cultural tool that changes and adapts through social groups. In Everett's pioneering investigation, it is argued that language is embedded within and inseparable from its specific culture.[92,93]

This insight led me to explore the individual involvement of people in a distributed community like TED's population of speakers. My idea was to study the culture of a small society built *in silico* in terms of linguistic participation, detecting the changes in its symbolic cultural landscape.

Sense-making

In essence, we can regard language as just symbols and a grammar. The American philosopher and mathematician Charles Sanders Peirce defined three kinds of signs.[212]

- **Indices:** signals that are physically connected to what they represent, like smoke is a vivid indication of the presence of fire, as our five senses evolved for us to be able to read them.
- **Icons:** signals that have no physical connection but some resemblance or allusion to a quality in particular, like the shape of a heart representing love, affection, and tenderness.
- **Symbols:** fundamentally abstract signs that are culturally connected to their meaning. Just like the number four (spelled in letters, numbers, or showed by quantity or gestures) points to a precise notion, the cross in Christianity started off as an icon and later on became the symbol of a religion and a structured set of beliefs, values, and traditions.

Humans have a core characteristic that has not changed in history, that is, the need for a story to understand the world around us. Narratives help us weave meanings in a context, giving them order

and direction. They create a network of symbols that can explain the complexity of our reality. When we think about something we need or want to express, we draw lines in our imagination, connect symbols, and with them, individual values and experiences. Language is hence much more than a protocol to transfer information from one's mind to another. While symbols and grammar provide the building blocks and the methodology to assemble a sentence, the final result is always much more than the sum of its parts. It is the product of our understanding and our need for interaction, cooperation, and knowledge sharing.

In the early stage of my research, I aimed to match linguistic expression and human behaviour, trying to glimpse into the distribution of symbolic universes in order to derive socio-linguistic clues. These are essentially the outcomes of cultural dynamics in action and emerge as patterns when we observe people acting in concert. Tracing these behaviours at scale, we can grasp a sense of people's culture around a certain topic, highlight their worldviews and systems of values, and segment their personality traits. Regularities in the choice of words, I assume, often reflect how people frame their world and make sense of it, revealing actionable indicators to interpret and describe these conventions.

I chose the archive of TED talks as cultural milieu because it is the perfect social arena for my investigation. It is a virtual, diversified knowledge market where speakers act and share their personal experience in the form of a monologue. In doing so, they reproduce and elaborate on their distinctive symbolic universes, each emerging as an individual contribution to the general cultural arrangement. Although TED is known for giving voice to social niches and minorities, speakers must adapt their utterances to the established format of the show. They are specifically taught how to address the audience, yet they can still express their ethos and personality. In light of this, it is essential to frame this type of analysis with ample domain knowledge. An accurate interpretation of the general intentions and contextual social norms is indeed extremely helpful in putting results into perspective.
As I alluded to in the previous section, culture emerges from human minds, and so does language. We shape symbolic universes to procure meanings for our experiences and motivate our feelings, thoughts, attitudes, and actions.

Semiotics

My research approach is much inspired by the *Semiotic Cultural Psychological Theory* (SCPT),[251] which conceives cultural dynamics as ongoing processes of sense-making. These acts of constant symbolic interpretation shape our perception and comprehension of reality. They are guided by generalised, affect-laden meanings embedded in the cultural milieu and work as basic intuitive assumptions concerning the world. According to Salvatore et al.: «*These intuitive assumptions channel lower generalised meanings, namely specific concepts and opinions concerning facts and objects of the social and physical world, values, beliefs, attitudes. SCPT adopts the term symbolic universes to denote such systems of assumptions*».[251] The framework highlights two main characteristics of these systems. First, they have an affective, pre-semantic valence. It means that they are used in socially suggested directions before being linguistically articulated and therefore rationally justified. Second, they encompass the entire field of experience rather than single parts of it. As they create their symbolic system, people are entirely embedded in it.[251] The dynamics of sense-making ineluctably make up the reality of the individual. They do not shape the world but the manner of experiencing it. They are guided and determined by the symbolic universe the individual identifies with, affecting both the outer and inner environments of the self.

The SCPT conceives people as members of particular communities whose cognitive processing is grounded and immersed culturally. From this perspective, the mind is not just individual but inherently transactional. «*It works in the interplay between the individual and the cultural milieu in terms of the communication dynamics substantiating social practices*».[250,253] People do not respond to their reality through invariant cognitive rules; instead, interpretation is channelled by generalised meanings embedded within the cultural arrangement of the population. Sense-making makes up the actual content of the experience and social identity of the individual. As the name suggests, the framework analyses the relation between the individual and their culture in semiotic terms, proposing «*a dynamic and performative view of meaning*»,[253] which is not a static entity but fluid and dynamic. The focus is on the interpretative activity from which meaning emerges, consistently with Peirce and Austrian linguist and philosopher Ludwig J. J. Wittgenstein,

who introduced the philosophical concept of *Sprachspiel* (language game) to argue that a word or a phrase has meaning depending on the rules of its context.[312] The analogy is to demonstrate that words have meaning depending on their uses in the varying activities of human life. In his posthumous *Philosophical Investigations*, Wittgenstein entirely rejected his previous theory of meaning while making one of his most powerful contributions to it. According to him, language gains its definitions from *how* it is used in specific cases: it is a game, and we learn its rules by actually playing.[8,312] The very notion of a universal definition is hence an artifice, a bit of subterfuge, stating that we cannot talk about what words really mean; we can only use them and observe their natural occurrence in the wild. There is no point in striving to interpret language for Wittgenstein.[8]
For example, if we can read a road sign, we instantly understand its meaning. It is an element of symbolic decoding, an affordance, a construal. We do not need to go deeper: the road sign is a social convention, and because of it, we read it as such. What is interesting, is that we can recognise an object to be a road sign — all thanks to culture, «*the immanent form of human phenomena*», «*the dynamic gestalt where human events come to life and develop*».[253] The same happens with words in ordinary conversations, when we rely on meanings that we know are shared with our interlocutors.
In Wittgenstein's theory of meaning, words are not defined by reference to the objects they designate, nor by mental association, but again by how they are conventionally used, contrasting with the descriptivist notions of sense and reference of German mathematician F. L. Gottlob Frege.[98]

In SCPT, the context is understood as an embedded system of generalised meanings of which the subject is part and works as a «*generative matrix of individual cognition*».[253] The dynamics of sense-making are ubiquitous and inevitable. Still, it does not mean that everything has to be considered a cultural phenomenon, as social and environmental conditions do exert their impact through and within the constraint of cultural mediation.

For Canadian psycholinguist Steven A. Pinker there is a level of fine-grained conceptual structure, which we automatically and unconsciously compute every time we utter a sentence, that governs our use of language.[217,218] These fundamental concepts, such as space, time, causation and intention determine the semantic construction of our sentences — reminiscent of the kinds of categories that German philosopher Immanuel Kant argued to be the basic configuration of human thought.[137]

Through language, we can speak of concrete and abstract things altogether using the same linguistic construction. For instance, we can say that we *gather* our ideas (as if they were material objects) and *put* them into words, which the listener *receives* and *unpacks* to *extract* their *content*. This kind of verbiage is usual and typical of many of our interactions. «*It is not the exception, but the rule*» observes Pinker.[217] It is very unlikely to find a language that does not rely on concrete metaphors to convey abstract conceptualisations. In our evolution as a social, knowledge-intensive species, we indeed managed to abstract our intellective repertoire of material concepts and apply it to new domains,[218] from religion, law, and literature to mathematics and economics. This perspective highlights a fundamental feature of human thought and the basis for our argumentations: people do not differ much on the facts as on how they ought to be construed, setting the attention on the words (and collective meanings) we use to articulate our thoughts.
For example: *ending a pregnancy* versus *killing a foetus*, *liberating Ukraine* versus *invading it*, or *redistributing wealth* versus *confiscating earnings*.

Language offers a window on the cognitive processes tWittge which we conceptualise the world and arrange interactions and social roles. As a matter of fact, on the report of American anthropologist Alan P. Fiske,[96] communication is a way of negotiating relationships and igniting cross-cultural variations. His *Complementarity Theory* states that human fitness and well-being depend on social coordination, characterised by reciprocal actions in conjunction with cultural paradigms and specific, highly structured, evolved proclivities.[51]

As a tool for social interplay, language has to satisfy two conditions: convey actual, intelligible content, and negotiate the kind of relationship with the interlocutor.[217] This happens explicitly, at a literal level, and implicitly, considering the context and other meaningful factors. A polite request, for instance, is a veiled, tactful imperative: we do not intend to exert dominance, and for this, the phrasing is measured, thoughtful, and affable.
In the big picture, our utterances reveal much of us, our culture, personality, and intentions.

Symbolic Universes

In the course of the previous sections, I often considered both culture and language in evolutionary terms. In this regard, American cognitive scientist Daniel C. Dennet identified three essential criteria for evolution to operate: variation, heredity, and differential survival.[75] Culture has great variation, is inherited across generations, and comprises alternatives that compete and differentially survive (e.g. languages, professions, ideologies, habits and traditions). From an evolutionary perspective, American biologist Mark D. Pagel argued that our habitual cooperation is a paradoxical feature that clashes with the notion of natural selection.[208] He deemed phenomena like language and religion to be the tools to reinforce cooperation through a shared sense of cultural identity and emphasised that *social learning* is the seed of cultural evolution and our intelligence.[208] Thanks to our ability to learn by comparison with others, we build on the wisdom and experience of our peers, our ideas accumulate, and our technology progresses.

As language evolves to enhance the benefits of cooperation and coordination, so do the symbols we use. For Austrian sociologists Peter L. Berger and Thomas Luckmann, reality is socially constructed, and language plays a pivotal role in shaping it and transcending its meanings.[28] They were influenced by the symbolic interactionism of American sociologist George H. Mead, who insisted that «*the individual mind can exist only in relation to other minds with shared meanings*».[188] In his perspective, reality is actively created by our interactions in and towards the world in dynamic processes that are somewhat redolent of the cybernetic framework. American sociologist Herbert Blumer, student of Mead, further proposed that meanings are derived from social interaction and altered through interpretation. It is hence from this interplay that common symbols are created, by approving, arranging, and redefining them.[31]

In *The Social Construction of Reality*, authors Berger and Luckmann introduced the notion of *symbolic universe*, a set of beliefs known and accepted by every member of a social group, providing legitimation and structure to the organisation.[28] In this section, I will deepen the concepts of symbolic universe and semiotic force, and present the results I gathered from the preliminary analysis of my dataset of TED talks.

In their ample study of the current sociopolitical dynamics of the future of European societies, *Symbolic Universes in Time of (Post) Crisis,* editors S. Salvatore, V. Fini , T. Mannarini, J. Valsiner, and G. A. Veltri brought together an innovative approach to politics and policy-making on the recognition of the prominent role of culture. In their quest to untangle the fundamental system of values at the core of the *Western Weltanschauung*, they developed the SCPT framework on the view of the human being as *homo semioticus*, «*a subject constantly engaged with the need to make meaningful ordinary experiences, as well as participation in society*».[253] In this view, symbolic universes are inherently cultural and internalised by those exposed to them. They are not just collections of ideas but embodied systems of assumption that comprise and foster the subjective experience of the world. They channel the way of feeling, thinking, and making decisions.[51] Meaning is the enactment of desire; conversely, desire is the way of providing subjective substance to the interpretation of experience.[249] Succinctly, a symbolic universe is something one *is*, not something one *has*.[253] It envelopes the entire experience of the individual, since affects, in their semiotic conceptualisation, are hyper-generalised, homogenising forms of sense-making: any individual feels, thinks, and acts from within and through their symbolic universe, and in doing so, they tend to reproduce it.[254] Thus, it is not changed by experience but the other way around.

Each symbolic universe is composed of an affect-laden, global, implicit, and only partially conscious worldview that operates as an embodied system of generalised assumptions. It channels and fosters the attitudes and behaviours of the individual, providing a consistent snapshot of how the world is, or ought to be, and their position towards it.[253] In SCPT, sense-making is not only symbolic; it links language, actions, and feelings, along with the formal level of cognition and the emotional level of experience. Its recursion allows us to recognise that culture is the root of psychological experience and social identity.[253] As meanings play a role in channelling individual and social cognition, the SCPT provides the frame to interpret the iterative pre-symbolic dynamics that underpin and fuel the emergence and stabilisation of such concepts.[253] In this view, the cultural milieu is, therefore, the inherent organisation of sense-making.

The authors of the study emphasised how today's complex scenarios cannot be puzzled out without considering people's subjectivity and bearing. However, it is crucial to stress that the cultural aspect of sense-making is constrained by contextual conditions (e.g. spatial, institutional, economic, demographic). Such variation leads interpretations in the light of a specific symbolic universe to distinct feelings, attitudes, and decisions with respect to the local context.[253]

As symbolic universes are hyper-generalised systems of meaning, they have no specific content and are expressed through social interpretations. Sense-making hence results from their encounter with the context-specific contents of social life. Whilst the subject matter of the representation reflects the contingencies of the context, the semantic structures underpinning them are similar across contexts, especially for topics involving high affective arousals.[252] In their work, Salvatore and his colleagues proposed a two-way interpretation: first, they focus on how culture affects the way people interpret the socio-political landscape and react accordingly; then, they complement the analysis with a hypothesis about the contextual conditions that have triggered and constrained the cultural dynamics.[253]

Their approach has been a great source of inspiration to frame my research. In particular, their view of culture is not presented in consensual terms but instead presumes a sense of *variability* inherent to the group. Each symbolic universe emerges as a particular interpretation of culture,[55] each one consisting in making certain basic cultural dimensions salient while de-emphasising others.[253] Symbolic universes are heterogeneously distributed, which means that individuals belonging to the same culture vary from each other due to their position in the cultural milieu. Culture is the "common ground" shared by a social group, yet it does not mean that members have the same feelings, ideas, and behaviours — these are manifestations that occur as a result of the arrangement of the symbolic universes. In this regard, culture is not a mere collection of self-contained symbolic universes but a network, with each of them defining its meaning by virtue of its similarities with the others. The authors hence shifted the focus of their cultural analysis from commonality to the traits that make people different from each other.

The SCPT conceives culture as the variability organisation of the individual trajectories of sense-making characterising a certain group,[253] scilicet the landscape that defines the movements of feeling, thought, and behaviour that is possible for a given society.

Language provides a helpful analogy to clarify this view. If a social group shares the same vernacular, it does not mean that its individuals produce the same statements; they will indeed articulate dissimilar assertions between which a relation can be found because they all originate from the same culture. Therefore, language is the set of shared rules that define the conditions and constraints of linguistic variability. In final analysis, language is a second-order form of sharing, as it generates and reflects the differences among those who share it. Every form of linguistic process can hence be seen as having a set of relations maintained over time through a constant variation of its elements.

This model of analysis merges two different aims: identifying the symbolic universes that are active in the cultural milieu and understanding them in terms of lines of semiotic force. The study of cultural grounds is not confined to the content description – that is, the emotional and cognitive output of the individual sense-making channelled by the cultural milieu – but is interested in modelling the dynamics that bring about these outputs, namely «*the latent organisation of culture that underpins the salience of symbolic universes, or what enables them to work as semiotic attractors*».[253] The ambition of the SCPT is indeed to move the cultural analysis from a descriptive to an explicative form of knowledge. Modelling culture as a semiotic field opens our understanding on two complementary levels:[253]

- the analysis of the inherent organisation of the semiotic field as a whole, tracing the lines that 1) work as the underpinning latent dimensions of sense, 2) foster the salience of symbolic universes, and 3) constrain the individual trajectories of sense-making;

- the identification of symbolic universes, each consisting of a set of generalised meanings substantiating a particular worldview: an implicit conception in which the relation of the individual with the world is interpreted and experienced.

Developmental psychologists also claimed that contextual factors related to the cultural specificity of a social group affect how people think and behave — as attested in the works of Estonian Jaan Valsiner, American Michael Cole, and Soviet Lev S. Vygotsky.[58,290,291,299] The processual view of culture that we can derive from such semiotic framing[248] emphasises once again the ongoing dynamics of sense-making and the role of generalised networks of meaning in framing people's minds.[321] Symbolic universes, however, are not just cognitive models: their embodied nature is consistent with the more general interpretation proposed by the *embodied cognition* perspective,[319] which gained momentum in cultural psychology with contributions from phenomenology and *enactivism*.[13,296] Accordingly, the embodied, affect-laden pre-semantic nature of symbolic universes takes into account their regulative function in terms of procedural habits concerning the preferences in the modes of activation and reaction.[29,97,248]

In the first phase of my research, I analysed the cultural milieu of TED speakers one topic at the time to highlight the aroused symbolic universes, in my case expressed as lexical patterns and semantic regularities. Given my expansive dataset of talk transcripts, I was interested in collating the density of semiotic connections evoked by a topic in two distinct decades (2002-12 and 2012-22) and describing the changes in perspective.

Symbolic universes are not directly observable: we can only detect them through abduction, id est analysing their effects in relation to patterns in the available data. Formulated and advanced by Charles S. Peirce, *abductive reasoning* consists in observing a phenomenon and drawing the simplest and most likely explanation, expecting an inexorable remnant of uncertainty. Instead of relying on surveys like in the case of Salvatore, Veltri, Ciavolino, Venuleo, and Mossi,[51,293,294,295] my intention was to build a semantic network from the linguistic insights provided by the transcripts.

To do so, I defined the computational logic to process each text, extracting and aggregating the lemmatised versions of nouns or verbs, then sorted in descending order of frequency (Notebook 5). Next, the algorithm computes the cosine similarity between the vector representations of the query word and the most common lemmas in the set. This determines an index of semantic closeness of each word to the chosen query, ranging from zero to one. Using the relative frequency of occurrence, I also compute the normalised weight of each word in the set. All these metrics are necessary to assemble a semantic graph depicting the semiotic lines of a symbolic universe (Figures 3.1, 3.2, 3.3). The query lies at the centre of the figure, whilst the other nodes represent the related concepts: the closer they are, the more semantically kindred their relationship. If the similarity score exceeds a defined threshold (0.65), an edge is drawn connecting the two nodes.

The result is a concise, interpretable map with some noticeable patterns mirroring the nuances of a symbolic universe. The network illustrates the linguistic regularities in the TED talks around a given argument. It fundamentally collects the most frequent lemmatised words that happen to co-occur with the query topic, associated and arranged in the space with regard to their semantic adjacency. The details regarding the word embeddings will be analysed in section Five.

Despite the inherent diversities, the network provides a snapshot of the query-related topics of discussion across the community, allowing us to advance a general explicatory framing of the cultural milieu. This technique is particularly useful to observe how general argumentation changes over time, as we can detect the incidence and diversity of interconnected topics.

Let us consider the word *creativity* as our query. Figure 1 shows the yearly frequency distribution of TED talks that have been tagged as creative. We can notice that the decade 2012-22 has more talks than 2002-22: this is a noteworthy factor to consider when comparing the two maps, since a larger corpus can affect the intrinsic cultural variability. In this regard, Figure 3.2 renders the differences between the two sets of concepts in terms of normalised frequency of occurrence. From the picture, we can infer that the talks associated with *creativity* in the decade 02-12 are distinctively driven by a sense of artisanal craftsmanship: there are conspicuous words that stand out, such as emotion, motivation, genius, spirit, mastery, desire, and creation.
Conversely, in the decade 12-22, we observe that far more attention is paid to ideas, intuitions, and innovation, showing greater emphasis on concepts like learning and skills. Nevertheless, the use of words like imagination, talent, sense, and patience remained unvaried.

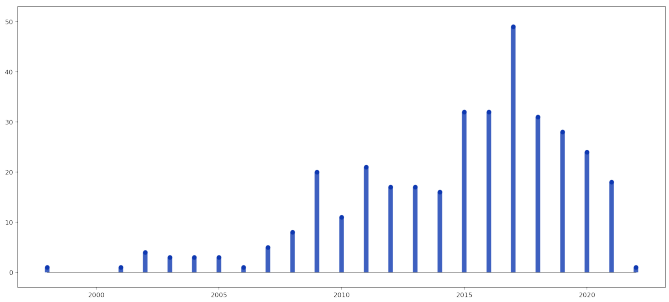
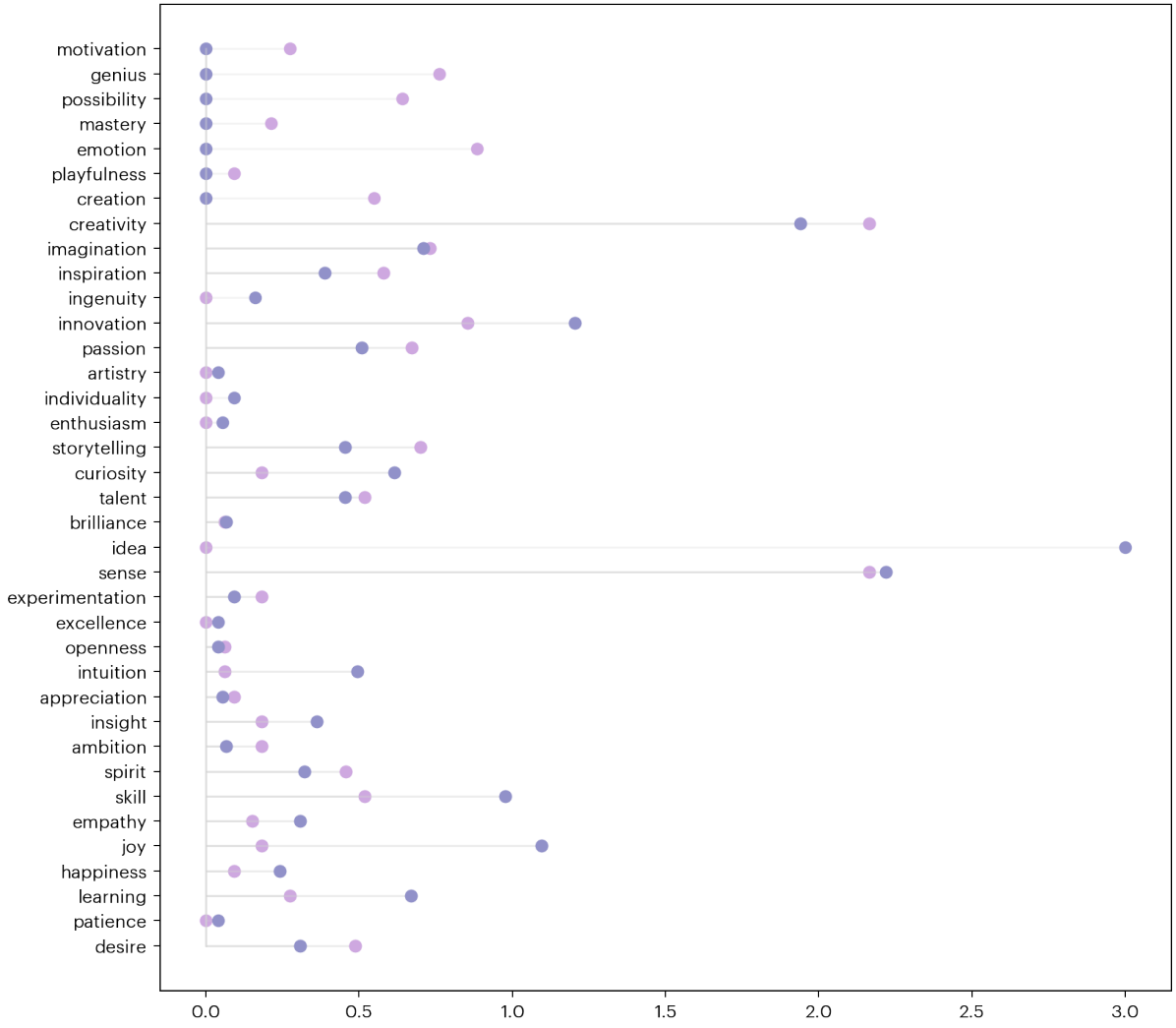


Figure 3.1, left. Frequency distribution of the talks tagged as creative, aggregated by year.

