Quantitative Methods in the Humanities and Social Sciences

Arjuna Tuzzi *Editor*

Tracing the Life Cycle of Ideas in the Humanities and Social Sciences

Springer

Quantitative Methods in the Humanities and Social Sciences

Quantitative Methods in the Humanities and Social Sciences is a book series designed to foster research-based conversation with all parts of the university campus – from buildings of ivy-covered stone to technologically savvy walls of glass. Scholarship from international researchers and the esteemed editorial board represents the far-reaching applications of computational analysis, statistical models, computer-based programs, and other quantitative methods. Methods are integrated in a dialogue that is sensitive to the broader context of humanistic study and social science research. Scholars, including among others historians, archaeologists, new media specialists, classicists and linguists, promote this interdisciplinary approach. These texts teach new methodological approaches for contemporary research. Each volume exposes readers to a particular research method. Researchers and students then benefit from exposure to subtleties of the larger project or corpus of work in which the quantitative methods come to fruition.

More information about this series at http://www.springer.com/series/11748

Arjuna Tuzzi
Editor

Tracing the Life Cycle of Ideas in the Humanities and Social Sciences

Springer

Editor
Arjuna Tuzzi
Department of Philosophy, Sociology, Education
and Applied Psychology
University of Padova
Padova, Italy

ISSN 2199-0956 ISSN 2199-0964 (electronic)
Quantitative Methods in the Humanities and Social Sciences
ISBN 978-3-319-97063-9 ISBN 978-3-319-97064-6 (eBook)
https://doi.org/10.1007/978-3-319-97064-6

Library of Congress Control Number: 2018956144

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Some years ago, I made, together with my students, some experiments aimed to test the Piotrowski-Altmann law on textual data from newspapers. The Piotrowski-Altmann law explains and describes the dynamics of the spread of new elements in a language and the dynamics of how elements of a language disappear. The formula which represents this law is

$$p(t) = \frac{1}{1 + ae^{-bt}}$$

It can be obtained as the solution to a differential equation which describes the dynamics of language change as a function of time. Apparently, the parameter b represents the velocity of change and can be interpreted as a bunch of linguistic and extralinguistic factors. The results of these tests gave perfect support to the hypothesis on language change and showed various forms of temporal behaviour of the function. Some words were on the increase; others could be observed while they were losing momentum. A special group reached a peak within 1 day and started decreasing the next day. Of course, there was no hope to single out the individual factors which contributed to the empirical values of the parameters and thus to a detailed interpretation of our results. We were happy enough with the empirical support to the law and a catalogue of several progression forms we found and could interpret in individual cases.

When Arjuna Tuzzi told me that she was planning a project based on distant reading using a quantitative approach aimed at data on the "history of ideas" in several scientific disciplines, I was not very optimistic at a first thought. It was clear that the search of such a history of concepts was methodologically very similar to the dynamics of linguistic elements because the concepts, or ideas, as taken from texts, are found in the form of terms in texts. I remembered my impressions from the experiments with my students. The results were excellent from a pure scientific point of view but did not look useful with respect to a chance to apply them. But then I thought: "What about if someone smarter than I am turned the process the

other way round? Starting from one or two extra-linguistic factors and analysing the frequency dynamics of words or chunks found in the texts?". This was exactly the idea behind Arjuna Tuzzi's plan. And now I became enthusiastic.

A member of the scientific community has always some knowledge about his/her discipline: there are concepts, research questions, pioneers and important personalities, significant publications, debates and controversies, leading paradigms, failures and many more, which an informed colleague will be familiar with. On the other hand, no one is able to cover a discipline totally. The older a discipline, the harder a good picture on the basis of individual descriptions will be. After some decades, even a relatively young science becomes not even remotely comprehensible by a single person. Young colleagues are not yet able to gain an overview; older ones are less open to new developments. Thus, personal knowledge of a discipline is always incomplete and biased. A more complete picture can be obtained, of course, by reading as many relevant original books and articles as possible. This would become a project for decades, while the corresponding discipline keeps changing. Such a situation calls for statistics—the only method to collect reliable information in spite of fragmentary data. The project Arjuna Tuzzi was talking about suddenly seemed to provide the only possible way to achieve a "history of ideas" in several disciplines from texts and other data sources.