Figure 3.2, below. Stem plot representing the differences in normalised word count between the corpora of decades 02-12 (lighter purple) and 12-22 (darker purple). The metric reflects the prominence of a lemma in the se: the higher the value, the more relevant the concept.



The semantic graphs presented in Figure 3.2 give an additional reading of the cultural milieu of TED talks around the notion of *creativity* across the two distinct decades. The network of the decade 02-12 is scattered, with fewer nodes and edges than its counterpart. This is in part caused by the size of the original corpus, yet much of the talks centre around the concepts in the right-hand cluster (passion, sense, mind, emotion, desire), as hinted by the normalised density metrics. The cluster highlights some interesting connections with other relevant nodes in the graph (ability, understanding, knowledge). Decade 12-22 is instead congested with semantic interrelations among nodes, some of which are new, more empathetic and selfless.

In Figure 3.3 we observe the same operation but with *innovation* as query. The two images evince a substantial change in framing, which is eminently economical: the essential focus of the talks of the decade 02-12 is around investments, business, expertise, research, creativity, and engineering;

by contrast, those of the decade 12-22 are more concerned with entrepreneurship, growth, success, leadership, strategy, and productivity. In particular, looking at the normalised weights of the words in the second decade, we notice the emergence of now popular macro-trends in the field of innovation: sustainability, governance, development, and collaboration.

This preliminary experiment shows how we can derive symbolic universes from text, reinforcing my assumption that it is possible to extract general cultural insights from sheer utterances. In the following sections I will investigate the regularities in language and inquire into the computational aspects of natural language processing to procure refined units of culture.

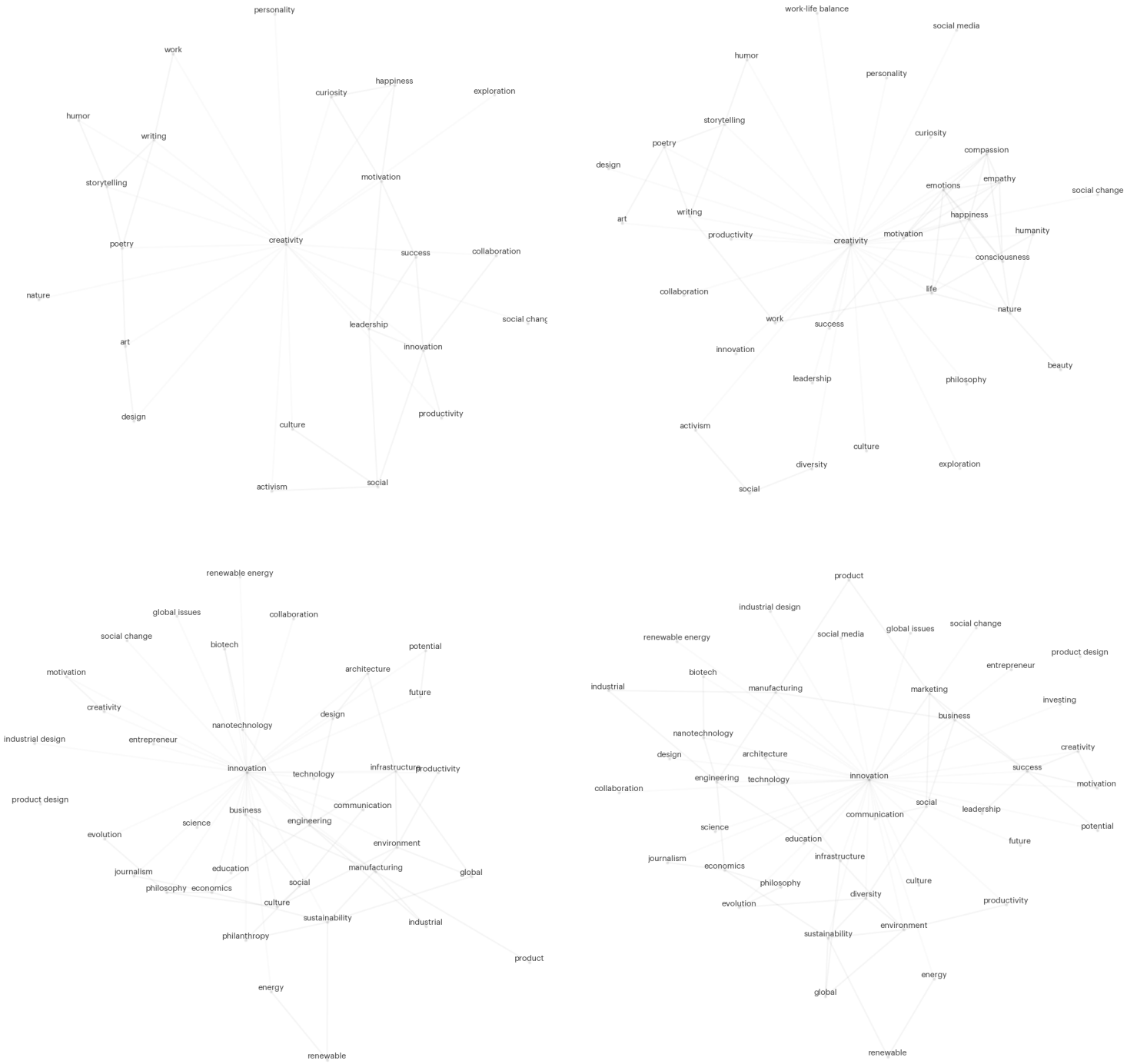


Figure 3.1. Semantic connections between tags occurring with the query, aroused from the words “creativity” (first row) and “innovation” (second row), showing decades 02-12 (left) and 12-22 (right).

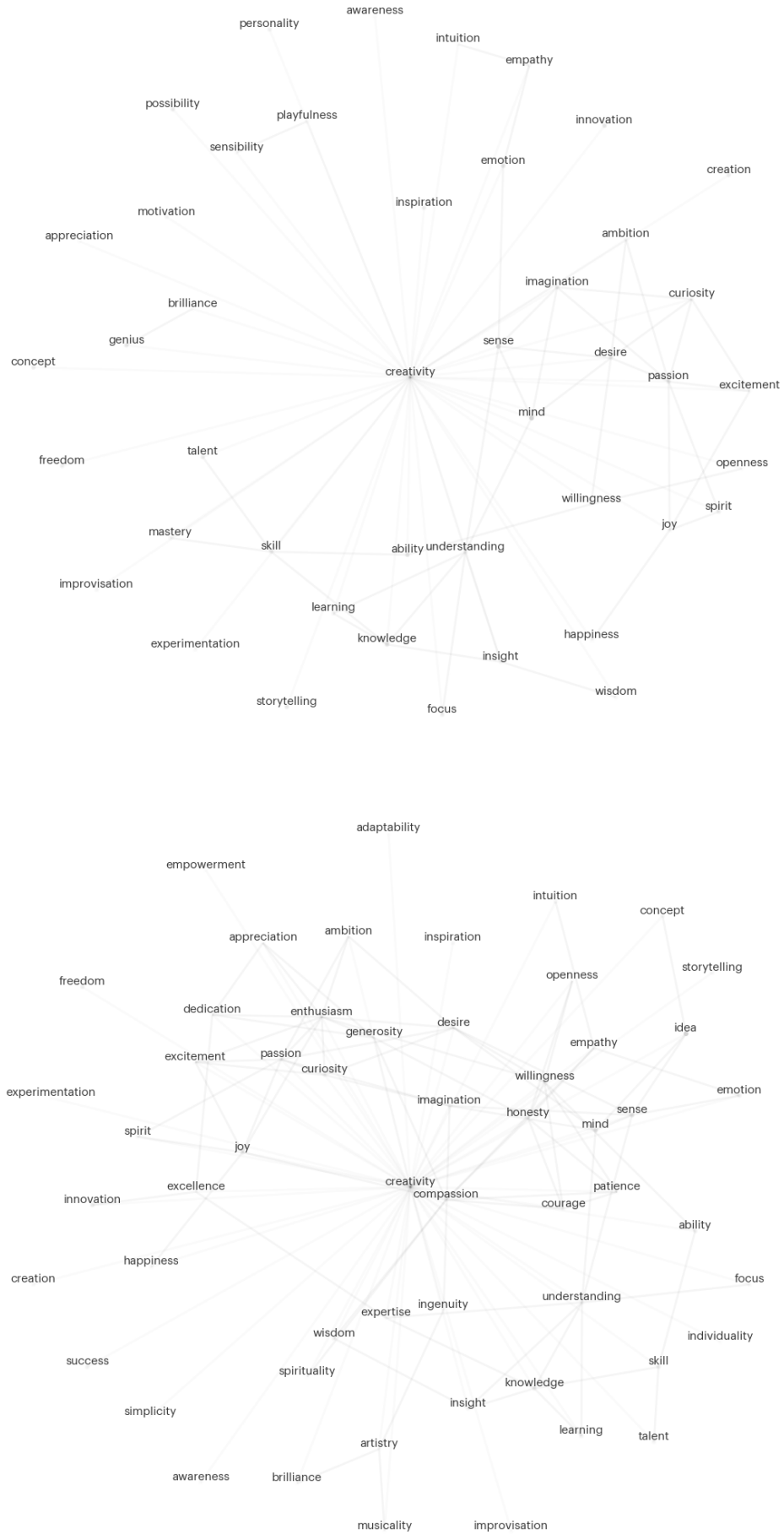


Figure 3.2. Semantic connections of the most frequent lemmatised nouns aroused from the word “creativity”. Top: decade 02-12 (92 talks). Bottom: decade 12-22 (248 talks).

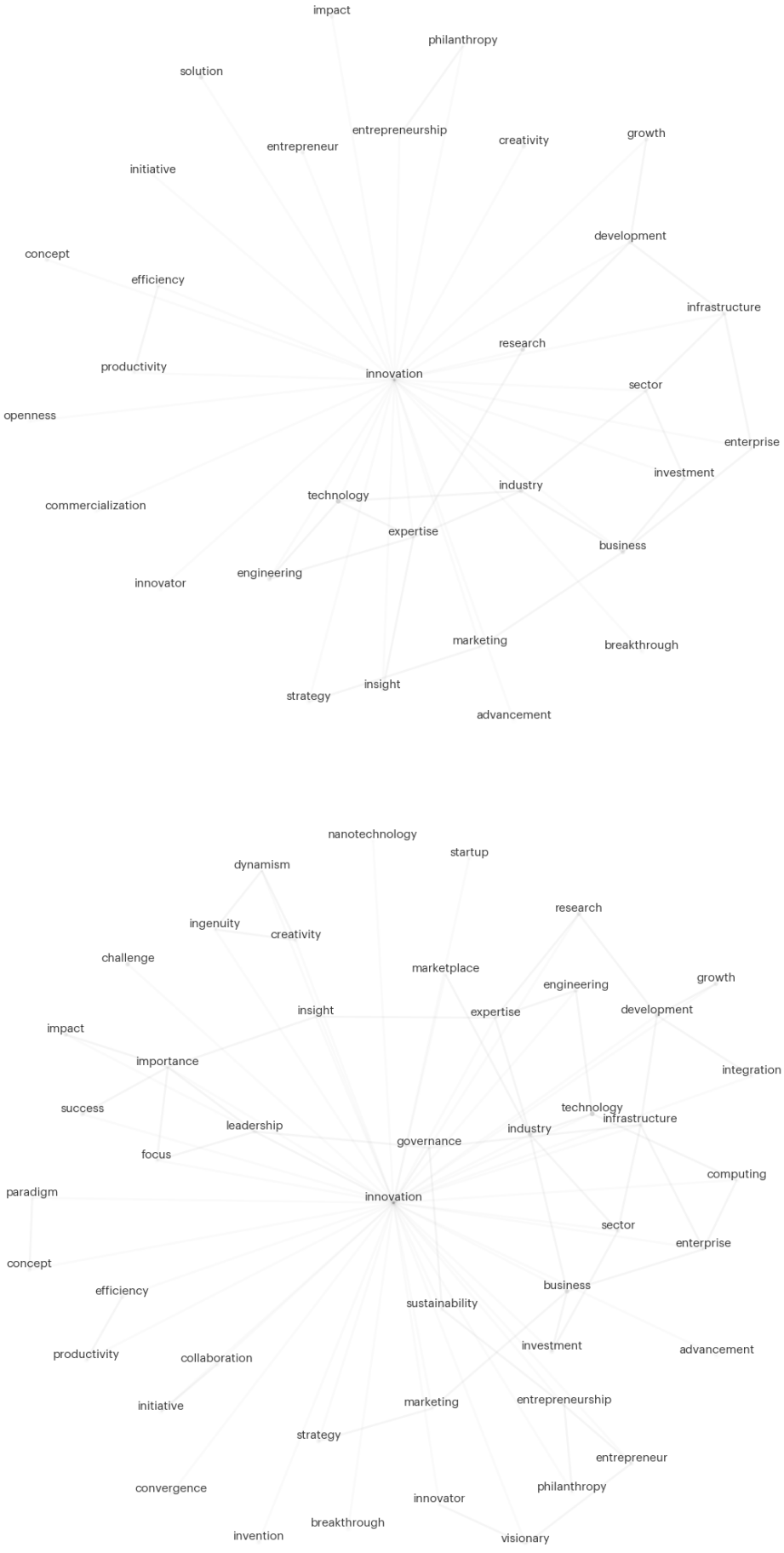


Figure 3.3. Semantic connections of the most frequent lemmatised nouns aroused from the word “innovation”. Top: decade 02-12 (66 talks). Bottom: decade 12-22 (376 talks).

The Economy of Words

Exploring the frequency distribution of the words in the entire collection of TED talks, a pattern is clearly distinguishable. It is not trivial: almost every time we rank word occurrences in a corpus of any language, we encounter the same regularity in the data. The frequency of a word is proportional to the multiplicative inverse of its rank. Hence, the second most used word will appear about half as often as the most used, and so on. The phenomenon (depicted in Figure 4.1) is known as *Zipf's Law* and fundamentally applies to any corpus of text. Formally, it predicts that, out of a population of N elements, the normalised frequency of the element of rank k, defined as $f(k;s,N)$, is equal to

$$\frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$$

where s is the value of the exponent characterising the distribution.

It is surprising how something as tangled as reality can be conveyed by something as creative as language in such a predictable way. Language is personal, intentional, and idiosyncratic. We are still far from understanding the underpinnings of such complex behaviour, but we are able to observe and model it. Zipf's Law has also been detected in city populations, protein sequences and immune receptors, earthquake magnitudes, ingredients per recipe, amount of traffic per website, academic citations, and chess moves. This empirical law was popularised by American linguist and philologist George Kingsley Zipf at Harvard University. He studied the statistical occurrences of words across different languages, proposing a discrete form of the continuous *Pareto distribution*, whose principle states that 20% of the causes are responsible for 80% of the outcome. In language, the most frequently 20% of words account for over 80% of word occurrences. Going further, Zipf conceived that the word frequency distribution is a consequence of the tendency for life and objects to follow the path of least resistance. In his trailblazing book from 1949, *Human Behaviour and The Principle of Least Effort*,[320] Zipf hypotheses that language is the optimised product of two main forces. Referring to the words of a text as an organised, sequential set of tools directed to the attainment of an objective, the author broaches the question of the economy of speech.

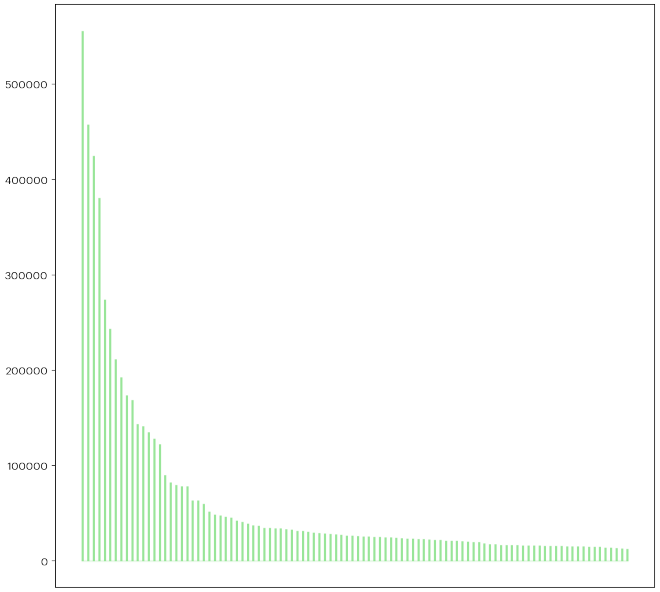


Figure 4.1. Frequency distribution of the hundred most common words in the TED corpus of talks (Notebooks 4 and 5).

All animals communicate, but it seems that only humans do it by means of language. We are the only species having elaborate, symbolic, grammatical systems. According to Zipf, if we focus upon the possible internal mechanics of language, we can hope to catch a glimpse of its inherent nature. In doing so, we have to consider two perspectives.

- The speaker is the one who has to select the meanings to convey and so the apt words for the task. There is a crucial latent economy in the lexicon they use: using fewer words spares the necessary effort to acquire and maintain an extensive vocabulary and select the words for their particular meanings.

- The listener is the one who has to disentangle the meaning of the words from the context, trying to understand what the speaker wants to say. Their interest is to rely on a vocabulary large enough to attain the word meanings better.

In essence, the speaker benefits from reducing the vocabulary size, whilst the listener from increasing it. The two attitudes are hence driven by two opposing drives: the *force of unification* and the *force of diversification*.[320] They determine the compromise, that is, the number of words in the vocabulary and the distribution of meanings across them. The speaker will therefore seek a balance between the economy of a small, wieldy vocabulary of more general reference on one side,

and the economy of a larger, more comprehensive one on the other. This theory is the closest attempt to unravel why language is so *zipfy*. To a certain extent, communication is deterministic: utterances and topics occur based on what was said and meant before. Zipf's Law appears to be built into our brains, describing how our thoughts ebb and flow and how meanings are distributed across the words we conventionally use. This happens via *preferential attachment processes*, which change according to how they have previously operated. Once a term is used, it is more likely to be used again soon until the topic changes.

Our inclination to minimise effort and the natural way in which a discussion follows preferential attachment processes are both responsible for the relationship between word rank and frequency.[320] For this reason, if a word has only been found once in the entire known collection of an ancient language, it is very difficult to come up with a meaning. Conversely, the abundance of textual corpora enabled the training of large-scale computational models to process natural languages.

The *principle of least effort* theorises that any human activity, including verbal communication, seeks to expend the least amount of effort to accomplish a task. That being so, language evolves as individuals simplify their speech in various ways. Abbreviations, for example. The reasons are manifold: from reducing the phonemes to articulate (*math* for *mathematics*), to removing irregular morphological forms to remember (*showed* instead of *shown*), language constantly undergoes a process of simplification and adaptation to the social environment, affecting words and the meanings we conventionally associate with them. These changes, writes German linguist Florian Coulmas, are generally «*utilitarian and economic in nature*».[61] From this perspective, the principle of least effort (PLE) provides an adequate explanation for many isolated changes (like *God be with you* becoming *good-bye*), and plays a role in most systemic variations, like the loss of inflections in English.[195]

Theoretically, the PLE governs the behaviour of both the individual and the collective group. Zipf first addressed the concept of minimised work through the study of words and their meanings from the viewpoints of both the speaker and the listener. He hence presented the concept of the *Economy of Words*, which is controlled by two contrasting forces – those of unification and diversification – that shape the vocabulary of a language. As people talk for a reason, speech can hence be likened to a set of tools that are engaged in achieving objectives.[76] It is indeed invariably directed to the attainment of certain purposes, yet directed enough to be considered as a tool, or a means to an end. Using this analogy, Zipf presented the case of *formal semantic balance* by plotting quantitative data on the frequency of word occurrence (Figure 4) and of the meanings in a stream of speech for the sake of showing the organised arrangement of the phenomenon.

If we concentrate our attention on the possible internal economies of language, we may hope to catch a glimpse of their inherent nature. He established beyond doubt the orderliness in human speech[76] and pointed out two consistent tendencies:

- The *Law of Abbreviation*, the direction of reducing the magnitudes of the speech entities by correlating the entities of smaller size with the classes of more frequent occurrence.

- The *Law of Diminishing Returns*, the direction of minimising the number of activities performed, according to the three economical principles of versatility, permutation, and specialisation.

These two proclivities maintain a formal-semantic, organic balance.[76] On the report of Zipf, we may expect to find that more frequent words tend to be shorter, more nuanced and more versatile.

In his study of natural language and human ecology, Zipf investigates the notion of symbolic process and culture. Emphasising the role of language in communicating social status and negotiating the social norms and conventions, he proposes a perspective of culture as «*a unit system of social signals and correlated social responses*». In this view, a social group acts like an individual's sensory system, establishing discrete objective criteria for the classes of action of its members.[320] At the same time, through the sheer act of living, everyone signals information about themselves and their intents. Cultural correlations of society are actually sundry and far-reaching: their ramifications condition human behaviour in any social group, setting a code of meanings and conventions, in many cases reflected by language.

Semantic Embeddings

Shortly after the publication of Zipf's pioneering work, the seminal paper *Computing Machinery and Intelligence*[289] was brought out, written by English theoretical biologist, mathematician, cryptanalyst, and computer scientist Alan M. Turing. It was the first to contemplate whether a machine can feature a notion of intelligence comparable to that of humans. Through the so-called *imitation game*, Turing's interest was to determine the ability of a computer program to impersonate a human in a real-time conversation with another individual so well that it is impossible to differentiate the program from a real human. An unconventional stance, since his focus was not on a machine that could think but rather on one that could *act like* a thinker. The resonance of the experiment and his refined theory of computation set the foundations of what we know today as Artificial Intelligence (AI),[158,176] the research field of study originally concerned with mimicking human cognitive skills, now centred around intelligent agents that demonstrate capabilities like automated reasoning, knowledge representation, and language processing.[202,221,241] Considering that the prominent purpose of my dissertation is to derive cultural insights from natural language, I will devote the next sections to covering my investigation in the subfield of AI called Natural Language Processing (NLP), intersecting mathematics, computer science, and linguistics.

Symbolic NLP

Just like Zipf aimed at uncovering the patterns in human language using statistics,[320] NLP seeks to understand it through these same regularities, acquiring knowledge from textual sources and applying it for classification, information retrieval, question answering, and machine translation.[176,241] From the 1950s to the early 90s, NLP has been entirely symbolic: the computer had to apply a given collection of rules to emulate natural language understanding (NLU) or other tasks through pattern matching.[303] The 70s were characterised by conceptual ontologies intended to structure real-world information into computer-processable data, creating the first interactive agents.[56,57,63,64,257] During the 80s, the research in NLP was chiefly focused on rule-based parsing by means of a *head-driven phrase structure grammar* (HPSG). Developed by the American linguists Pollard and Sag, a HPSG is a highly lexicalised, constraint-based formal grammar commonly used for knowledge representation in computational linguistics.[220] Other salient areas of NLP research included two-level morphology (comprising the lexical combination of words with roots and affixes, and their actual realisation in the corpus),[146] linguistic reference in light of the centring theory,[109,131] and rhetorical structure for text generation and summarisation.[182,283]
In the semantics branch, American computer scientist Michael E. Lesk introduced the eponymous algorithm for *automatic word sense disambiguation* (WSD) using machine-readable dictionaries.[160] Based on the assumption that words in a given section of text are likely to share a common topic, the algorithm compares the dictionary definition of an ambiguous word with the terms contained in its neighbourhood. «*To tell a pine cone from an ice cream cone*», the algorithm considers the words that constitute the definitions of the two objects and chooses the sense that has the larger number of contextual words.[160] Despite its glaring limitations, the Lesk algorithm propelled several extensions and variants for WSD over the years.[140,181,200,279]

Statistical NLP and Linguistics Disputes

Up to the 80s, NLP was essentially rule-based. The inception of ML techniques, marked by the steady increase in computational power resulting from Moore's Law,[198] considerably disrupted the panorama of the time with the introduction of algorithms able to learn patterns in the data and automate the procurement of *if-then* rules that until then were only defined by hand. The focus of the NLP research consequently shifted to the refinement of statistical models yielding soft, probabilistic decisions based on the assignment of real-valued weights to the features of the input data. These advancements concurred with the progressive surpassing of the Chomskyan stance of the *transformational-generative grammar* (TGG) that dissuaded corpus linguistics — that is, the study of language articulated in textual corpora collected in a natural context,[270] which are an essential component for the training of linguistic models. While NLP researchers focused on the systematic analysis of typical phenomena occurring in real-world data, the linguistic examination encouraged by American cognitive scientist Noam Chomsky was centred on investigating the corner cases that stress the limits

of the theoretical models in thought experiments. Regardless of this contrast, his elaboration of *Syntactic structures* represents an influential milestone in linguistics, arguing the independence of syntax from semantics,[45,47] considered by American linguist Charles Voegelin «*a Copernican revolution within linguistics*».[298]

The generative grammar propounded by Chomsky presumes a biological take on the structuralist theories of linguistics[49,147] and hypothesises an innate structure of explicit rules whose application allows to articulate countless sentences, contrasting with the previous structural and functional models.[90] *Syntactic structures* marked the epoch[163] and had a startling impact,[12] winning the endorsements of distinguished British linguists John Lyons[180] and Robert H. Robins.[235]

The study influenced the psycholinguistic research as well: contrasting with the behaviourist model of American psychologist and social philosopher Burrhus F. Skinner[273] – which presented language acquisition in terms of conditioned responses to outside stimuli and reinforcement – Chomsky argued that language is created by humans using separate syntactic and semantic components inside the mind, being the generative grammar a coherent abstract description of the underlying psycholinguistic reality.[43] Specifically, Skinner's so-called *operants* and behavioural reinforcement could not account for people being able to speak and understand sentences that they never heard before.[43] Despite the general refutation that followed, verbal behaviouralism might be analogous to cultural evolution and operant conditioning. For Skinner, they are all cases of parallel processes of *selection by consequences*, exhibiting the three aforementioned criteria of replication, variation, and environmental interaction.[271]

Chomsky is furthermore credited for his theory of *universal grammar* (UG),[50] which postulates the existence of innate constraints on the grammar of any natural language. During the course of language acquisition, claims Chomsky, children adopt specific syntactic rules to conform to the UG.[44] Proficiency comes with the knowledge of which expressions are acceptable and which are not. He thus advanced the controversial argument of the *poverty of the stimulus*[46,267] to motivate that all languages conform to the same structural principles, however, we are unable to acquire every feature of the language we are exposed to. In other words, the context in which a word occurs does not provide complete semantic information, and it is up to our minds to generalise and define meanings — a problem closely related to Quine's *indeterminacy of translation*.[224]

Analysing the faculty of language with American evolutionary biologists Marc Hauser and William T. S. Fitch,[111] Chomsky – in the light of his concept of UG – emphasised that the computational mechanism of recursion has evolved solely in humans, highlighting our unique adaptation for language.[48] The lack of an upper bound on the possible grammatical sentences an individual can build is indeed explained as a consequence of recursion in natural languages, as argued by Chomsky, Pinker, and Jackendoff.[216,219]

This generally accepted idea has been recently challenged by Daniel Everett. He claimed that syntactical and semantical features (like recursion, embedded clauses, quantifiers, and colour terms) are not necessarily tied to the Chomskyan UG[50] or Hockett's design features of language,[121] but have cardinal cultural connotations.[91,259] For Everett, a UG is not impossible in principle, but there is not much evidence for it. Instead, he staunchly puts forward his perspective of culture playing a paramount role in structuring the way we talk and the topics we talk about.[187]

This dispute is fairly reminiscent of the *linguistics wars* that took place in the 60s and 70s. Chomsky and other generative grammarians backed their thesis that the meaning of a sentence derives from its syntax. Conversely, the generative semanticists Postal, Ross, Lakoff, and McCawley argued that syntax is derived from meaning.[151,152,185,222,238] Their research program developed out of the TGG and eventually stood in opposition to it, spawning the linguistic paradigm that we know today as *cognitive linguistics* (CL), which attempts to correlate language understanding with cognitive concepts such as memory, attention, and perception. It hence offered a scientific first principle direction for quantifying *states-of-mind* through NLP.[143] Instead of a collection of structural protocols governing composition, grammar is seen by CL as the rules for the linguistic arrangement that best serve the communication of human experiences.[62] These rules are derived from conventional observations that seek to understand the sub-context of language patterns.[65] Combining this grammaticalisation with the analysis of word occurrences in a sentence,

formed the foundational strategy of *computational linguistics* — the interdisciplinary field of research tackling the computational modelling of natural language (e.g. text processing, parsing, part-of-speech tagging, and machine translation).

The 90s and 00s saw the consolidation of these disciplines and the concurrent growth in both volume and variety of multilingual textual corpora for model training. Most NLP systems, however, remained contingent on datasets specifically developed for specific tasks. A great deal of effort went into the development of supervised ML techniques that proved to learn more effectively from limited amounts of data, in the so-called *statistical revolution* in computational linguistics.[130] Alternatively, the abundance of raw, unannotated corpora prompted the researchers to focus on semi-supervised and unsupervised algorithms.

Neural NLP

At the beginning of the 21st century, the established ML approach in NLP relied heavily on statistical inference to automatically learn linguistic regularities. These techniques have the advantage of expressing the relative probability of many different possible answers rather than a single one, inspiring the composition of larger, more articulated structures. Statistical models also yield more reliable results when integrated into more complex systems and are generally more robust to unfamiliar or incorrect inputs. Yet, they require elaborate feature engineering and fine-tuning to get satisfactory performances. The issue was solved by representation learning[24] and deep neural networks,[104,107] which achieved brilliant results across various NLP tasks while setting new, higher standards. Artificial neural techniques are today's established state-of-the-art solutions for higher-level tasks (e.g. classification, translation, summarisation, question answering, and text generation). The shift from statistics to neural networks entailed substantial changes in the design of NLP systems, often reckoned a distinct new paradigm.

Meanings as Vectors

In all likelihood, the crucial challenge in NLP consists in translating the nuanced semantic definition of a word into a mathematical representation, id est, a real-valued vector of arbitrary size. This entity should encode the subtle set of meanings of the word, which is flexible[144] and context-contingent. It should also be able to preserve the many socio-cultural interpretations and utter interdependencies. To refer to such representations, we use the term *word embeddings*.[134] As these vectors stand for the meanings of a word, those that are closer in the vector space are expected to be similar in meaning. The concept of similarity between word embeddings is conventionally measured in terms of *cosine distance*:

$$\cos(\theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A_i and B_i are components of vectors A and B respectively. This similarity metric is defined as the vectors' dot product, divided by the product of their lengths. It is not dependent on the magnitudes but only on the angle of the two vectors involved, belonging to the interval $[-1, +1]$, yet neatly bounded in $[0,1]$ to estimate the relatedness or semantic closeness of two word embeddings like in the case of the symbolic universes in section Three (Figures 3.1, 3.2, 3.3). The concept of word embedding can also be extended to multi-word terms or even entire documents, as I will discuss in the next sections.

Semantic embeddings are procured with a set of language modelling and feature learning techniques, including probabilistic models of word co-occurrences,[101] context representations,[157] dimensionality reduction,[165,166170] and neural networks.[193] They led to a significative boost in the performance of tasks like syntactic parsing[275] and have been used for knowledge representation[243] in *distributional semantics*, the quantitative methodological approach that aims to understand meaning in observed language.

The first iteration of semantic spaces consisted of the algebraic modelling of vectors of identifiers, often employed in information retrieval, content filtering, relevancy ranking, and data indexing.[246] Such vector space models (VSMs) always incur largely sparse embeddings,[247] which are "cursed" with very high dimensionality.[20,21] The application of matrix factorisation techniques like Singular Value Decomposition (SVD)[211] or Principle Component Analysis (PCA)[83,122] mitigated the problem and motivated the design of Latent Semantic Analysis (LSA),[136] a distributional semantics tool that analyses the relationships between documents and the terms they contain, producing a set of related concepts.

The technique assumes that words similar in meaning occur in analogous pieces of text; in the words of English linguist J. R. Firth, «*a word is characterised by the company it keeps*».[95]

Canadian computer scientist Yoshua Bengio and colleagues R. Ducharme, P. Vincent, and C. Jauvin advanced a neural probabilistic language model, tackling the joint probability function of sequences of words in a corpus.[25] To fight the curse of dimensionality, they proposed to simultaneously learn a distributed representation for each word, together with the probability function for the word frequencies that are expressed in terms of these representations.[26] The key is generalisation, which is obtained «*because a sequence of words that has never been seen before gets high probability if it is made of words that are similar [...] to words forming an already seen sentence*».[25]

A statistical language model can be represented by the conditional probability of the next word given all the previous ones, since

$$\hat{P}\left(w_1^N\right) = \prod_{n=1}^{N} \hat{P}\left(w_n \,|\, w_1^{(n-1)}\right)$$

where w_n is the nth word, and $w_i^j = (w_i, w_{i+1}, \ldots, w_j)$.

The idea of using a neural architecture to model high-dimensional discrete distributions was already found useful for learning joint probabilities of random variables (decomposed as a product of conditional probabilities).[22,23,27] However, the method had to be updated to share parameters across time and deal with left-to-right sequences of variable length,[86,116] learning several symbolic relations.[207] Bengio's work focused on learning a statistical model of the distribution of word sequences, rather than the role of words in a sentence, and pushed the idea to a larger scale. Each word is thus associated deterministically or probabilistically with a discrete class, signalling that they are similar in some respect.[26] Similarity is represented with learned distributed feature vectors, or more simply, word embeddings. Like in Latent Semantic Indexing[71] or information retrieval, these vectors are procured on the basis of their probability of co-occurring in the corpus.[260]

These advancements brought two different flavours of semantic embeddings: one in which words are expressed as static vectors learned from their statistical co-occurrences, and the other in which words are expressed as vectors of linguistic contexts in which they occur. Nevertheless, after the seminal work of Bengio, Hinton and colleagues,[197,199] state-of-the-art word embedding techniques rely today on neural solutions instead of probabilistic models.

Given a corpus of text, today it is reasonably easy to obtain word embeddings with tools like Gensim or FastText. However, research literature shows that we can make use of sizeable pre-trained models (PTMs) featuring universal language representations that are fruitful for downstream NLP tasks.[302] It is the case of `word2vec`:[190,193] an algorithm that employs a neural network to learn word associations from a large corpus, detect synonyms, and yield semantic embeddings. Published in 2013 by Czech computer scientist Tomáš Mikolov during his time at Google AI, his core idea was to reconstruct the linguistic context of words, producing a vector space of several hundred dimensions where words are positioned in a way that resembles their common conditions. The quality of these representations was measured in a word similarity task and compared to previously best-performing neural techniques. His team observed large improvements in accuracy at a much lower computational cost.[190] Moreover, the algorithm produces quality vectors that do not only reflect the similarity of the words that are close in meaning but provide multiple degrees of similarity.[194] Somewhat surprisingly, the information captured by `word2vec` embeddings goes beyond simple syntactic regularities,[189,192] enabling the user to avail themself of word offset techniques to perform simple algebraic operations with word meaning.[194]

A popular example:

```
vec("King") – vec("Man") + vec("Woman")
```

results in a vector that is closest to the vector representation of the word "Queen".

Different types of models were proposed to estimate continuous representations of words, including Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), yet neural networks showed that they perform significantly better than LSA, preserving linear regularities among words,[194,317] while LDA becomes computationally much more expensive on large datasets.[193,191] For the *Efficient Estimation of Word Representations in Vector Space*,[190] Mikolov and colleagues first advanced the use of the

Feed-forward Neural Net Language Model (NNLM) by Bengio, Ducharme, and Vincent,[25] noting that the density of the values in the projection layer made the computation laborious and demanding. To avoid such complexity and overcome the limitations of the NNLM, the team presented the Recurrent Neural Net Language Model (RNNLM). RNNs can efficiently represent more complex patterns than shallow neural networks and do not have a projection layer. The recurrence of the model allows for some kind of short-term memory, whereas information in the NNLM is represented by the hidden layer state, which gets updated based on the current input and the state of the hidden layer in the previous time step.[190]

To minimise computational burden, `word2vec` takes advantage of two new model architectures producing a distributed representation of words.

- **Continuous Bag-of-Words Model**
  Similar to the feedforward NNLM, without the non-linear hidden layer. The projection layer is shared for all words and not just the projection matrix. This way, words get projected into the same position and their vectors are averaged. The model predicts the current word from a window of surrounding context words. It is called "bag" because the order of words in the history does not influence the projection, as future words are also used.

- **Continuous Skip-gram Model**
  Similar to CBOW, but instead of predicting the current word based on the context, it tries to maximise the classification of a word based on another word in the same sentence. Each word is used as input to a log-linear classifier with a continuous projection layer, while the model is set to predict words within a certain range before and after the current word. If the range increases, the quality of the word vectors improves at the cost of growing computational complexity. On the assumption that distant words are usually less related to the current one than those close to it, word weights are inversely proportional to their distance. In other words, the model uses the current word to predict the surrounding window of context words, yielding more refined vectors for infrequent words.

It is still not very clear why `word2vec` embeddings are so successful. Yoav Goldberg and Omer Levy estimated that its objective function causes words occurring in similar contexts to be represented by similar vectors as per cosine similarity.[106] This is utterly congruous with Firth's distributional hypothesis,[95,164] but still reckoned a sort of black box. The two assessed that the superior performance of `word2vec` in downstream tasks is not a result of the models per se but instead the choice of specific hyperparameters.[167] Although it proved to be very successful in capturing fine-grained semantic and syntactic regularities using vector arithmetic, the origin of these regularities has remained opaque.

To analyse the necessary model properties for these traits to emerge in the word embeddings, Jeffrey Pennington, Richard Socher, and Christopher D. Manning introduced a new global log-bilinear regression model combining the advantages of global matrix factorisation and local context window methods. Published in 2014, a year after `word2vec`, GloVe efficiently leverages the statistical information of non-zero elements in a word co-occurrence matrix – instead of an entire sparse matrix or individual context windows – and outperformed related models on similarity tasks and named entity recognition.[213]

Pre-Trained Models

PTMs are essentially huge dictionaries of embeddings that have been previously computed on massive, diversified corpora and freely released to the public. Thanks to open-sourced PTMs, the user can do without training a new language model from scratch[288] and take advantage of the shared effort. They date back to the 10s and have been deemed a staple tool in NLP for years.[193] GloVe provides high-quality distributed vector representations for word types that are learned from co-occurrence statistics on extensive datasets of unlabelled texts. These embeddings capture a large number of precise syntactic and semantic word relationships, with two inherent limitations: their indifference to word order and their inability to represent idiomatic phrases.[193]

At the time, large-scale static PTMs marked an outstanding improvement in the generalisation of downstream models and allowed way better similarity operations than custom-trained models.[223,302] Static word embeddings are considerably lighter than language models, faster, and easy to deploy. Also, they do not require a demanding use of resources for training.

Text Categorisation

A large body of empirical work is now actively seeking to understand the quality and efficacy of the most recent language models, especially when attention-based architectures are involved.[191,192,193] Word-level contextual representations from these kinds of models are able to encode sentence structure across a range of syntactic, semantic, local, and long-range occurrences. It is observed that these state-of-the-art language models produce strong representations for conditional phenomena but only offer comparably small improvements on semantic tasks over a non-contextual baseline,[285] like in the experiment I am about to discuss in this section.

Contextualised word vectors are the last revolution in the field.[135,236,313] They are obtained using an encoder module, usually an LSTM[120] or a Transformer,[292] and are trained on massive corpora, requiring intensive computational power at a prohibitive cost for the user. In this section, I will analyse whether they are a true improvement over static, more conventional semantic vectors.

Problem setting

After analysing the latent symbolic universes that can emerge from language (section Two) and assessing the economy of words that is so typical of our linguistic behaviour (section Three), I used two different pre-trained language models to compute semantic embeddings of each TED talk in the dataset: a statistical one and a neural one.

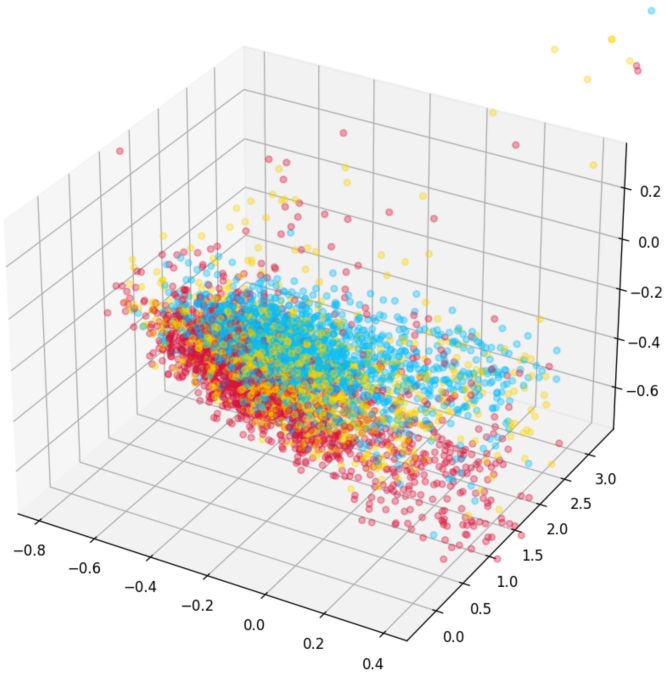


Figure 6.1. 3D spatial arrangement of the static document embeddings obtained with PCA. Colours represent the three categories: "science and innovation" (blue), "culture and society" (red), "economy and environment" (yellow).

My intention is hence to compare and examine the performances of the two models across different ML algorithms and draw some actionable considerations of the approaches they epitomise. After deriving a document-level semantic embedding of every transcript, the task is to categorise each talk based on its respectively annotated topic area. The first part of the experiment deals with essential baseline models and subsequently includes three custom neural architectures: a multi-layer perceptron (MLP), a feed-forward neural network (FFNN), and a convolutional neural network (CNN).

The objective of the first section of my experiment is organised in three phases:

1. Process each transcript in the dataset, obtain the word vectors, and compute a document embedding to encode the talk's content.
2. Use the embeddings to train a set of baseline models for multi-classification.
3. Design and implement a FFNN and a CNN for the same task, describe their architecture, and compare the results across all models.

NLP Pipeline and Document Embeddings

To process the raw text of a talk transcript and procure its linguistic features, I used SpaCy coupled with two distinct pre-trained language models: `en_core_web_lg` and `en_core_web_trf`. While both have the same processing workflow for multi-task learning, the two PTMs are very different in terms of how vectors are obtained. The first employs `tok2vec` to yield static embeddings from a pre-computed word vector table.[213] The second makes use of a pre-trained transformer model based on RoBERTa,[174] a large and powerful neural network that provides more accurate vectors after word-piece realignment (Notebook 1).

While word vector tables are only used as static features, the vectors produced by a transformer are *dynamic*, in the sense that the same word can have a different representation depending on the context where it occurs. For example, given two phrases – "*Apple pie*" and "*Apple stocks*" – the first model returns the same vector for the words "*Apple*", whereas the second returns two dissimilar vectors as the word refers to the fruit in the first phrase and to the company in the second.

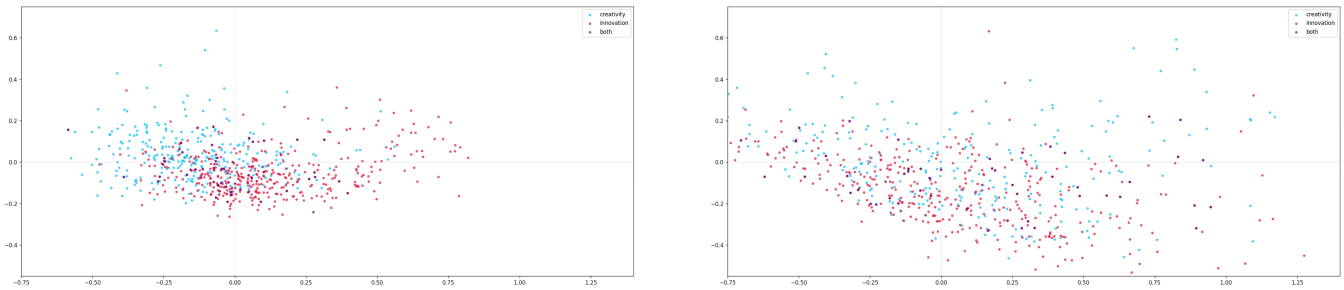


Figure 6.2. Static (left) and contextual (right) document embeddings in a 2D vector space obtained with PCA.
Blue dots indicate that the talk has the tag "creativity", red dots the tag "innovation", and the purple dots both tags. It is visually convincing how static embeddings retain a more defined spatial arrangement, while contextual vectors (although richer in conditional semantics) are more scattered and difficult to huddle. The code used to produce these plots is in Notebook 2 and Notebook 3.

The document embedding of an entire transcript is computed by averaging all the word vectors it contains (Notebook 1). The method is extremely simple, intuitive, and provides good performance in many applications. It requires no parameters and is straightforward to carry through. Essentially, we summarise the local information of a collection of word embeddings, capture the general sense of the whole text, and encode it in a single vector. The document embedding is nothing else than the average vector of the collection. Owing to the fact that there is no correspondence between the dimensions of the pre-trained word vectors, the validity of the approach may not seem obvious. Nevertheless, it is empirically demonstrated that averaged embeddings retain semantic information by preserving the relative distances between word vectors.[54] The operation, however, has the potential consequence of *diluting* individual meanings due to the high cosines between semantically unrelated words; scilicet, a document embedding is likely to give more resonance to more prominent vector directions.

Using the first principal components for plotting, it is possible to assess the positioning of all document embeddings in a compressed vector space with respect to the category to which they belong (as in Figures 6.1 and 6.2; Notebook 2).

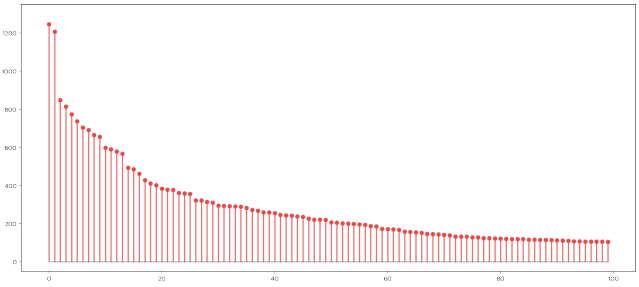


Figure 6.3. Frequency distribution of the original 347 tags.
Only the first hundred are shown.

Baseline Models

Once all document embeddings had been computed and assigned to a category, I arranged a diversified collection of classification algorithms from Scikit-Learn to gather some initial performance results as a baseline for both PTMs. The collection included nearest neighbours, linear models, support vector machines, decision trees, random forests, extreme gradient boosting, and a multi-layer perceptron. All hyper-parameters of the estimators were fine-tuned on a validation set to optimise the performance, and models with random initialisation were run several times to check the robustness of the results they yielded before assigning a random seed for consistency.

In every training situation, the static embeddings have shown greater accuracy than their context-enhanced alternative, along with better results in terms of precision and recall. Looking at the confusion matrices of the validation scores, it is clear that static vectors generalise better than contextual ones when dealing with document classification. Apparently, as "classic" word embeddings are the average result of multiple semantic occurrences, they retain a more restrained signal when averaged a second time to procure a document representation. Static embeddings are also more clustered if analysed in a dimensionality-reduced vector space, while the transformer-obtained counterparts are more scattered and difficult to categorise (Figure 6.2).

Nearest neighbours and ensemble methods have been pivotal in outlier detection (Notebook 6). Notably, pruning the outliers identified by the Isolation Forest (IF) algorithm[172,173] led to a significant increase in accuracy for the ensemble learners, regardless of the PTM. In every other training scenario, outlier detection did not achieve a performance improvement. An explanation can

be traced in their graphical spatial arrangement. By looking at the scatter plots, there are very few outliers that can actually impair training.

The best-performing baseline classification models are XGBoost, RidgeClassifier, LinearSVC, LDA, and SGD. The best estimator results are reported in the two following tables.

Linear models (Notebook 7) exceeded my expectations. In particular, the regularisation strategy of the RidgeClassifier normalised the data and then used SVD to compute the regression coefficients and achieved a good result. Because of the multi-classification setting, the estimator relied on three one-versus-all classifiers, taking advantage of the multi-variate response support in Ridge. Linear Discriminant Analysis also proved very effective without the need to remove any outlier or even tune any hyper-parameter. These results suggest that the three classes have varying covariances that are best traced by linear models. Despite class weighing, Support Vector Machines did not excel in this context (Notebook 8). SVMs are known for being one of the most robust prediction methods, flexible, generalisable, and extensible through kernel functions.[110] They do not get stuck at local minima, have few model parameters to select, final results are stable, reproducible, and largely independent of the optimiser.[35] Only LinearSVC achieved a relatively good result, behaving almost like LDA. The best model, however, is an ensemble learner: XGBoost,[37] a regularised gradient boosting framework that includes proportional leaf shrinking, automatic feature selection, and parallelised computing (Notebook 9). The model configuration is set for multi-class classification using the softmax objective, outputting the predicted probability of each data point belonging to each class. Following the principles of stochastic modelling,[119] tree-based ensemble methods emerged with the intent to arbitrarily increase complexity in order to improve accuracy. The intention is to build multiple predictors in randomly selected areas of the feature space:[225] they generalise the classification, and their combined effort can be monotonically improved. Optimising an ensemble algorithm on a suitable differentiable cost function generally leads to a significant decrease in generalisation error,[33] which is indeed the core of gradient boosting: use a set of weak learning methods to create a single, robust one.[138]

Model	Accuracy
Nearest Centroid Classifier	67.98% (with IF)
K-Neighbours Classifier	72.80% (with IF)
Logistic Regression	74.08%
Ridge Classifier	75.63%
Stochastic Gradient Descent Classifier	74.60% (with LOF)
Linear Discriminant Analysis	75.35%
Quadratic Discriminant Analysis	72.38% (with LOF)
Linear Support Vector Classification	75.39% (with LOF)
C-Support Vector Classifier	72.58% (with IF)
Nu-Support Vector Classification	73.52%
eXtreme Gradient Boosting Classifier	**75.65% (with IF)**
Decision Trees Classifier	63.37% (with IF)
Random Forest Classifier	74.56% (with IF)

Table 6.1. Results of the static vectors on the baseline models. Legend of the outlier detection: IF = Isolation Forest Classifier; LOF = Local Outlier Factor Classifier;

Model	Accuracy
Nearest Centroid Classifier	43.16% (with IF)
K-Neighbours Classifier	53.87%
Logistic Regression	69.25%
Ridge Classifier	71.50%
Stochastic Gradient Descent Classifier	68.28%
Linear Discriminant Analysis	69.67%
Quadratic Discriminant Analysis	55.27% (with IF)
Linear Support Vector Classification	68.54%
C-Support Vector Classifier	38.36%
Nu-Support Vector Classification	69.53%
eXtreme Gradient Boosting Classifier	65.83% (with IF)
Decision Trees Classifier	45.83%
Random Forest Classifier	62.76%

Table 6.2. Results of the static vectors on the baseline models.

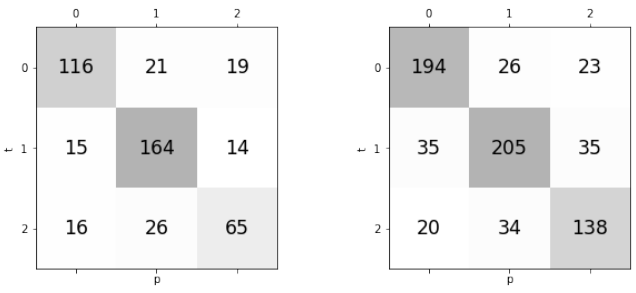


Figure 6.4. Confusion matrices of the results obtained from XGBoost (left) and Ridge Classifier (right).

Neural Networks

Further analysing the performance of the pre-trained static and contextual word vectors for document classification, I adopted deeper learning solutions, starting with a Multi-Layer Perceptron.[240] MLPs are the simplest design for an Artificial Neural Network, consisting of at least three layers of neuron-like processing units: the input and output layers, with at least a "hidden" one in between that collects internal representations of the training data. They map similar input patterns to similar output patterns, a characteristic that allows them to make reasonable generalisations and increase performance results.[115] These internal representations are constantly weighted in the supervised learning phase, and the direction of such update is pointed out using gradient descent procedures.[159] From a random initialisation, at every step, each connection computes the derivative, with respect to its strength, of a global measure of the error in the network, eventually reaching convergence through the process of backpropagation.[102] The fine-tuned structure of my MLP (Notebook 10) comprises a hidden layer of one hundred neurons, the rectified linear unit as activation function, the default L2 regularisation term, and a standard adaptive learning rate that is kept constant as long as training loss keeps decreasing. Results do not differ from previous ones: static word vectors proved once again to be superior in performance when dealing with document classification. While the MLP achieved 75.21% accuracy using classic embeddings, contextual vectors only scored 65.86%. It is noteworthy, however, that the MLP beats XGBoost in terms of precision and recall: the first reached 75.26% and 74.28% against the 74.32% and 73.36% of the latter. This was a good incentive to experiment with deeper neural architectures.

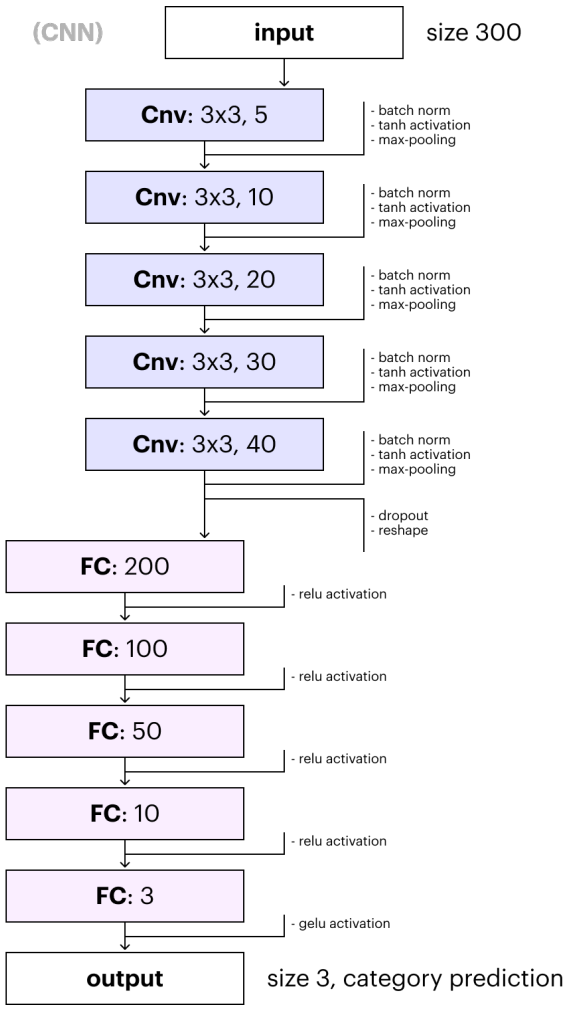
Therefore, I moved to Convolutional Neural Networks (CNNs), a class of regularised versions of MLPs. Based on the shared-weight architecture of the convolution kernels, they learn filters that slide along the input vector and provide translation-equivariant responses known as feature maps. In the past, pattern classification algorithms were mainly based on linear mappings. Later on, Feed-Forward Neural Networks (FFNNs) overcame the non-linear mapping tasks, such as the XOR problem, with the assumption that deeper architectures lead to increased adaptability in signal processing.[304] (Notebook 11).

It ultimately brought to the idea of the CNN: stack a series of processing modules that are able to optically learn the patterns in the input on top of a FFNN.[316] Years ago, CNNs were considered one of the most powerful tools to detect position-agnostic regularities in huge amounts of data.[1] Inspired by biological vision processes, CNNs introduced layers that convolve the input, abstracting it to a feature map. The network can hence recognise stimulus patterns based on their geometrical similarity without being affected by their position.[99] Similar to the biological neuronal response in the visual cortex, each convolutional neuron processes data only for its receptive field. This technique was successful in a wide range of fields, from CV to NLP.[59]

My implementation of a CNN (Notebook 12) has five mono-dimensional 3x3 convolutional layers followed by batch normalisation, each one activated with a hyperbolic tangent function and fed to a max-pooling layer to reduce resolution and, therefore, complexity. The output of this first part is then passed to a regularisation layer that applies a low-probability dropout to alleviate overfitting, and it is finally processed by five fully-connected layers that are activated by gaussian error linear units. Batch normalisation makes the CNN faster and more stable while easing the parameter updates, allowing a higher learning rate.[127] Re-centering and re-scaling the signal is reported to be useful to mitigate the problem of internal covariate shift and makes the optimisation landscape significantly smoother, inducing a more predictive and stable behaviour of the gradients.[256] Dropout, on the other hand, is an inexpensive approximation of the bootstrap aggregation and prevents complex co-adaptations of feature detectors by randomly omitting some of them.[117] The network is trained using the cross-entropy loss, the standard criterion for multi-classification and equivalent to the combination of the log-softmax and the negative log-likelihood loss. When instantiated, it is provided with a vector of class weights that have been obtained to balance the computation of the loss for each category. Instead of regular Adam,[142] which I used for the MLP, my optimiser of choice for the CNN is AdamW,[175] a decoupled weight decay regularisation approach for the adaptive gradient algorithm that improves generalisation performance and significantly reduces the number of required epochs to reach good outcomes.

After an ample set of trials, the model struggles to properly generalise, and results are poor if compared with the baseline models. Training loss is fairly easy to reduce, however, both classic and modern embeddings provide no substantial grasp for the convolutions to yield good accuracy. It is not necessarily a surprise: in experiments like this one, word vectors give their best when matched with techniques that take into account and value their spatial information, as previously shown by ensemble methods and linear models. Static vectors have a bigger advantage because of their statistically-synthesised arrangement in the vector space, being document representations just better centroids compared to their transformer-based counterparts. Hereby, my custom CNN does not exceed ~73% accuracy with static embeddings and ~64% with their contextualised alternative.

In the next section, I will introduce the transformer architecture and the revolutionary new take it brought in natural language modelling, testing some of its offshoots in this same experimental setting.



The Transformer Architecture

The deep learning techniques I covered so far were limited because of the way the text was encoded as their input. In this section of my dissertation, I no longer avail myself of PTMs to derive word vectors, but rather look at the raw text as a chain of inputs. Recurrent Neural Networks have been a popular and powerful general-purpose sequence learning architecture for probabilistic transduction modelling.[108,240] They put forth the combination of high-dimensional multivariate internal states and non-linear state-to-state dynamics to offer more expressive prediction power. Intrinsically, RNNs introduced the concept of *memory* in sequential learning — later on enhanced and developed by LSTMs,[120] where an artificial neural cell can efficiently adjust the flow of sequential information with the ability to retain in its internal state the valuable and forget the unnecessary.[100,129] These architectures have been firmly established as leading-edge approaches for language modelling and machine translation.[226] Further efforts pushed the boundaries of these architectures, assuming an Encoder-Decoder structure to learn more semantically and syntactically meaningful representations of linguistic phrases.[42] Such a singular architecture is composed of two RNNs: one that encodes the input sequence into a fixed-length vector and the other that decodes the intermediate representation into the output. The two are jointly trained to maximise the conditional probability of the target given a source while adaptively remembering and forgetting. As RNNs factor computation along with the symbol positions of the input, a sequence of hidden states is generated, precluding parallelisation within training samples. This becomes critical at longer sequence lengths, as memory constraints limit batching. Factorisation tricks[149] and conditional computation[268] improved efficiency and performance, however, the fundamental constraint persists. Assuming that the in-between fixed-length vector known as the “code” is just a bottleneck that hampers the whole system, the Encoder-Decoder architecture has been extended

Figure 6.5. A graphical abstraction of one of the many module combinations I tested to improve the performance of the CNN. Batch normalisation and dropout helped preventing overfitting, but it is hard for the model to grasp actionable information from a document embedding. Filters larger than 3x3 were vain, while increasing the number of channels (and so, requiring deeper matrix computations) made the training more steady.

to allow an automatic soft-search for the entities in the input sequence that are relevant to the prediction.[14] Taking inspiration from human cognition once again, attention-like mechanisms were introduced in the 1990s and have been recently adapted to artificial neural networks to mimic the cognitive process of selectively devoting more focus to a discrete aspect of information, ignoring the rest. They make use of soft weights that are learned during training, and this flexibility enables outstanding categorical inference and rich structural dependencies,[141] making self-attention a predominant technique in almost every deep learning application.

Attention is the key

The result of a systematic investigation that involved eight researchers of the Google Brain Team, *Attention Is All You Need*[292] was presented in 2017, marking a breakthrough for sequence mappings. Deemed by many a milestone in NLP and beyond, the paper captures a snapshot of the cutting-edge techniques of the time and provides a succinct yet exhaustive analysis of the established recurrent approaches for sequence modelling and transduction problems, unfolding their flairs and shortcomings. Clearly, the essence of the study is the monograph on self-attention and its application in the Transformer, a novel model architecture «*eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output*».[292] The very title of the paper reflects this stalwart confidence that attention is enough to compute high-quality semantic representations dispensing with convolutions and sequence-aligned recursions entirely. A confidence that is backed by cogent experiments on two machine translation tasks that highlight how the Transformer is superior in quality while being more parallelisable, hence requiring significantly less time to train.[292] It scored 28.4 on the English-to-German translation task of the Bilingual Evaluation Understudy (BLEU),[209] outperforming the previous models by two points, and reaching 41.8 on the English-to-French translation task. This is achieved after a training session of 3.5 days on eight GPUs, which can be exorbitant for many practitioners but at the same time only a small fraction of the training costs of the best models in literature.[292] The authors noted that the Transformer generalises well to other tasks, parsing both large and limited training data. In the paper, they dissected the architecture of the Transformer and emphasised the contribution of its components: their reporting is thorough and compelling, a fluent reading, solid, and comprehensive in its organisation, whereas stiff for the less experienced reader. Although the exposition provides enough information to appreciate and validate the design choices, some aspects are so briefly described that it is difficult to re-implement the system without additional documentation.

I reckon that the paper is more of a framework than a blueprint. It aims at setting the general principles of the transformer class of models,

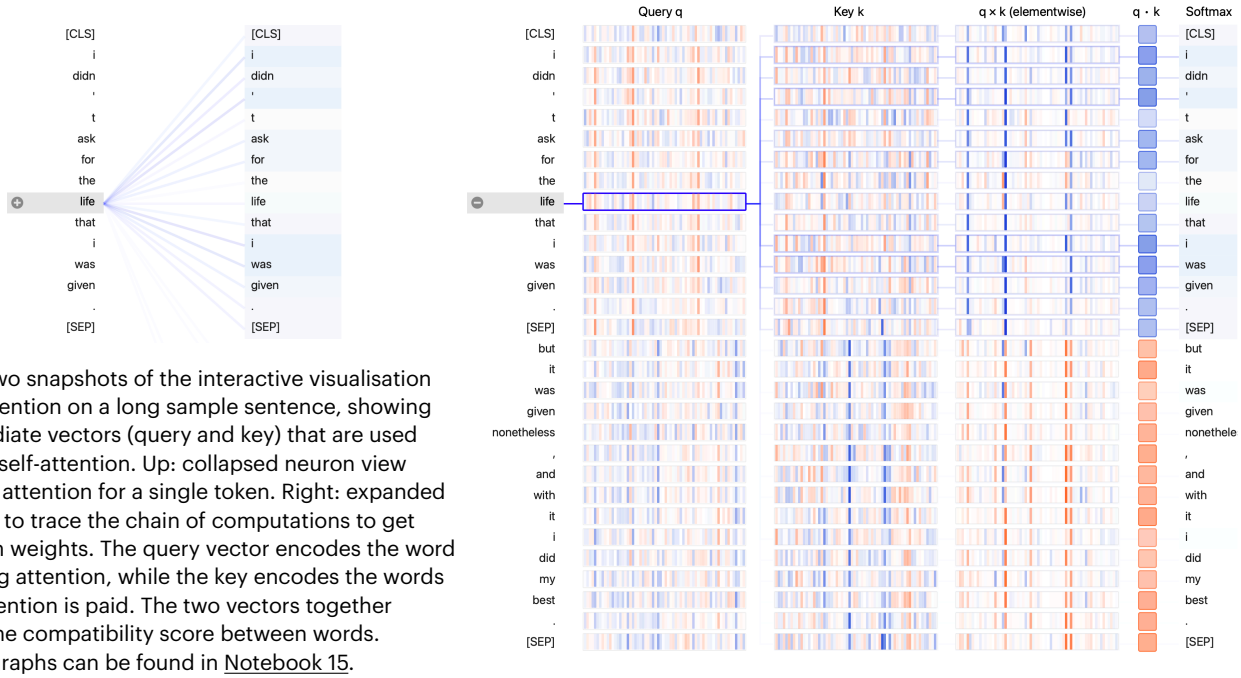


Figure 7.1. Two snapshots of the interactive visualisation of BERT's attention on a long sample sentence, showing the intermediate vectors (query and key) that are used to compute self-attention. Up: collapsed neuron view highlighting attention for a single token. Right: expanded neuron view to trace the chain of computations to get the attention weights. The query vector encodes the word that is paying attention, while the key encodes the words to which attention is paid. The two vectors together determine the compatibility score between words. Interactive graphs can be found in Notebook 15.

giving more prominence to the role of attention to speed up training while reducing computational complexity. The excitement of the authors and their commitment to future extensions somehow foresaw the ample success of their concept and the broad adoption that followed shortly thereafter.

Attention modules

Jakob Uszkoreit had the initial idea of replacing RNNs with self-attention, a mechanism that allows items of an input sequence to interact with each other and find out which they should pay more attention to (Figure 7.1), using learnable weights and a combination of three vectors (query, key, and value). The attention function is defined as mapping a query and a set of key-value pairs to an output that is computed as the sum of the values, weighted by the compatibility of the query with the corresponding key. Here is where the paper gets abstruse and fails in elucidating the inner workings of the attention procedure. Essentially, the three vectors are derived by multiplying a word embedding by three distinct trainable matrices. They are smaller abstractions of the embedding, used to calculate the attention score of each word. It is obtained by taking the dot product of the query vector with the key vector of the respective word we are scoring. The score is then divided by the square root of the dimension of the key vectors to make the gradients more stable and passed through a softmax operation. The value vectors, still untouched, are normalised with softmax as well. Final step, weighted value vectors are summed up, producing the self-attention layer for the position of the embedded word. This enables to simultaneously compute a set of queries in a rather simple succession of matrix operations.

Noam Shazeer proposed to refine the dot-product operation with a scaling factor to counteract the large magnitude that would push the softmax function into regions with extremely small gradients. He also proposed the concept of *multiple attention heads*: another fundamental contribution of the paper to the new generation of sequence models. Instead of performing attention only once, the authors found it beneficial to linearly project queries, keys, and values for a given number of times and compute the attention on each of these projected versions in parallel, yielding multi-dimensional output values that are chained and projected once again to get the final output. Multiple heads let the model focus jointly on different subspaces at different positions, composing a sort of ensemble learner. The Transformer described in the paper employs eight heads for a resulting vector dimension of 512.

The Transformer

The paper provides exhaustive responses to the core questions that the authors set for their research. They deepened the efficacy of attention and provided convincing improvements to established approaches, like the scaling factor and the multi-head module (Figure 7.2). Their brilliant exposition is completed with the debut of the Transformer architecture. It is not a radical new design but a different take on the encoder-decoder structure[226,227] and the first transduction model relying entirely on self-attention. Moreover, it is auto-regressive, in the sense that it consumes the previously generated symbols as additional input when generating the next (Figure 9)[292].

The attention mechanisms featured in the paper set the standard for future research, thanks to the interpretability of the yielded models, their computational efficiency, and the broader span of attentive focus. The adoption of Attention Neural Networks led to dramatic improvements in almost every area of NLP. Today we can notice two main families of transformers: GPT (OpenAI)[226] and BERT (Google).[76]

Attention Neural Networks demonstrated that massive datasets and trailblazing training regimes can further increase the

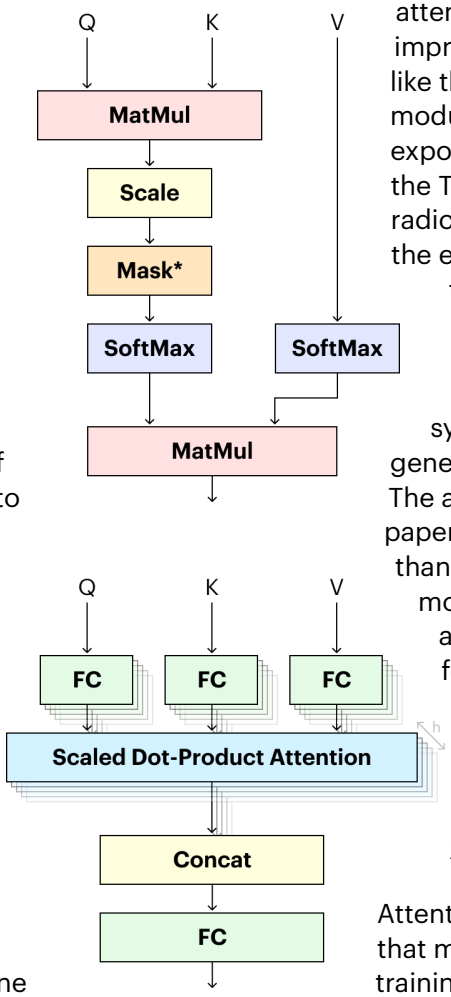


Figure 7.2. Above: conceptualisation of the scaled dot-product attention (SDPA) and multi-head attention (MHA) modules.

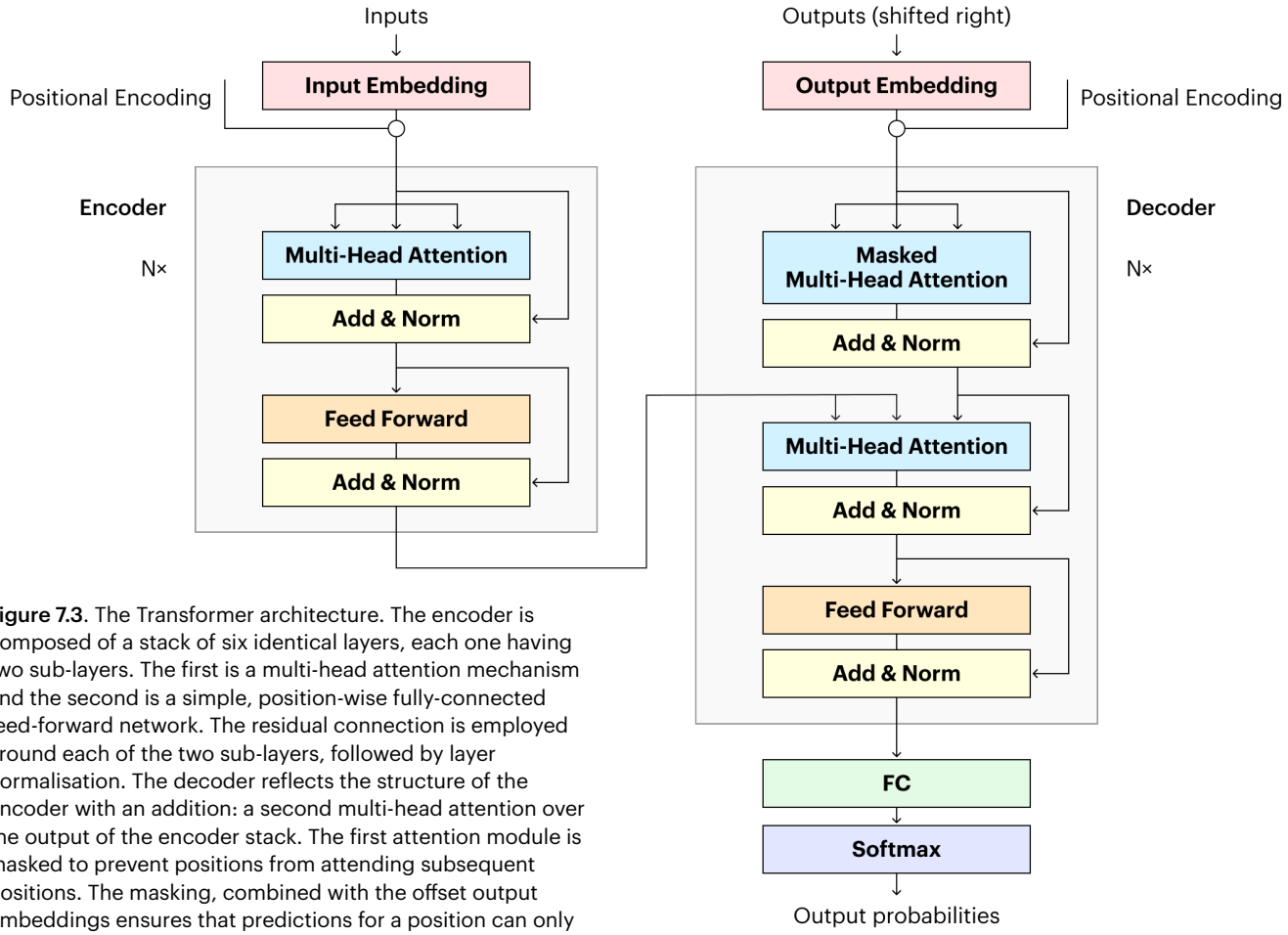


Figure 7.3. The Transformer architecture. The encoder is composed of a stack of six identical layers, each one having two sub-layers. The first is a multi-head attention mechanism and the second is a simple, position-wise fully-connected feed-forward network. The residual connection is employed around each of the two sub-layers, followed by layer normalisation. The decoder reflects the structure of the encoder with an addition: a second multi-head attention over the output of the encoder stack. The first attention module is masked to prevent positions from attending subsequent positions. The masking, combined with the offset output embeddings ensures that predictions for a position can only depend on the known outputs of previous positions.

accuracy of ANNs, and their inception marked a thriving period for PTMs. These are hefty transformer architectures that can be trained once and exported for general use.

The Generative Pre-Trained Transformer (GPT) introduced minimal task-specific parameters and was trained on downstream tasks by simply fine-tuning all pre-trained parameters. We could infer that an intricate enough Transformer might be able to assimilate general language representations through a generative process, entailing an extraordinary similarity between how humans and machines learn. The claim that a transformer architecture bears resemblance to human learning is, however, hardly tenable. Human cognition happens in time, and this particular aspect is much better modelled by recursive neural networks than transformers.

BERT thereupon pioneered the bidirectional language representation. Its authors argued that unidirectional models are limited in training and offered an alternative that has two main objectives, crowning the intentions of the first Transformer: a single, versatile architecture and a robust pre-training regime that ensure a final language model that is apt for any NLP task. Such adaptability spurred several new forks of the project that tailored it for distinct natural languages (UmBERTo, HerBERT,[242] CamemBERT[184]) and different use cases (FinBERT,[9] BERTweet[201]).

It is impressive the impact and the following of the original definition of the Transformer in just a few years. This family of ANNs broadened the panorama of deep learning tasks such as named entity recognition, part-of-speech recognition, lemmatisation, sequence classification, information extraction, sentence similarity estimation, question-answering, translation, summarisation, and text generation in over one hundred languages. Transformer models can also tackle multi-modal problems, like table question answering, activity detection, speech analysis, and optical character recognition.

BERTology[236]

The field of study concerned with investigating the inner working of large-scale transformers like BERT – often referred to as BERTology – is rapidly expanding with profuse attainments, especially in providing better, deeper understandings of the attentive mechanics.[52,210,284,297] BERT performs well on tasks that require sensitivity to linguistic structure. Its representations are hierarchical rather than linear in the higher layers[171] and encapsulate something akin to a syntactic tree in addition to the word order organisation. Its embeddings encode linguistically relevant aspects of hierarchical structure (e.g. part of speech, syntactic chunks, and roles), though they do not appear to show the sharp sensitivity that is found in human processing of reflexive anaphora.[171] Self-attention weights do not directly encode syntactic dependencies, which are recovered from the word vectors.[114,124] BERT is also able to gather some knowledge of the semantic roles, yet with comparably small improvements over the non-contextual baseline.[194] In particular, it struggles to create good collective embeddings from token representations, as attested in my experiment. Although its scores on probing tasks are high,[194] out-of-the-box pre-trained BERT is surprisingly brittle to named entity replacements, suggesting that the model does not actually form a general idea of the entities. An issue that is practically fixed with model fine-tuning: we collect the general-purpose model (previously trained on massive corpora, with ample resources, and for extensive time) and effortlessly train it a second time on a smaller, task-specific dataset. An utter advancement that lowered the barrier for practitioners and empowered them to use top-notch high-performance models with a unified API framework for relatively inexpensive training, evaluation, and production.

BERTology exhibited a conspicuous list of shortcomings that limit BERT's potential. The team at Facebook AI studied the pre-training procedure and measured the impact of its hyperparameters, ultimately averring that BERT is significantly undertrained.[174] RoBERTa, their best shot at optimising the design of BERT, featured a prolonged training session, longer sequences, dynamic pattern masking, and the removal of the next sentence prediction.[174] The trained models are inevitably heavier and laborious to fine-tune than BERT but achieved state-of-the-art results on GLUE,[300] RACE,[150] and SQuAD.[228] Research accomplishments are growing vigorously towards the continuous refinement of these architectures, aiming for higher, human-like scores. Notwithstanding, resuming my initial experimental setting, I prefer to address a more practical, production-oriented transformer design. In the following section, I will describe the principles behind SqueezeBERT: a lighter, undemanding bidirectional transformer that runs more than four times faster than BERT-base while achieving competitive accuracy.[126]

SqueezeBERT

«*Humans read and write hundreds of billions of messages every day. […] Out of these, more than half of the world's emails are read on mobile devices*»; this is an excerpt of the suggestive opening premise of the authors of SqueezeBERT.[126] Most of us can safely assert that the majority of the written content we consume every day is digital, literally.* Considering the vast stream of text at our fingertips, NLP technology has unlimited application possibilities. Countless approaches achieved considerable results, but very few had the flexibility to be deployed at scale (including smartphones and lean back-end server infrastructures), a concern that is inhibiting mass adoption of custom embedded models.
The intention of the authors is thus to target mobile devices and revise BERT to derive a computationally inexpensive alternative from the insights of the Computer Vision community. Developed to achieve faster inference, MobileBERT[281] was a good reference for its strong accuracy on the GLUE benchmark.[300] Analogous to ResNet[112] in CV, bottleneck layers are adopted to reduce the number of parameters and, therefore, the computational cost of the attention layers. Moreover, residual connections are added between the higher layers to retain the signal throughout the network and enable higher information flow.[125] SqueezeBERT leverages two more ideas from CV literature to accelerate NLP. The first is convolutions. Used since the 1980s, convolutional layers are quite flexible and well optimised to dilate a layer to perform up- or down-sampling. The second is grouped convolutions.[148] Extensively used in efficient CV networks like MobileNet,[123] ShuffleNet,[318] and EfficientNet,[156] the term refers to the process of using different sets of convolution filter groups on the same input.

* **Digital**. Origin: late 15th century, from Latin *digitus* 'finger, toe'.

This way, we can learn more features with two or more sub-models that train and back-propagate in parallel. It consists in creating a deep network with a number of layers and then reusing it to have more pathways for convolutions on a single image. Convolutions are so relevant because of their significant speedups in CV networks. BERT-base and RoBERTa-base have the same self-attention encoder architecture and incur approximately the same latency.

Stage	Module type	FLOPs	Latency
Input	Embedding	0%	0.26%
Encoder	PWFCs in SDPA	24.30%	18.9%
Encoder	SoftMax in SDPA	2.70%	11.3%
Encoder	PWFCs in FFNN	73.00%	69.4%
Final Classifier	Additional FCs	0%	0.02%

Table 7.1. Breakdown of computation (in floating-point operations) and latency in BERT-base on smartphone.

Recalling the encoder structure (Figure 9), it is composed of a stack of blocks, each one hosting a scaled dot-product attention (SDPA) module followed by three position-wise fully connected (PWFC) layers, known as feed-forward neural network (FFNN). Each SDPA module contains three separate PWFC layers, which are used to generate the query, key, and value activation vectors for each position in the feature embedding independently. From the computation breakdown shown in the table, it is evident that PWFC layers account for more than 97% of the FLOPs and over 88% of the latency.[126] To address such inefficiency, the authors replaced the PFC layers in the attention modules – which were first introduced in the Transformer and later used in GPT and BERT – with grouped monodimensional convolutions of kernel size equal to one, all this without altering the networks' numerical properties or behaviour.[126]

SqueezeBERT was pre-trained on a combination of Wikipedia and BookCorpus, adhering to the indications of ALBERT,[153] with two prediction objectives: masked language modelling and sentence ordering. It is then fine-tuned on GLUE, a set of nine NLU tasks that provide a good approximation of the generalisability of a model. The first training iteration of SqueezeBERT followed the default training scheme without distillation. Initial pre-training used the LAMB optimiser,[315] for layer-wise adaptive learning rates and little hyperparameter tuning. When training BERT, the optimiser allows the use of large batch sizes, reducing the training time considerably without any degradation in performance. Fine-tuning instead uses AdamW[175] without momentum or weight decay but performing hyperparameter tuning on the learning rate and dropout rate.

Model	QNLI	SST-2	Speedup
BERT-base[9]	92.2%	92.7%	1.0x
ALBERT-base[75]	—	90.3%	1.0x
MobileBERT[68]	88.2%	90.1%	3.0x
SqueezeBERT[67]	90.5%	92.0%	4.3x

Table 7.2. Comparative results of two tests in the dev-set of the GLUE benchmark with relative speedup in training.

SqueezeBERT runs significantly faster than its competitors and achieves comparable results. The average GLUE score of BERT is 85.1%, whilst SqueezeBERT's is 82.4%. To improve the training and increase the final accuracy, the authors reviewed the notion of *knowledge distillation*.[118] In lieu of making predictions using an ensemble of neural models, which is definitely cumbersome and anything but deployable at scale, distillation connotes the compression of knowledge in a single composite model. Essentially, a compact model (the student) is trained to reproduce the behaviour of a larger, more complex one (the teacher). It is the case of DistilBERT,[255] coached by distilling BERT's pre-training phase, making it 40% lighter and 60% faster while retaining 97% of its original language understanding capabilities. The authors of SqueezeBERT opted for a relatively simpler form of distillation, employing it only to the final layer and only during fine-tuning. In addition, inspired by STILTS[215] and ELECTRA,[53] they applied transfer learning from one GLUE task to the others, and observed the following results.

Model	QNLI	SST-2	Speedup
DistilBERT[78]	89.2%	91.3%	2.1x
Turc[81]	89.4%	91.1%	2.1x
Theseus[82]	89.5%	91.5%	2.1x
MobileBERT[68]	91.5%	92.5%	3.0x
SqueezeBERT[67] (d.)	90.9%	92.5%	4.3x

Table 7.3. Comparative results of two tests in the test-set of the GLUE benchmark with relative speedup in training. Tables are adapted from (Iandola et al., 2020).[126]

The task in question is the multi-genre natural language inference (MNLI)[310] for broad sentence understanding. SqueezeBERT is pre-trained like before but fine-tuned on the MNLI. The computed weights are then used as initial student weights for the other GLUE tasks, except for CoLA,[301] the unbalanced binary classification task of acceptability judgments for grammatical knowledge assessment.[47]

The teacher model is a BERT-base model, pre-trained using the ELECTRA method,[53] fine-tuned on MNLI and then again on each GLUE task independently. The task-specific teacher weights are finally used for distillation. The results of the distilled models in the previous table show that SqueezeBERT kept its speedup advantage at the cost of a modest decrease in accuracy if compared to MobileBERT. This is true for question-answering (QNLI)[228] and sentiment analysis (SST-2),[276] while SqueezeBERT beat MobileBERT in semantic equivalence (MRPC).[77]

Model	Distillation	GLUE score	Speedup
BERT-base	—	78.3%	1.0x
SqueezeBERT[67]	**—**	**76.9%**	**4.3x**
Theseus[82]	DESW	77.1%	2.1x
MobileBERT[68]	DEW	78.5%	3.0x
SqueezeBERT[67]	**DS**	**78.1%**	**4.3x**

Table 7.3. Comparative results of GLUE scores on the test-set with relative distillation techniques and speedup in training.
Legend: D = distillation of final layer; E = distillation of encoder layers; S = transfer learning across GLUE tasks (aka STILT); W = per-layer warmup.

In both evaluation and test settings, the distilled version of SqueezeBERT sees an improvement of two percentage points on the GLUE score. Looking at the accuracy of the distilled models:

- SqueezeBERT outperforms MobileBERT on four tasks (QQP, STS-B, MRPC, RTE).
- MobileBERT outperforms SqueezeBERT on four other tasks (MNLIs, QNLI, SST-2, CoLA).
- On the reading comprehension task (WNLI),[162] both models predicted the most frequently occurring category.

To improve the accuracy on the WNLI task, data augmentation approaches were taken into consideration,[144] however, the authors refused to use it for fairness against the baselines.

Heedful of the reasons behind its design and the results it attained, we can ascertain the success of SqueezeBERT in borrowing grouped convolutions from CV and enacting the concept of a new, efficient NLP transformer architecture to be comfortably deployed at scale. Furthermore, soft distillation improved accuracy to such a degree that it is competitive even with the original implementation of BERT.

For a final review, I applied the transformer architectures mentioned in this section to the experimental setting presented in the previous. Using the raw corpus of TED talk transcripts, I fine-tuned four models for a total of twenty epochs ($8{\times}10^4$ sample observations) without outlier detections (Notebook 14). All models used their specific tokeniser and the standard PyTorch sequence classification head on top, that is, a linear FC layer on top of the pooled output of the transformer.

Model	A	P	R	Size
XGBoostClassifier[21]	75.6%	74.3%	73.3%	1.6
BERT-base-c[9]	93.2%	93.3%	92.7%	433.3
RoBERTa-base[10]	85.2%	85.9%	83.7%	498.7
DistilBERT-base-c[78]	92.4%	92.2%	92.5%	263.2
SqueezeBERT-uc[67]	**94.5%**	**94.6%**	**94.2%**	**204.5**

Table 7.3. Performance results of the transformer architectures on the text categorisation task presented in section Six.
Legend: A = accuracy, P = precision, R = recall. Size is expressed in megabyte (MB). Note: model names ending with *-c stand for "cased", whilst *-uc for "uncased".

The table evinces the staggering performance of the models in this experiment: in absolute terms, all four transformers had far superior accuracy compared to the baseline models (XGBoost is included for comparison); in relative terms, there is a glaring difference between the leading model across all three metrics, SqueezeBERT, and the runner-up, BERT. It is not just the best-performing model but also the lightest. This is not a coincidence: SqueezeBERT is a finely efficient architecture, with a condensed number of parameters that makes it very pliable when performing fine-tuning. Similarly, RoBERTa is so stiff and tight that requires a more vigorous training session and an earlier stop to prevent drastic overfitting. Yet, the issue of efficient training remains, as the four models required roughly the same amount of time.

Zero-Shot Learning

Until now, I have always considered NLP in traditional, supervised terms. My experiment consisted in a multi-classification task, carried out first by a set of baseline statistical models, then by more complex neural architectures. The first part of the experiment was determined on comparing the performance of pre-trained language models in yielding more informative document embeddings. Static word vectors proved far more effective than their modern, context-sensitive counterparts. This brought to the second part of the experiment: rather than relying on document-level vector representations, I used pre-trained transformer architectures to process the text samples from the start, fine-tune the model, and perform the categorisation. The experiment has hence been the setting to discuss the fine distinctions in semantic multi-classification and examine different members of the transformer family. Now, resuming the core interest of the dissertation, I want to use the transformers to derive cultural insights from text. To do so, I had to overturn the conventional paradigm of how we conceive a ML task, bridging supervised and unsupervised learning. In this section, I will explore Zero-Shot Learning in the context of text classification (ZSTC).

In recent years the world of NLP has been thriving. As I analysed in the previous sections, the community of researchers figured out advanced learning methods from extensive sets of unlabelled textual data. The success of transfer learning allowed outstanding achievements in language modelling and is continuously pushing the boundaries of what can be attained. As current state-of-the-art PTMs are trained on massive corpora, I assumed that they "know" already quite enough about our language that we can exploit them as estimators of units of culture. Fine-tuning a large language model to recognise tens of labels would be tremendously expensive and time-consuming, not to mention the need for an annotated corpora to enable the operation. Instead, we can query such a model and expect it to estimate a sense of closeness of a descriptor to a given input text. This is the fundamental idea of Zero-Shot Learning (ZSL): «*get a model to do something that it was not explicitly trained to do*».[67] In other words, the model has to learn how to recognise new concepts by just having a description of them.

In this setup of «*dataless classification*»,[36] the learning protocol makes use of general knowledge to induce classifiers without the need for any labelled data. The key technical direction in NLP builds on the model's ability to understand the descriptors that are provided with the input text, scilicet, representing the classes provided in the same semantic space as that of the document and estimating how much they are close in meaning.[154] The concept is somewhat similar to the semantic networks I draw in section Three to describe the symbolic universes around a tag in TED's cultural milieu. Given a class, the ZSTC model processes the input text and computes a relevance score between the two. However, it should not just compute a similarity computation. Indeed, the model can be fine-tuned on an annotated Natural Language Inference (NLI) dataset, modelled via sequence-pair classification. The objective of the task is to consider two sentences, a premise and a hypothesis, and determine whether the latter is true (entailment) or false (contradiction) given the first.[314] This enhancement is reported to free the trained ZSTC model from establishing mere semantic similarities, improve understanding, and refine its conceptual entailment capabilities.[310]

As demonstrated by GPT-2, sizeable language models are unsupervised multitask learners.[227] When trained on extensive datasets and considering their ample set of parameters, they can learn tasks (e.g. question answering, machine translation, reading comprehension, and summarisation) without any direct supervision. GPT-3 then consolidated this outlook:[34] extremely large and general-purpose PTMs perform competitively well on downstream tasks with far less task-specific samples than would be required by smaller models. These two achieve an astonishing performance, yet they are amongst the most demanding ever designed, for the number of their parameters, training time, and computational resources required, placing them beyond the reach of most users. However, smaller transformer models like BERT[76] proved to be capable of encoding a tremendous amount of information in their weights. Whilst learning linguistic regularities, recent transformer-based PTMs can store relational knowledge that can be accessed by conditioning on latent context representations, or using the original weights to initialise a task-specific model that has to be fine-tuned.[214]

The ZSL paradigm is inspired by how human beings are able to identify a new object from its description, leveraging similarities between the two entities and previously learned concepts.[237] Similarly, zero-shot approaches are designed to learn this intermediate semantic layer and apply the attributes it gathers at inference time to estimate the relation between a text and a descriptor. In this setting, the attribute description of a class is referred to as *signature*.

Zero-shot learning is inherently a two-phase process. In the first, training, information about the attributes is captured. In the second, inference, the latent knowledge is used to categorise instances among new sets of classes. Transformer-based models like BERT proved to have many advantages over structured knowledge bases.[214] For instance, they are easily extendable with more data and require no schema engineering nor human supervision during training. Without fine-tuning, BERT learns certain types of factual knowledge faster than other approaches, features relational information competitive with more traditional NLP methods, and has a satisfactory performance on open-domain question-answering against a supervised baseline.[214] In particular, BERT-large is so accurate in knowledge capturing that it is comparable to an oracle-based entity linker gold standard.[76] It consistently outperforms other language models in recovering factual and common-sense information while being more robust to the phrasing of a query. In addition, factual insights can be derived surprisingly well, however, some relations can be very poor without apt NLI refinement.[214]

Diagnosing syntactic heuristics in natural language inference, McCoy et al.[186] found that models like BERT tend to rely heavily on fallible syntactic heuristics, suggesting that there is substantial room for improvement in NLI systems. However, Y. Goldberg assessed that BERT can learn English syntactic phenomena through naturally occurring or manually crafted linguistic regularities with remarkable results.[105]

In my quest to derive cultural indicators from texts, I used BART:[168] a denoising auto-encoder for pre-training seq-to-seq models. It combines bidirectional and auto-regressive transformers and is built to be applicable to a very wide range of end tasks, including NLI. Trained through arbitrary text corruption, it learns by reconstructing the original document. This makes it particularly effective in comprehension tasks, matching the performance of RoBERTa on GLUE and SQuAD in a comparable training setting while achieving state-of-the-art results in abstractive dialogue and summarisation tasks. It also achieves stunning impressive results in machine translation.[168]

A decisive advantage of this setup is the noising flexibility, which generalises BERT's original word masking and next sentence prediction objectives, forcing BART to make longer range transformations to the input.[168] It employs the standard Transformer architecture[292] (Figure 7.3) with the sole exception that all rectified linear units are replaced by Gaussian error linear units.[113] The architecture is hence closely related to the one used in BERT, with two differences: each layer of the decoder performs additional cross-attention over the final hidden layer of the encoder, and the additional feed-forward network before word prediction is removed. Compared to an equivalently sized BERT model, BART contains roughly 10% more parameters. The training objective consists in optimising the reconstruction loss of a corrupted document, that is, the cross-entropy between the original text and the output of the decoder.

BART is not tailored on a specific noising scheme, allowing the application of several types of document corruption: token masking[76] and deletion, text infilling,[132] sentence permutation, and document rotation.[168] Besides more traditional fine-tuning tasks like text classification, sequence generation benefits from BART's autoregressive decoder. Closely related to the denoising training objective, the model catches the salient features in the input text and adapts them for summarisation, or matches them with previous latent knowledge for question answering.

In my opinion, these characteristics make BART the best candidate to procure the units of culture from a series of documents. Whilst in my previous experiment I lamented the fact that larger, more intensive transformer models like RoBERTa[174] were too stiff to be fine-tuned on my dataset of TED talks for semantic categorisation, now I appreciate such intricate architectures for the new task at hand. In the next section, I will use a pre-trained BART-large[168] language model, fine-tuned on the MultiNLI dataset,[314] and implemented in a custom pipeline for zero-shot estimation of personality traits and units of culture.

Towards The Units of Culture

In this final stage of the investigation, I will dive into the details of my findings. Since the very beginning of the study, the results obtained with BART as a zero-shot text classifier (ZSTC) have been exceptional. Given a collection of previously unseen descriptors, the ZSTC provides solid estimates about their meta-semantic relatedness with the input document. It does not just act as a simple classifier, recognising whether a talk is about innovation or the environment. Instead, drawing from its internal latent understanding of language, the model is capable of reckoning profound connections between a candidate label and a text at a striking performance, outputting a quantitative assessment of their relation.
For example, in one of my earliest tests, the model was able to estimate different levels of caution and suspicion in small propositions. Increasing in expressive complexity, BART always responded with very impressive responses, proving that it can extract sense and intentions of the speaker, analyse their sentiment, interpret irony, and even recognise veiled insults. An evident nuisance that emerged is the utter lack of viable evaluation methods. In months of trials, the model never committed any blatant mistake in prediction, yet there is no other way to fix a bias than adjusting the contents of the corpus and restarting the operations of training and fine-tuning. It may indeed be very problematic to rely on a dataset skewed in a particular direction. However, volume and variety of the training data are the best bet to mitigate the risk of involuntary favouritism, especially if the task is to derive personal insights. BART-large is pre-trained on a massive collection of documents: the entire English Wikipedia, eleven thousand unpublished books, sixty-three million news articles (crawled between 2016 and 2019), the OpenWebText dataset (used to train GPT-2), and a subset of CommonCrawl. After an intensive pilot study, I am led to believe that BART is sufficiently unbiased for my purposes. As for the evaluation of the results, I resorted to personal supervised checks, often followed by a confrontation with colleagues for impartiality.

Personality Traits

According to the approach advanced by the *dispositional theory* in psychology, habitual patterns of behaviour, thought, and emotion can be studied in relation to our personality. Prominent avant-gardist in this field is American psychologist Gordon W. Allport, who developed an eclectic and influential conception of personality that gives prominence to the uniqueness of the individual and the unavoidable influence of the present context.[4,7] He particularly rejected both the psychoanalytic and behavioural approaches to personality, and hypothesised the figurative interplay of two forces that determine how we behave and communicate:[2]

- **Genotypes**: the internal drives associated with the way the individual retains information and uses it to interact with the external world;
- **Phenotypes**: the external drives associated with the way the individual accepts their environment and others influence their behaviour.

As Zipf[320] conceived language as the optimised result of the confrontation between the forces of unification and diversification, Allport saw genotypes and phenotypes affect our culture.[5] In his involvement in the dispositional theory,[6] he spurned the idea of classifying people by their quirks and maintained that every individual is unique and distinguished by their peculiar *traits*. These are habits of social significance,[3] very predictive of our nuanced cultural peculiarities. They frequently reflect our proclivities and our personally embedded systems of meaning.

Following this perspective, personality traits are quintessential to describe the current transitory disposition of an individual. Allport devised a three-level hierarchy to ease the study of personality traits across varying cultures:[3]

- **Cardinal traits**: rare ruling passions and obsessions that determine and control the behaviour of the individual;
- **Central traits**: general aspects that set the foundation for conventional behaviour;
- **Secondary traits**: marginal aspects noticeable only in certain circumstances.

The reason of adopting such framework is to separate one from the other: ignoring culture enables to focus on the personal traits and their ties with the individual.[183] Indeed, Allport's theory centres its focus on the individual over the situation in which they are in.[196] In my research, I aspired to do the same, capturing an individual's set of central dimensions from their utterance.

Taking into account the dispositional theory as a solid methodological approach for the study of human dispositions, I designed the BART ZSTC to estimate the so-called *Big Five personality traits*, a reviewed taxonomy of psychometric factors that can be used to describe an individual's idiosyncrasies and comprehend the relationship between personality, values, and behaviour.[286,103] They represent five overarching domains subsuming the most known central traits and a basic structure behind all of them (Table 9.1).[206,60]

It is reported that factor analysis on personality survey data reveals significant semantic associations.[103] In my general assumption, these linguistic regularities can be examined without the need for a questionnaire. As I reviewed in the previous sections, language models can estimate the same semantic associations, yielding precise outcomes for a given document in a matter of seconds. This is fascinating: using words, or linguistic symbols, to elicit a quantitative report of the cultural descriptors of an individual, leveraging the latent meta-semantic understandings of an unsupervised computational model fine-tuned on information entanglement.

For each of the big five traits, I devised a set of four descriptors, two positives and two negatives. This way, obtaining a total of twenty indicators in the range $[0,1]$, it is possible to compute five indicators in the range $[-1,+1]$, by calculating the difference of their respective sum:

$$T = \Big[(p_1 + p_2) - (n_1 + n_2) \Big]$$

Prompted with a document and the twenty personality descriptors, my algorithm feeds the text to the BART ZSTC, collects the estimates, and aggregates them into the five indicators, returning a table of results and a radial plot (Figure 9.1). Looking at the single results, indicator scores might not appear reasonable at first. It is indeed hard for a human to determine how much a talk is curious or confident from zero to one. This is why numbers need some interpretation: we have to consider exceptionally high or low values that diverge from average. In this case, those that are significantly distant from zero, in both positive and negative directions (Notebook 16).

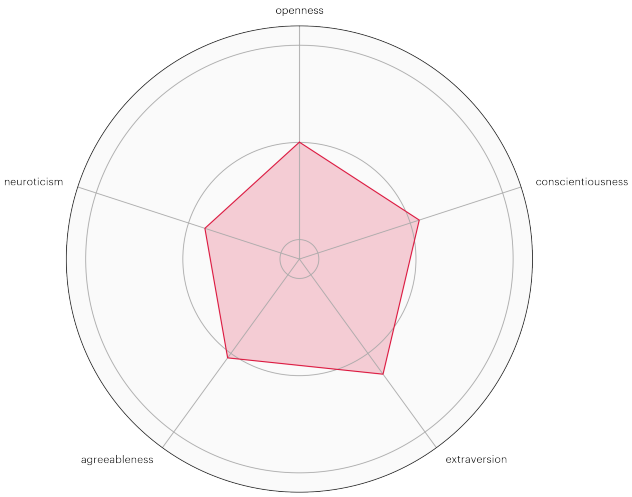


Figure 9.1. Radial plot showcasing the (big five) personality traits of Sir Ken Robinson's talk *Do schools kill creativity?* The smaller and the larger circles represent the limits of the spectrum (-1 and +1); the one in the middle is the zero. The talk scores high in extraversion (0.264) and low in neuroticism (-0.176), so the speaker might have been passionate, affectionate and quite confident.

Trait	Positive/Negative Indicators		Dimensions
Openness	*inventive, curious*	*consistent, cautious*	active imagination, aesthetic sensitivity, rituality, absorption, adventurousness, intellectual curiosity, challenging authority, attentiveness to inner feelings, etc.
Conscientiousness	*efficient, organised*	*extravagant, careless*	diligence, efficiency, organisation, self-discipline, need for achievement, carefulness, thoroughness, deliberation, attentiveness to potential danger, etc.
Extraversion	*outgoing, energetic*	*solitary, reserved*	enthusiasm, gregariousness, discretion, assertiveness, need for outside gratification, search for social interaction, etc.
Agreeableness	*friendly, compassionate*	*critical, rational*	sympathy, trust, cooperation, consideration, empathy, altruism, modesty, straightforwardness, competition, compliance, etc.
Neuroticism	*sensitive, nervous*	*resilient, confident*	worry, anxiety, fear, anger, frustration, envy, jealousy, guilt, self-consciousness, shyness, depression, loneliness, etc.

Table 9.1. The Big Five personality traits, with the twenty indicators used in my investigation, and some explanatory dimensions.

Established that the algorithm yields good quality results, I applied it to the entire corpus of TED talks looking for meaningful correlations with the topics of discussion (Figure 9.2). First, it is important to address the format of a talk and its communicative intention. The average guest aims at being as persuasive as possible, which translates into higher values of conscientiousness and extraversion (Figure 9.3). Speakers often display a planned, disciplined demeanour, accompanied by careful, diligent utterances. Despite losing some spontaneity, the conventional TED talk features a captivating, sociable way of expression. Within this framing, there are many aspects to examine (Notebook 17).

For instance, observing topics like *society*, *politics*, and *global issues*, we can notice a low score in organisation. Indeed, it is hard to advance an ordered plan or solution to deal with these matters, and language reflects this uncertainty. Yet, a similar subject like *climate change* tells a different story: it has high scores in openness and conscientiousness, stressing aspects like creativity, curiosity, and also organisation, but it scores low in confidence and resilience. This suggests communal relevance and ambition, and a certain degree of apprehension for the future. *Social change* is thus a middle ground: talks are sensitive, restive, and clear-minded, but not very methodical.

Following the examination of the big picture, I analysed the cultural idiosyncrasies through agglomerative clustering (Figures 9.5-9).

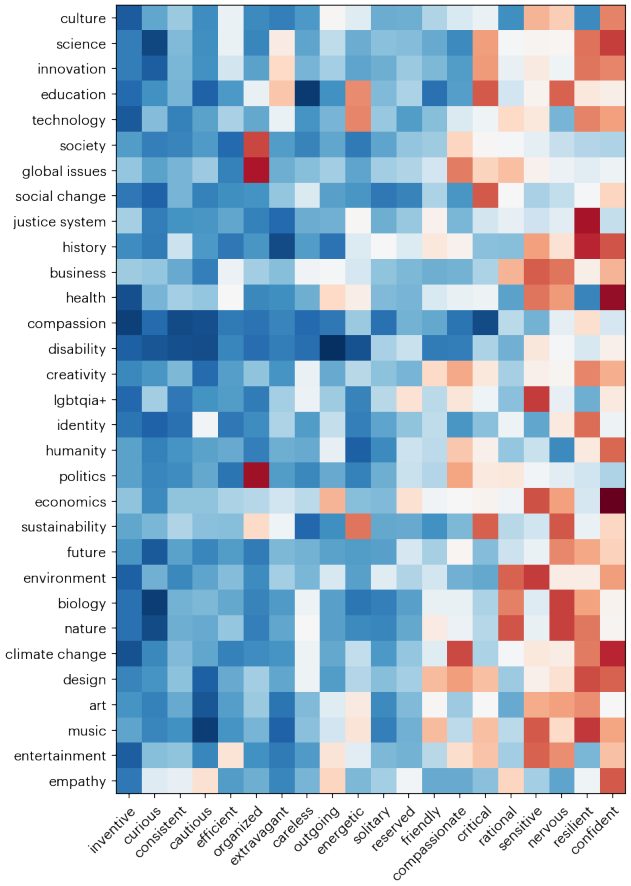


Figure 9.2. Diverging heatmap of the personality indicators associated to a selection of topics in the TED talks dataset. Shades of blue indicate higher scores, reds the lower.

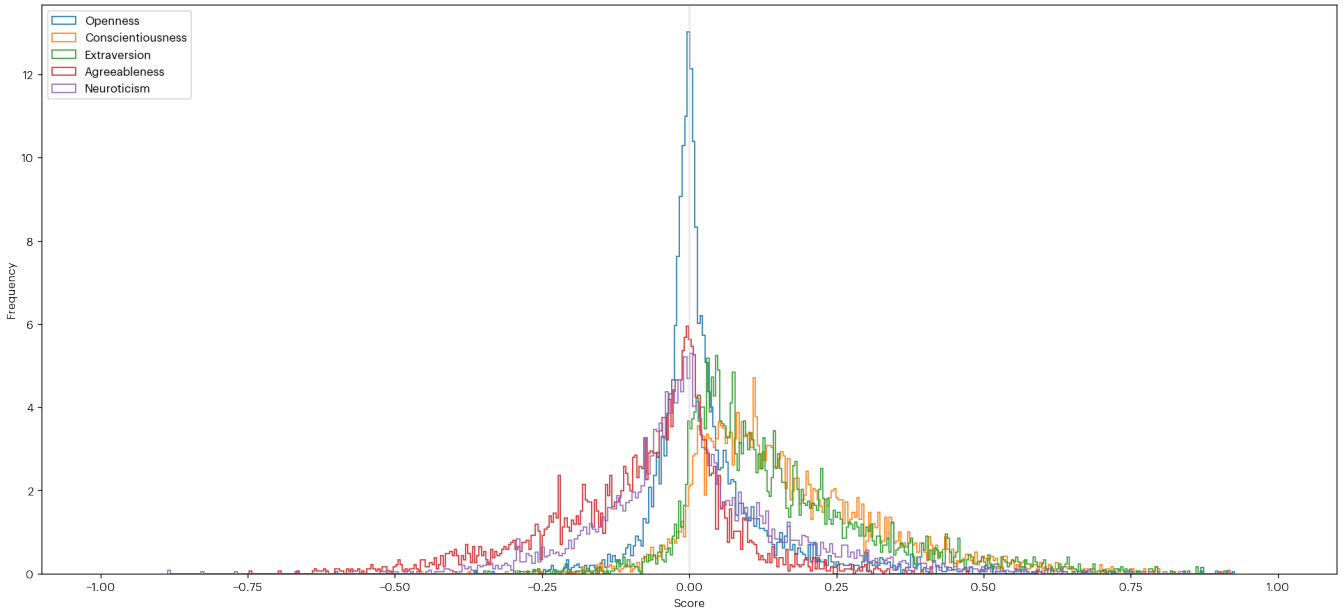


Figure 9.3. Frequency distribution of the aggregated personality scores of the talks in the TED dataset. Values of conscientiousness and extraversion are predominantly above zero, which is comprehensible given the format of the show.

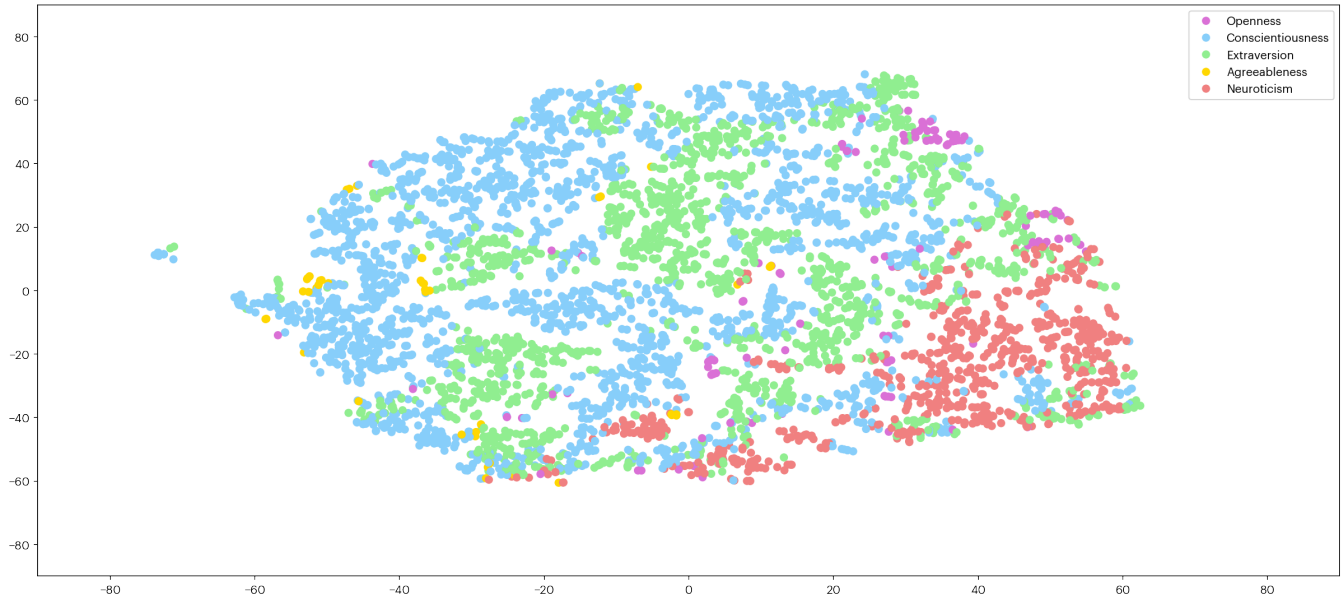


Figure 9.4. Arrangement of the TED talks in a 2D scatterplot obtained from the vector of personality traits, compressed using T-SNE. The colour of every dot corresponds to the personality trait that has higher score, indicating a specific inclination of the speaker; spacial disposition gives a clue about the other personality traits. For example, highly neurotic talks are at the opposite side of the most agreeable ones, in a spectrum that includes different shades of conscientiousness and extraversion.

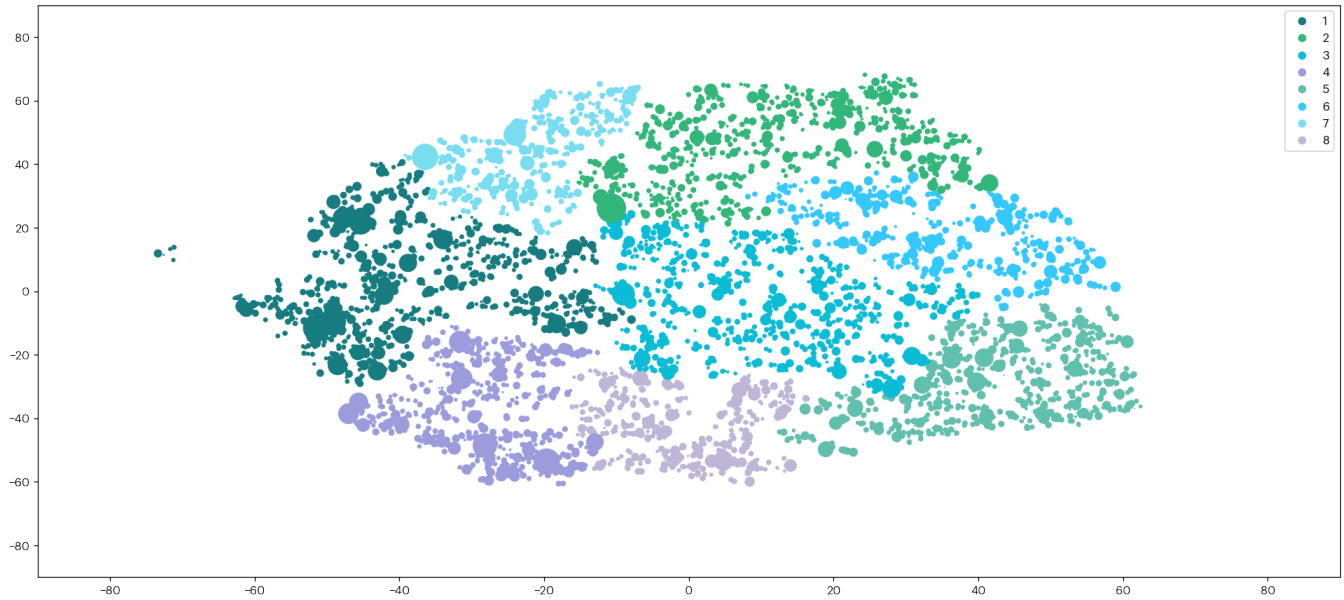


Figure 9.5. Agglomerative clustering performed on the personality traits of every talk and presented in the same arrangement of Figure 9.4. The colour of every dot indicates a cluster, while size is adjusted based on the view count of the talk to reflect the talk prominence in the cultural milieu.

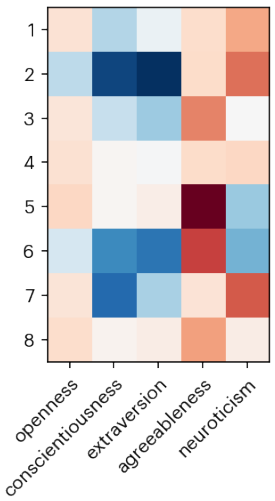


Figure 9.6, right. Divergent heatmap of the big five personality traits of the eight clusters depicted in the scatterplot of Figure 9.5. Diversity suggests differences in the symbolic universes of the speakers.

Talks in cluster 4 (bottom left) are plain and devoid of any particular inclination, while on their right, those in cluster 8 (bottom), score lower in agreeableness. Moving to the right side of the graph, talks in cluster 5 (bottom right) have the lowest score in agreeableness and the second-highest in neuroticism. Just above, talks in cluster 6 (centre right) show a trend reversal of the trend: they are still low in agreeableness, and higher in neuroticism, mainly in extraversion and conscientiousness. Talks in cluster 3 (centre) and in cluster 1 (centre left), exhibit this same pattern, only with more subtle intensities progressing to the left. Talks in cluster 7 (top left) and cluster 2 (top) have the lowest scores in neuroticism and are the strongest in conscientiousness and extraversion.

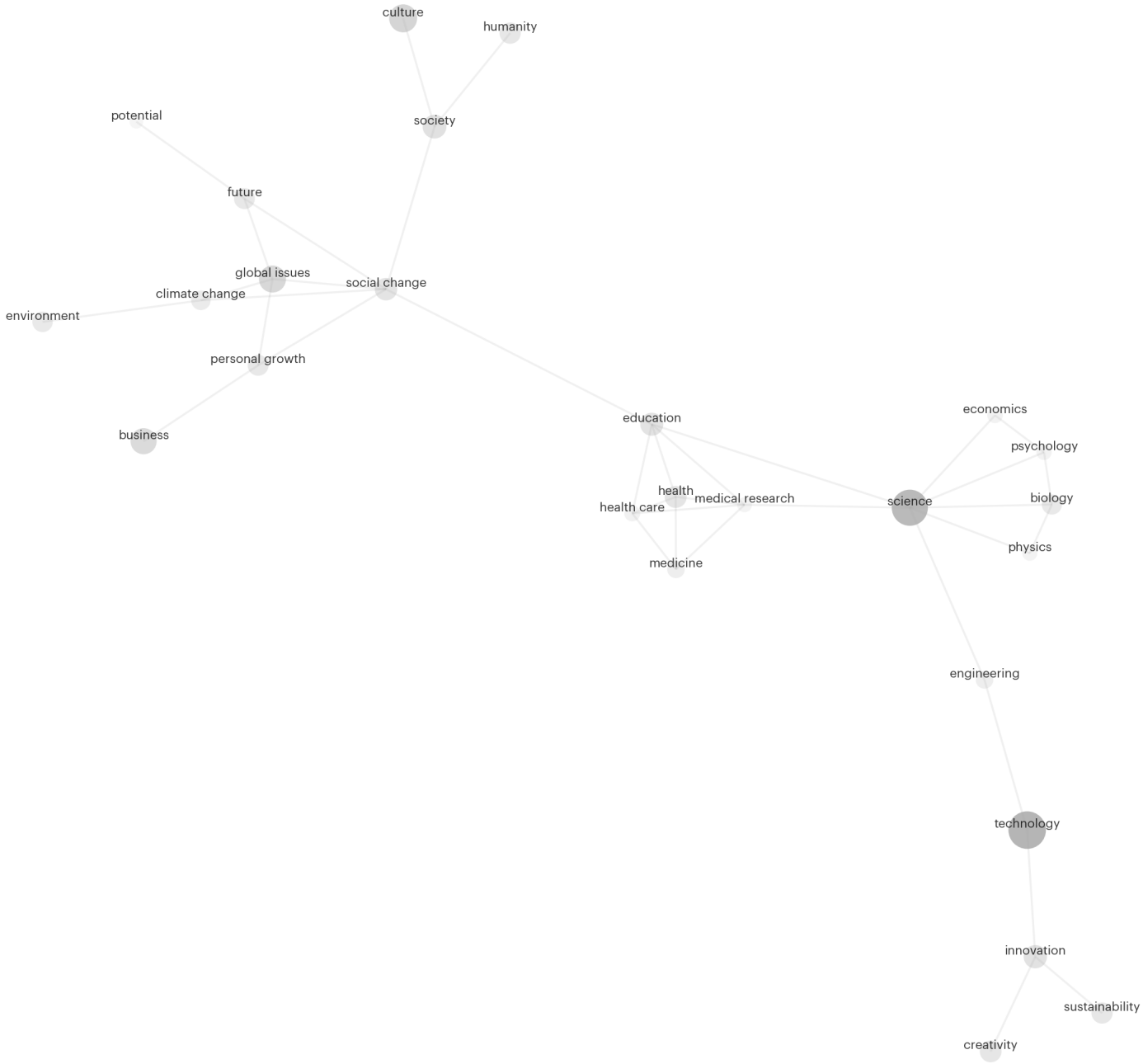


Figure 9.7. Semantic network summarising the prominent topics of the talks in cluster 1. Size and shade represent frequency.

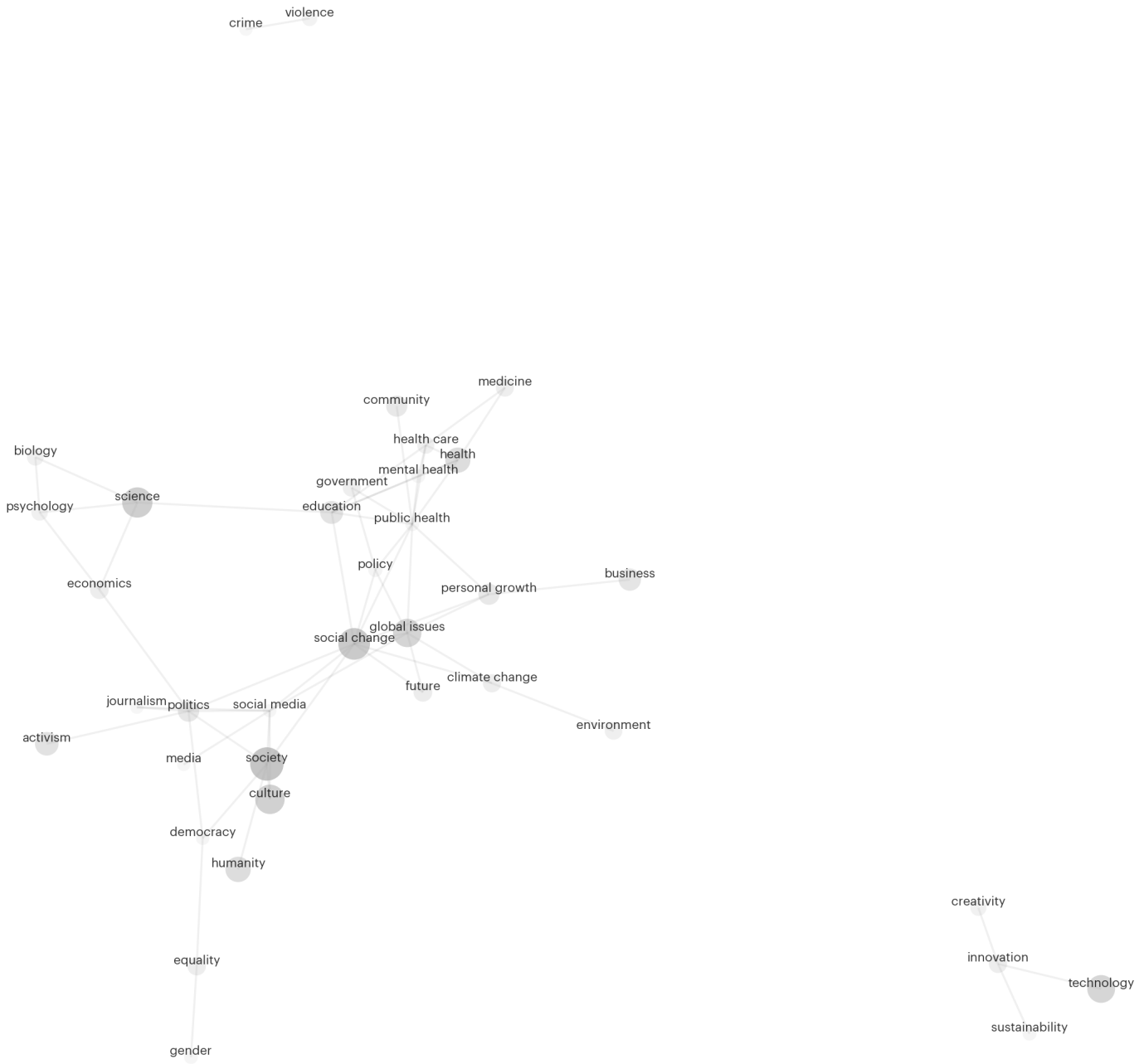


Figure 9.8. Semantic network summarising the prominent topics of the talks in cluster 5. Size and shade represent frequency.

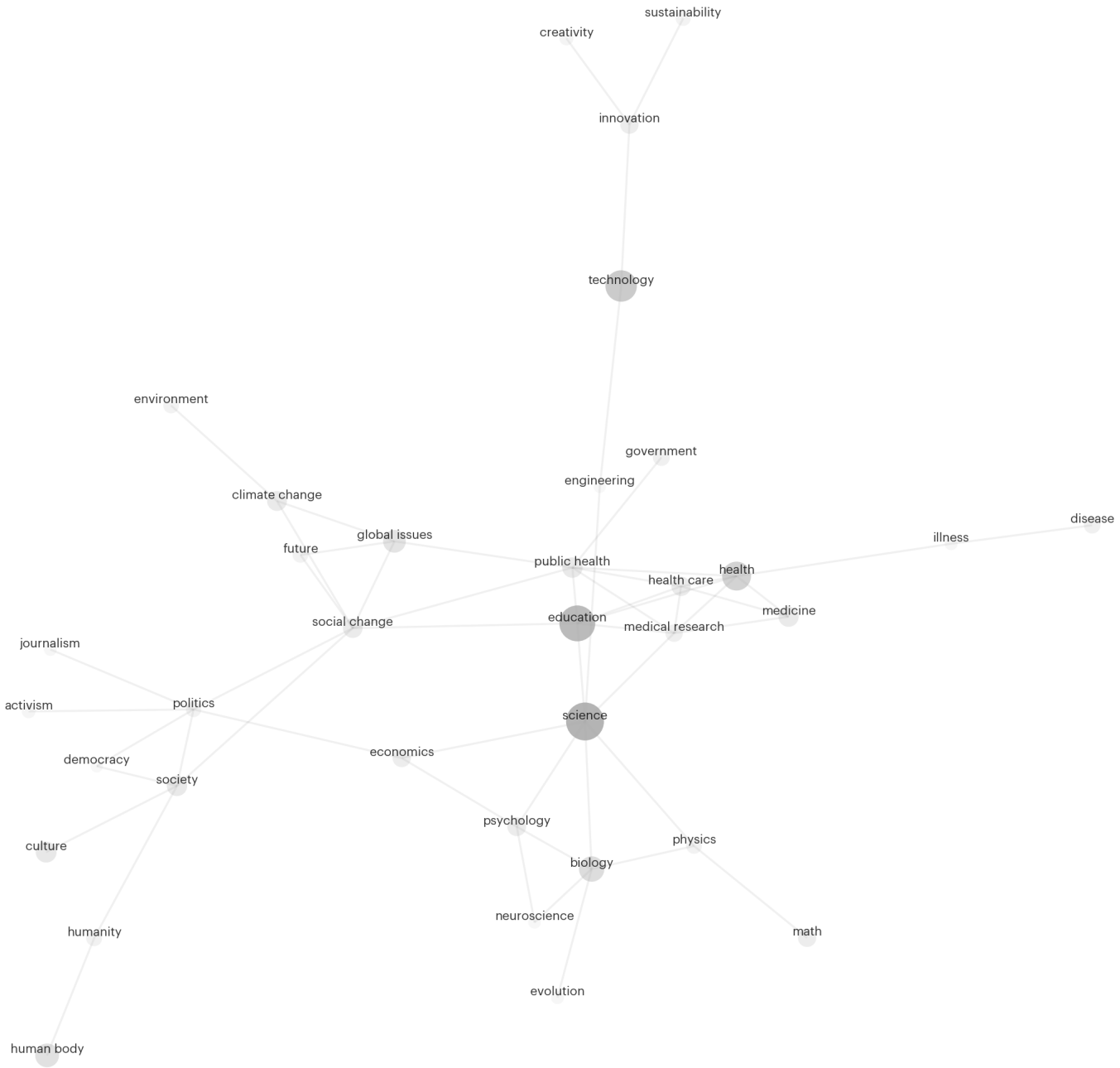


Figure 9.9. Semantic network summarising the prominent topics of the talks in cluster 6. Size and shade represent frequency.

Culture originates from people. If we can measure what goes on inside of their mind, it is possible to estimate what will emerge from them. By looking at culture as a distribution of personal attributes that then plays out in the wider world, we can start to understand people's proclivities. My experiment showcased that personality traits are a good metric for understanding attitudes and behaviour. In particular, it demonstrated how the words we use to describe our experiences provide a wealth of implicit insights.

I came up with an additional set of sixteen indicators (and relative antonyms) to get a more granular understanding of every talk in the dataset and check whether this predictive framework can generalise well (Figure 9.10, 9.11). BART proved an incredibly convincing performance in detecting cultural traits and other nuances across a wide range of challenges. Using adjectives as descriptors is crucial. For example, combining the two antonyms *progressive* and *conservative* yields a simple yet very robust indicator of someone's political orientation. The same goes with *optimistic* and *pessimistic*, or *idealist* and *pragmatic*. The single value does not mean much by itself, but a collection of these indices can bring to a cogent snapshot of reality. With a tailored array of classes, applications are manifold (Notebook 18).

Conclusions

In the course of this dissertation, I went on a quest to explain how we can derive cultural insights from language using the latent understanding of a computational model trained on a massive corpus of text without supervision.

After providing an actionable definition of culture and exploring the striking characteristics of human language, I showcased the symbolic universes that can be elicited from a cultural milieu, in my case, a custom archive of TED talks transcripts. Using a word as query, my pipeline creates a semantic map of the most prominent topics. This is possible thanks to semantic embeddings. Using the same dataset, I arranged a text categorisation challenge to review the performances of both baseline statistical machine learning models and state-of-the-art transformer architectures. To procure the personal traits of the TED speakers in the dataset, I devised a zero-shot text classifier employing a task-agnostic, pre-trained neural architecture fine-tuned for information entailment. The model is light in size compared to the alternatives, low in latency, and utterly generalisable on any set of labels. It yields impressively relevant intermediate results that, once aggregated, tell a compelling story.

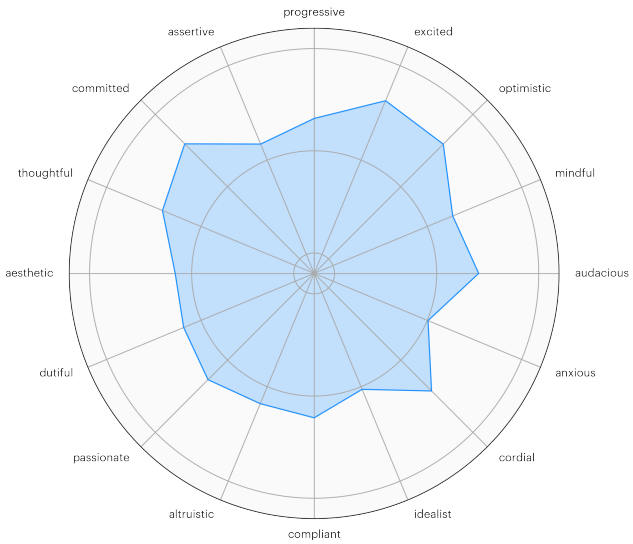


Figure 9.10. Radial plot showcasing the cultural insights of Sir Ken Robinson's talk *Do schools kill creativity?*

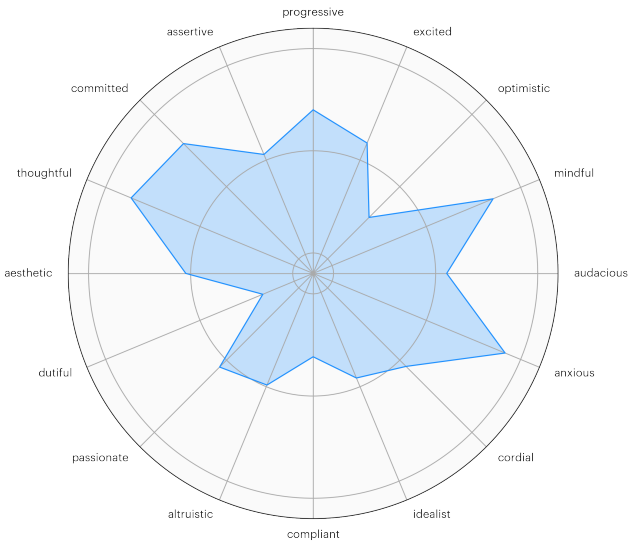


Figure 9.11. Radial plot showcasing the cultural insights of Cathie Wood's report on *Inventories & Deflation* of July 2022.

References

1. S. Albawi, T. A. Mohammed; Understanding of A Convolutional Neural Network, 2017
2. G. W. Allport; Becoming: Basic Considerations for A Psychology of Personality, 1955
3. G. W. Allport; Concepts of Trait and Personality, 1927
4. G. W. Allport; The Nature of Personality 1950 (ed. 1975)
5. G. W. Allport; The Person in Psychology, 1968
6. G. W. Allport; Pattern and Growth in Personality, 1961
7. G. W. Allport, P. E. Vernon; Studies in Expressive Movement, 1933
8. T. Andersen; Quantum Wittgenstein – Metaphysical debates in quantum physics don't get at 'truth' – they're nothing but a form of ritual, activity and culture, 2022
9. D. Araci; FinBERT: Financial Sentiment Analysis with Pre-trained Language Models, 2019
10. M. Arnold; Culture and Anarchy: An Essay in Political and Social Criticism is a series of periodical essays, 1869
11. W. R. Ashby; An Introduction to Cybernetics, 1956
12. E. Bach; Structural Linguistics and the Philosophy of Science, 1965
13. C. Baerveldt, T. Verheggen; Enactivism, 2012
14. D. Bahdanau, K. Cho, Y. Bengio; Neural Machine Translation by Jointly Learning to Align and Translate, ICLR, 2015
15. J. A. Ball; Memes as Replicators, 1984
16. J. A. Banks, C. A. McGee; Multicultural Education, 1989
17. G. Bateson; A Sacred Unity: Further Steps to an Ecology of Mind, 2005
18. G. Bateson; Mind and Nature: A Necessary Unity, 2002
19. G. Bateson; Steps to an Ecology of Mind – Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology, 1972 (rep. 2000)
20. R. E. Bellman; Adaptive Control Processes, 1961
21. R. E. Bellman; Dynamic Programming, 1957
22. Y. Bengio; New Distributed Probabilistic Language Models, 2002, Journal of Machine Learning Research
23. Y. Bengio, S. Bengio; Modeling High-Dimensional Discrete Data with Multi-Layer Neural Networks, 2000b; MIT Press
24. Y. Bengio, A. Courville, P. Vincent; Representation Learning: A Review and New Perspectives, 2013
25. Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin; A Neural Probabilistic Language Model, 2003
26. Y. Bengio, H. Schwenk, J. S. Senécal, F. Morin, J. L. Gauvain; A Neural Probabilistic Language Model – Studies in Fuzziness and Soft Computing, 2006
27. S. Bengio, Y. Bengio; Taking on the Curse of Dimensionality in Joint Distributions Using Neural Networks, 2000a; IEEE Transactions on Neural Networks, special issue on Data Mining and Knowledge Discovery
28. P. L. Berger, T. Luckmann; The Social Construction of Reality – A Treatise in the Sociology of Knowledge, 1966
29. M. Bickhard; Interactionism: A Manifesto, 2009
30. S. Blackmore; The Meme Machine, 1999
31. H. Blumer; Symbolic Interactionism: Perspective and Method, 1969
32. R. Boyd, P. J. Richerson; Culture and the Evolutionary Process, 1985
33. L. Breiman; Arching the Edge, 1997
34. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, et al.; Language Models are Few-Shot Learners, 2020
35. K. P. Bennet and C. Campbell; Support Vector Machines: Hype or Halleluja?, 2000
36. M. W. Chang, L. Ratinov, D. Roth, V. Srikumar; Importance of Semantic Representation: Dataless Classification, 2008
37. T. Chen, C. Guestrin; XGBoost: A Scalable Tree Boosting System, ACM SIGKDD, 2016
38. G. Chick; Cultural Complexity: The Concept and Its Measurement, 1997
39. G. Chick; Leisure, Labor, and the Complexity of Culture: An Anthropological Perspective, 1986
40. G. Chick; The Units of Culture, 2001
41. G. Chick; What's in a meme? The Development of the Meme as a Unit of Culture, 1999
42. K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio; Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, EMNLP, 2014
43. N. Chomsky; A Review of B. F. Skinner's Verbal Behavior, 1959, Readings in the Psychology of Language
44. N. Chomsky; Approaching UG from Below: Interfaces + Recursion = Language?, 2007
45. N. Chomsky; Aspects of the Theory of Syntax, 1965
46. N. Chomsky; On Cognitive Structures and their Development: A reply to Piaget, 1980
47. N. Chomsky; Syntactic Structures, 1957
48. N. Chomsky; The Galilean Challenge: Architecture and Evolution of Language, 2017
49. N. Chomsky; The language capacity: architecture and evolution, 2017
50. N. Chomsky; Tool Module: Chomsky's Universal Grammar, 1986 (ca.)
51. E. Ciavolino, R. Redd, A. Evrinomy, et al.; Views of Context. An instrument for the analysis of the cultural milieu, 2017
52. K. Clark, U. Khandelwal, O. Levy, C. D. Manning; What Does BERT Look At? An Analysis of BERT's Attention, 2019
53. K. Clark, M. T. Luong, Q. V. Le, C. D. Manning; ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators, ICLR, 2020
54. J. N. Coates, D. Bollegala; Frustratingly Easy Meta-Embedding – Computing Meta-Embeddings by Averaging Source Word Embeddings, 2018
55. W. W. Cobern, G. Aikenhead; Cultural aspects of learning science, 1997
56. K. M. Colby; Computer Models of Thought and Language, 1973, The American Journal of Psychology
57. K. M. Colby, F. D. Hilf, S. Weber, H. C. Kraemer; Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes, 1972
58. M. Cole; Cultural psychology. A once and future discipline, 1996, American Psychological Association
59. R. Collobert, J. Weston; A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning, 25th ICML, 2008
60. P. T. Costa, R. R. McCrae; NEO personality Inventory professional manual, 1992
61. F. Coulmas; The Influence of Language on Culture and Thought, 1991
62. W. Croft, A. Cruse; Cognitive Linguistics, 2004

63. R. Cullingford; Inside Computer Understanding, 1981, SAM
64. R. Cullingford; Script Application: Computer Understanding of Newspaper Stories, 1978
65. Ö. Dahl; Grammaticalization and the Life-Cycles of Constructions, 1998
66. L. Damen; Culture Learning: The Fifth Dimension on the Language Classroom, 1987
67. J. Davison; Zero-Shot Learning in Modern NLP, 2020
68. R. Dawkins; The Extended Phenotype, 1982
69. R. Dawkins; The Selfish Gene, 1976
70. R. Dawkins; The Selfish Gene (new edition), 1989
71. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman; Indexing by Latent Semantic Analysis, 1990
72. J. Delius; Of Mind Memes and Brain Bugs: A Natural History of Culture, 1989, The Nature of Culture
73. D. C. Dennett; Consciousness Explained, 1991
74. D. C. Dennett; Darwin's Dangerous Idea, 1995
75. D. C. Dennett; Darwin's Dangerous Idea: Evolution and the Meanings of Life, 1996
76. J. Devlin, M. W. Chang, K. Lee, K. Toutanova; BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018
77. W. B. Dolan, C. Brockett; Automatically Constructing a Corpus of Sentential Paraphrases, ACL, 2005
78. A. Dugin; The Noology of the Ancient Chinese Tradition, 2018, NOOMAKHIA
79. W. H. Durham; Coevolution: Genes, Culture and Human Diversity, 1991
80. W. H. Durham; Units of Culture, 1997, Human by Nature: Between Biology and the Social Sciences
81. W. H. Durham, P. Weingart; Units of Culture, 1997, Human by Nature: Between Biology and the Social Sciences
82. D. E. Durkheim; Sociology and Philosophy, 1924
83. C. Eckart, G. Young; The Approximation of One Matrix by Another of Lower Rank, 1936
84. S. Edunov, A. Baevski, M. Auli. Pre-trained language model representations for language generation, 2019;
85. T. S. Eliot; Notes Towards the Definition of Culture, 1948
86. J. L. Elman; Finding Structure In Time, 1990
87. J. Elster; Logic and Society, 1978
88. M. Ember; Evolution of the Human Relations Area Files, 1997, Cross-Cultural Research Journal
89. K. Erk, A. Herbelot; How to marry a star: probabilistic constraints for meaning in context, 2021
90. M. Everaert, M. A. C. Huybregts, N. Chomsky, R. C. Berwick, J. J. Bolhuis; Structures, Not Strings: Linguistics as Part of the Cognitive Sciences, 2015
91. D. Everett; Cultural Constraints on Grammar and Cognition in Pirahã: Another Look at the Design Features of Human Language, 2005
92. D. Everett; Dark Matter of the Mind: The Culturally Articulated Unconscious, 2016
93. D. Everett; How Language Began: The Story of Humanity's Greatest Invention, 2017
94. D. Everett; Language: The Cultural Tool, 2012
95. J. R. Firth; A synopsis of linguistic theory 1930–1955, 1957
96. A. P. Fiske; Complementarity Theory: Why Human Social Capacities Evolved to Require Cultural Complements, 2000, Pers Soc Psychol Rev
97. J. A. Fodor; The modulary of mind. An essay on faculty psychology, 1983
98. F. L. G. Frege; On Sense and Reference, 1892
99. K. Fukushima; Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position, 1980
100. F. A. Gers, E. Schmidhuber; LSTM recurrent networks learn simple context-free and context-sensitive languages, 2001
101. A. Globerson; Euclidean Embedding of Co-occurrence Data, 2007
102. X. Glorot, Y. Bengio; Understanding the Difficulty of Training Deep Feed-Forward Neural Networks, 2010
103. L. R. Goldberg; The structure of phenotypic personality traits, 1993
104. Y. Goldberg; A Primer on Neural Network Models for Natural Language Processing, 2016
105. Y. Goldberg; Assessing BERT's syntactic abilities, 2019
106. Y. Goldberg, O. Levy; Word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method, 2014
107. I. Goodfellow, Y. Bengio; A. Courville; Deep Learning, 2016
108. A. Graves; Sequence Transduction with Recurrent Neural Networks, 29th ICML, 2012
109. B. J. Grosz, A. K. Joshi, S. Weinstein; Providing a Unified Account of Definite Noun Phrases in Discourse, 1983
110. S. R. Gunn; Support Vector Machines for Classification and Regression, 1998
111. M. Hauser, N. Chomsky, W. T. Fitch; The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?, 2002, Science
112. K. He, X. Zhang, S. Ren, J. Sun; Deep residual learning for image recognition, IEEE-CVPR, 2016.
113. D. Hendrycks, K. Gimpel; Gaussian error linear units, 2016
114. J. Hewitt, C. D. Manning; A Structural Probe for Finding Syntax in Word Representations, NAACL-HLT, 2019
115. G. E. Hinton; Connectionist Learning Procedures, 1989
116. G. E. Hinton; Learning Distributed Representations of Concepts, 1986
117. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov; Improving neural networks by preventing co-adaptation of feature detectors, 2012
118. G. E. Hinton, O. Vinyals, J. Dean; Distilling the Knowledge in a Neural Network, NIPS, 2014-5
119. T. K. Ho; Random Decision Forests, 1995
120. S. Hochreiter, J. Schmidhuber; Long Short-Term Memory, Neural computation, 1997
121. C. F. Hockett; The Origin Of Speech, 1960
122. H. Hotelling; Analysis of A Complex of Statistical Variables into Principal Components, 1933
123. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, et al.; MobileNets: Efficient convolutional neural networks for mobile vision applications, 2017
124. P. M. Htut, J. Phang, S. Bordia, S. R. Bowman; Do attention heads in BERT track syntactic dependencies?, 2019
125. F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer; SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, ICLR, 2016-7
126. F. N. Iandola, A. E. Shaw, R. Krishna, K. W. Keutzer; SqueezeBERT: What can computer vision teach NLP about efficient neural networks?, 2020

127. S. Ioffe, C. Szegedy; Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, 2015

128. G. Jahoda; Quetelet and the emergence of the behavioral sciences, 2015

129. R. Jia, P. Liang; Data Recombination for Neural Semantic Parsing, ACL, 2016

130. M. Johnson; How the Statistical Revolution Changes (Computational) Linguistics, 2009

131. A. K. Joshi, S. Weinstein; Control of Inference: Role of Some Aspects of Discourse Structure-Centering, 1981

132. M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy. Spanbert: Improving pre-training by representing and predicting spans, 2019;

133. R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, Y. Wu; Exploring the Limits of Language Modeling, 2016

134. D. Jurafsky; M. H. James; Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2000

135. K. S. Kalyan, A. Rajasekharan, S. Sangeetha; AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing, 2021

136. P. Kanerva, J. Kristoferson, A. Holst; Random Indexing of Text Samples for Latent Semantic Analysis, 2000

137. I. Kant; Kritik der reinen Vernunft, 1781-7

138. M. Kearns; Thoughts on Hypothesis Boosting, 1988

139. R. Keesing; Cultural Anthropology: A Contemporary Perspective, 1976

140. A. Kilgarriff, J. Rosenzweig; English SensEval: Report and Results, 2000

141. Y. Kim, C. Denton, L. Hoang, A. M. Rush; Structured Attention Networks, ICLR, 2017

142. D. P. Kingma, J. Ba; Adam: A Method for Stochastic Optimization, ICLR 2015, 2014-7

143. O. N. E. Kjell, K. Kjell, D. Garcia, S. Sikström; Semantic measures: Using Natural Language Processing to Measure, Differentiate, and Describe Psychological Constructs, 2019

144. V. Kocijan, A. M. Cretu, O. M. Camburu, Y. Yordanov, T. Lukasiewicz; A surprisingly robust trick for winograd schema challenge, ACL, 2019

145. A. N. Kolmogorov; Selected works, 2005

146. K. Koskenniemi; Two-level morphology: A general computational model of word-form recognition and production, 1983

147. S. Kotowski, H. Härtl; Recursion and the language faculty - on the evolution of the concept in Generative Grammar, 2011, Norddeutsches Linguistisches Kolloquium

148. A. Krizhevsky, I. Sutskever, G. E. Hinton; ImageNet Classification with Deep Convolutional Neural Networks, NeurIPS, 2012

149. O. Kuchaiev, B. Ginsburg. Factorization tricks for LSTM networks, ICLR, 2017;

150. G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy; RACE: Large-scale ReAding Comprehension Dataset From Examinations, 2017

151. G. P. Lakoff; On generative semantics, 1971

152. G. P. Lakoff; Toward generative semantics, 1963 (ed. 1976)

153. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut; ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, ICLR, 2020

154. H. Larochelle, D. Erhan, Y. Bengio; Zero-data Learning of New Tasks, 2008

155. A. Lavelli, F. Sebastiani, R. Zanoli; Distributional term representations: an experimental comparison, 2004

156. M. Tan and Q. V. Le; EfficientNet: Rethinking model scaling for convolutional neural networks, ICML-PMLR, 2019

157. R. Lebret; R. Collobert; Word Emdeddings through Hellinger PCA, 2013

158. S. Legg, M. Hutter; A Collection of Definitions of Intelligence, 2007

159. C. Lemaréchal; Cauchy and the Gradient Descent, 2012

160. M. Lesk; Automatic Sense Disambiguation Using Machine Readable Dictionaries, 1986

161. D. W. Letcher; Cultoromics: A New Way to See Temporal Changes in the Prevalence of Words and Phrases, 2011

162. H. J. Levesque; The Winograd Schema Challenge, 2011

163. S. R. Levin; Langue and Parole in American Linguistics, 1965, Foundations of Language

164. D. Levinson, M. Ember; Encyclopaedia of Cultural Anthropology, 1996

165. O. Levy; Y. Goldberg; Linguistic Regularities in Sparse and Explicit Word Representations, 2014

166. O. Levy, Y. Goldberg; Neural Word Embedding as Implicit Matrix Factorization, 2014

167. O. Levy, Y. Goldberg, I. Dagan; Improving Distributional Similarity with Lessons Learned from Word Embeddings, 2015, TACL

168. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer; BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, 2019

169. T. G. Lewis; Cognitive stigmergy: A study of emergence in small-group social networks, 2012

170. Y. Li, L. Xu; Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective, 2015

171. Y. Lin, Y. C. Tan, R. Frank; Open Sesame: Getting Inside BERT's Linguistic Knowledge, ACL, 2019

172. F. T. Liu, K. M. Ting, Z. H. Zhou; Isolation-based Anomaly Detection, ACM KDD, 2012

173. F. T. Liu, K. M. Ting, Z. H. Zhou; Isolation Forest, 2008

174. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov; RoBERTa: A Robustly Optimized BERT Pretraining Approach, Repository, 2019

175. I. Loshchilov, F. Hutter; Decoupled Weight Decay Regularization, ICLR 2019, 2017-9

176. G. Luger, W. Stubblefield; Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 2004

177. H. P. Luhn; A new method of recording and searching information, 1953

178. C. J. Lumsden, E. O. Wilson; Genes, Mind, and Culture, 1981

179. A. Lynch; Thought Contagion: How Belief Spreads Through Society, 1996

180. J. Lyons; Review of Aspects of the Theory of Syntax by Noam Chomsky, 1966

181. K. Mahesh, S. Nirenburg, S. Beale, E. Viegas, V. Raskin, B. Onyshkevych; Word sense disambiguation: why statistics when we have these numbers?, 1997

182. W. C. Mann, S. A. Thompson; Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, 1988

183. A. J. Marsella, J. Dubanoski, W. C. Hamada; The measurement of personality across cultures: Historical conceptual, and methodological issues and considerations, 2000

184. L. Martin, B. Muller, P. J. O. Suárez, et al.; CamemBERT: a Tasty French Language Model, ACL, 2020

185. J. D. McCawley; Syntax and semantics 7: Notes from the linguistic underground, 1976

186. R. T. McCoy, E. Pavlick, T. Linzen; Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference, 2019

187. R. McCrum; Daniel Everett: "There is no such thing as universal grammar", 2012

188. G. H. Mead; The Individual and the Social Self: Unpublished Essays, 1982

189. T. Mikolov; Language Modeling for Speech Recognition in Czech, 2007

190. T. Mikolov, G. Corrado, K. Chen, J. Dean; Efficient Estimation of Word Representations in Vector Space, 2013

191. T. Mikolov, A. Deoras, D. Povey, L. Burget, J. Černocký; Strategies for Training Large Scale Neural Network Language Models, 2011

192. T. Mikolov, J. Kopecky, L. Burget, O. Glembek, J. Černocky; Neural Network Based Language Models for Highly Inflective Languages, 2009

193. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean; Distributed Representations of Words and Phrases and their Compositionality, 2013

194. T. Mikolov, W. T. Yih, G. Zweig; Linguistic Regularities in Continuous Space Word Representations, 2013

195. C. M. Millward; A Biography of the English Language, 1996

196. W. Mischel, Y. Shoda; Reconciling processing dynamics and personality dispositions, 1998

197. A. Mnih, G. Hinton; A Scalable Hierarchical Distributed Language Model, 2009

198. G. E. Moore; Cramming more components onto integrated circuits, 1965

199. F. Morin, Y. Bengio; Hierarchical Probabilistic Neural Network Language Model, 2005

200. R. Navigli; Word Sense Disambiguation: A Survey, 2009

201. D. Q. Nguyen, T. Vu, A. T. Nguyen; BERTweet: A pre-trained language model for English Tweets, EMNLP, 2020

202. N. Nilsson; Artificial Intelligence: A New Synthesis, 1998

203. D. Norman; Living with Complexity, 2010

204. D. Norman; The Design of Everyday Things, 2013

205. D. Norman; The Psychology of Everyday Things, 1988

206. B. P. O'Connor; A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories, 2002

207. A. Paccanaro, G. E. Hinton; Extracting Distributed Representations of Concepts and Relations from Positive and Negative Propositions

208. M. D. Pagel; Wired for Culture: origins of the human social mind, 2012

209. K. Papineni, S. Roukos, T. Ward, W. J. Zhu; BLEU: a Method for Automatic Evaluation of Machine Translation, ACL, 2002

210. Paul Michel, Omer Levy, Graham Neubig; Are Sixteen Heads Really Better than One?, NeurIPS, 2019

211. K. Pearson; On Lines and Planes of Closest Fit to Systems of Points in Space, 1901

212. C. S. Peirce; On a New List of Categories, 1867

213. J. Pennington, R. Socher, C. D. Manning; GloVe: Global Vectors for Word Representation, 2014

214. F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel; Language Models as Knowledge Bases?, 2019, FAIR

215. J. Phang, T. Févry, S. R. Bowman; Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks, 2018-9

216. S. Pinker; The Language Instinct, 1994

217. S. Pinker; The Stuff of Thought – Language as a Window into Human Nature, 2007

218. S. Pinker; Words and Rules – The Ingredients of Language, 1999, Science Masters Series

219. S. Pinker, R. Jackendoff; The faculty of language: What's so special about it?, 2005

220. C. Pollard; I. A. Sag; Head-driven phrase structure grammar, 1994

221. D. Poole, A. Mackworth, R. Goebel; Computational Intelligence: A Logical Approach, 1998

222. P. M. Postal; The best theory, 1972

223. X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang; Pre-trained Models for Natural Language Processing: A Survey, 2021

224. W. V. Quine; Word and Object, 1960

225. J. R. Quinlan; Induction of Decision Trees, 1985

226. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever; Improving Language Understanding by Generative Pre-Training, 2018

227. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever; Language Models Are Unsupervised Multitask Learners, 2019, Open AI

228. P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang; SQuAD: 100,000+ Questions for Machine Comprehension of Text, EMNLP-ACL, 2016

229. A. T. Rambo; The Study of Cultural Evolution, 1991, Profiles in Cultural Evolution

230. M. W. Ridley; The Origins of Virtue: Human Instincts and the Evolution of Cooperation, 1996

231. M. W. Ridley; The Rational Optimist: How Prosperity Evolves, 2010

232. M. W. Ridley; The Red Queen: Sex and the Evolution of Human Nature, 1993

233. J. M. Roberts; Explorations in Cultural Anthropology, 1965

234. J. M. Roberts; The Self-Management of Cultures, 1964

235. R. H. Robins; A Short History of Linguistics, 1967

236. A. Rogers, O. Kovaleva, A. Rumshisky; A Primer in BERTology: What we know about how BERT works, 2020

237. B. Romera-Paredes, P. H. S. Torr; An embarrassingly simple approach to zero-shot learning, 2015

238. J. R. Ross; On declarative sentences, 1970

239. J. Rothman; The Meaning of Culture, 2014

240. D. E. Rumelhart, G. E. Hinton, R. J. Williams; Learning internal representations by error propagation, 1985

241. S. J. Russell, P. Norvig; Artificial Intelligence: A Modern Approach, 2003

242. P. Rybak, R. Mroczkowski, J. Tracz, I. Gawlik; KLEJ: Comprehensive Benchmark for Polish Language Understanding, ACL, 2020

243. M. Sahlgren; A brief history of word embeddings, 2015

244. M. D. Sahlins; Culture and Practical Reason, 1976

245. M. D. Sahlins; Evolution and Culture, 1960

246. G. Salton; Some Experiments in the Generation of Word and Document Associations, 1962

247. G. Salton, A. Wong, C. S. Yang; A Vector Space Model for Automatic Indexing, 1975

248. S. Salvatore; Psychology in black and white. The project of a theory-driven science, 2016

249. S. Salvatore; Social life of the sign: Sensemaking in society, 2012

250. S. Salvatore, V. Fini, T. Mannarini, J. Valsiner, G. A. Veltri; Introduction to Symbolic Universes in Time of (Post) Crisis – The Future of European Societies, 2019

251. S. Salvatore, V. Fini, T. Mannarini, G. A. Veltri, E. Avdi, F. Battaglia, et al.; Symbolic universes between present and future of Europe. First results of the map of European societies' cultural milieu, 2018

252. S. Salvatore, T. Mannarini, E. Avdi, F. Battaglia, et al.; Globalization, demand of sense and enemization of the other. A psycho-cultural analysis of European societies' socio-political crisis, 2018b

253. S. Salvatore, J. Valsiner, G. A. Veltri; The Theoretical and Methodological Framework. Semiotic Cultural Psychology, Symbolic Universes and Lines of Semiotic Forces, 2019

254. S. Salvatore, T. Zittoun; Outlines of a psychoanalytically informed cultural psychology, 2011

255. V. Sanh, L. Debut, J. Chaumond, T. Wolf; DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, NeurIPS, 2019-20

256. S. Santurkar, D. Tsipras, A. Ilyas, A. Madry; How Does Batch Normalization Help Optimization?, 2019

257. R. Schank, N. Goldman, C. Rieger, C. Riesbeck; MARGIE: Memory Analysis Response Generation, and Inference on English, 1973

258. E. H. Schein; Organizational culture, 1990

259. J. Schuessler; How Do You Say "Disagreement" in Pirahã?, 2012

260. H. Schutze; Word Space, 1993

261. R. V. Scruton; A Short History of Modern Philosophy, 1982

262. R. V. Scruton; Art And Imagination: A Study in the Philosophy of Mind, 1974

263. R. V. Scruton; On Human Nature, 2017

264. R. V. Scruton; The Aesthetic Understanding: Essays in the Philosophy of Art and Culture, 1983

265. R. V. Scruton; The Politics of Culture and Other Essays, 1981

266. W. H. Sewell; Logics of History – Social Theory and Social Transformation, 2005

267. M. Shatz; On the development of the field of language development, 2007

268. N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean; Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, ICLR, 2017

269. G. Simmel; Philosophische Kultur, 1919

270. J. Sinclair; The automatic analysis of corpora, 1992

271. B. F. Skinner; The Evolution of Verbal Behavior, 1986

272. B. F. Skinner; Selection by Consequences, 1981

273. B. F. Skinner; Verbal Behavior, 1957

274. A. Smith; The Theory of Moral Sentiments, 1759

275. R. Socher, J. Bauer, C. Manning, A. Ng; Parsing with Compositional Vector Grammars, 2013

276. R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, C. Potts; Recursive deep models for semantic compositionality over a sentiment tree-bank, EMNLP, 2013

277. P. A. Sorokin; Contemporary Sociological Theories, 1928

278. P. A. Sorokin; Social and Cultural Dynamics, 1937-41

279. M. Stevenson, Y. Wilks; Combining Weak Knowledge Sources for Sense Disambiguation, 1999

280. D. Strinati; An introduction to theories of popular culture, 2004 (ed. 2012)

281. Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, D. Zhou; MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices, ACL, 2020

282. I. Sutskever, O. Vinyals, Q. V. Le; Sequence to Sequence Learning with Neural Networks, 2014

283. M. Taboada, W. C. Mann; Rhetorical Structure Theory: Looking Back and Moving Ahead, 2006

284. I. Tenney, D. Das, E. Pavlick; BERT Rediscovers the Classical NLP Pipeline, ACL, 2019

285. I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. Van Durme, et al.; What Do You Learn From Context? Probing for Sentence Structure in Contextualized Word Representations, 2019

286. E. C. Tupes, R. E. Christal; Recurrent personality factors based on trait ratings, 1961

287. I. Turc, M. W. Chang, K. Lee, K. Toutanova; Well-read students learn better: On the importance of pre-training compact models, 2019

288. J. Turian, L. Ratinov, Y. Bengio; Word Representations: A Simple and General Method for Semi-Supervised Learning, Association for Computational Linguistics, 2010

289. A. M. Turing; Computing Machinery and Intelligence, 1950

290. J. Valsiner; An invitation to cultural psychology, 2014

291. J. Valsiner; Culture in minds and societies. Foundations of cultural psychology, 2007

292. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al.; Attention Is All You Need, 2017

293. G. Veltri, R. Redd, T. Mannarini, S. Salvatore; The identity of Brexit: a cultural psychology analysis, 2019

294. C. Venuleo, P. G. Mossi, S. Salvatore; Educational subcultures and dropping out in higher education. A longitudinal case study, 2016

295. C. Venuleo, S. Salvatore, P. G. Mossi; The role of cultural factors in pathological gambling, 2015

296. T. Verheggen, C. B. Baerveldt; We Don't Share! Exploring the theoretical ground for social and cultural psychology: The social representation approach versus an enactivism framework, 2007

297. J. Vig; A Multiscale Visualization of Attention in the Transformer Model, ACL, 2019

298. C. F. Voegelin; Review of Noam Chomsky, Syntactic Structures, 1958

299. L. S. Vygotsky; Mind in society, 1978

300. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman; GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, ICLR, 2018-9

301. A. Warstadt, A. Singh, S. R. Bowman. Neural network acceptability judgments, 2018;

302. C. Wei, S. M. Xie, T. Ma; Why Do Pretrained Language Models Help in Downstream Tasks?, 2021

303. J. Weizenbaum; Computer Power and Human Reason: From Judgment to Calculation, 1976

304. B. Widrow, R. Winter; Neural Nets for Adaptive Filtering and Adaptive Pattern Recognition, IEEE, 1988

305. N. Wiener; Cybernetics: Or Control and Communication in the Animal and the Machine, 1948

306. N. Wiener; The Human Use of Human Beings, 1950

307. R. Williams; Culture and Society, 1958

308. R. Williams; Keywords: A Vocabulary of Culture and Society, 1983 (ed. 2012)

309. R. Williams; The Long Revolution, 1965

310. A. Williams, N. Nangia, S. R. Bowman; A broad-coverage challenge corpus for sentence understanding through inference. NAACL-HLT, 2018

311. E. O. Wilson; Consilience: The Unity of Knowledge, 1998

312. L. J. J. Wittgenstein; Philosophical Investigations, 1953

313. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, et al.; Transformers: State-of-the-art Natural Language Processing, 2019

314. W. Yin, J. Hay, D. Roth; Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach, 2019

315. Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, et al.; Large Batch Optimization for Deep Learning: Training BERT in 76 minutes, ICLR, 2020

316. W. Zhang, K. Itoh, J. Tanida, Y. Ichioka; Parallel distributed processing model with local space-invariant interconnections and its optical architecture, 1990

317. A. Zhila, W. T. Yih, C. Meek, G. Zweig, T. Mikolov; Combining Heterogeneous Models for Measuring Relational Similarity, 2013, NAACL HLT

318. X. Zhang, X. Zhou, M. Lin, J. Sun; ShuffleNet: An extremely efficient convolutional neural network for mobile devices, CVPR, 2018

319. T. Ziemke, J. Zlatev, R. R. Frank; Body, language and mind – Volume 1: Embodiment, 2007

320. G. K. Zipf; Human Behaviour and The Principle of Least Effort, 1949 (ed. 2012)

321. T. Zittoun; Transitions: Development through symbolic resources, 2006