Now, I am tracking the project with rapt attention.

University of Trier
Trier, Germany

Reinhard Köhler

Contents

Contributors

Michele A. Cortelazzo University of Padova, Padova, Italy

Pierdaniele Giaretta University of Padova, Padova, Italy

Giuseppe Giordan University of Padova, Padova, Italy

Stefano Ondelli University of Trieste, Trieste, Italy

Pasquale Pavone Università degli Studi di Modena e Reggio Emilia, Modena, Italy

Valentina Rizzoli University of Padova, Padova, Italy

Chantal Saint-Blancat University of Padova, Padova, Italy

Stefano Sbalchiero University of Padova, Padova, Italy

Giuseppe Spolaore University of Padova, Padova, Italy

Matilde Trevisani University of Trieste, Trieste, Italy

Arjuna Tuzzi Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Padova, Italy

Giovanni Urraci University "Ca' Foscari" Venice, Venice, Italy

Abbreviations

AJS	American Journal of Sociology
ASA	American Sociological Association
ASR	American Sociological Review
ATD	Analysis of textual data
CA	Correspondence analysis
CC	Curve clustering
ECU	Elementary context units
ETD	Emerging topic detection
EASP	European Association of Social Psychology
EJSP	European Journal of Social Psychology
EDA	Exploratory data analysis
ETD	Emerging topic detection
FD	Functional data
FDA	Functional data analysis
FPCA	Functional principal component analysis
GCV	Generalized cross-validation
HDP	Hierarchical Dirichlet process
IE	Information extraction
IR	Information retrieval
JASA	Journal of the American Statistical Association
JPSP	Journal of Personality and Social Psychology
KWIC	Keyword in Context
KBS	Knowledge-based system
LNRE	Large number of rare events
LDA	Latent Dirichlet allocation
LSI	Latent semantic indexing
ML	Machine learning
MWE	Multiword expression
MI	Mutual information
NLP	Natural language processing
POS	Part-of-speech

PLSA	Probabilistic latent semantic analysis
PASA	Publications of the American Statistical Association
QASA	Quarterly Publications of the American Statistical Association
RE	Regular expression
RMS	Root mean square
SVD	Singular value decomposition
TM	Text mining
TDT	Topic Detection and Tracking
WSD	Word sense disambiguation

Chapter 1
Introduction: Tracing the History of a Discipline Through Quantitative and Qualitative Analyses of Scientific Literature

Arjuna Tuzzi

Contents

Abstract The chapters of this book are concerned with learning of the evolution of ideas (theories, concepts, methods, and application domains) and of the history of a discipline, by means of the temporal evolution of word occurrences in papers published by scientific journals. The work carried out for each of the areas involved in the project (philosophy, sociology, psychology, linguistics, statistics) pursued different objectives: to obtain a first overview of the relationship between time and contents in order to observe latent temporal patterns; to identify relevant keywords; to cluster keywords portraying similar temporal patterns; to identify latent dynamics of cluster keywords; and to identify relevant topics as groups of related words. The

A. Tuzzi (✉)
Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Padova, Italy
e-mail: arjuna.tuzzi@unipd.it

A. Tuzzi (ed.), *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*, Quantitative Methods in the Humanities and Social Sciences,
https://doi.org/10.1007/978-3-319-97064-6_1

contributions identified and analysed the main subject matters that, at the time of publication, were considered relevant by mainstream journals and offer new viewpoints to read and understand the evolution of a discipline. The interdisciplinary debate triggered by this research work is innovative because quantitative methods for text analysis have been used in areas of human and social sciences, which are traditionally studied through qualitative approaches, and also represents a positive experience since new paths have been explored by pooling together the qualitative and quantitative research methods, traditions, and expertise of different disciplines.

Keywords History of ideas · Quantitative methods · Qualitative methods · Statistical analysis of textual data · Diachronic corpora · Scientific literature

1.1 Quantitative Methods and History of Ideas

Quantitative methods for the analysis of scientific literature have already been utilized by a variety of disciplines and the growing availability of large databases calls for fresh methods to deal with emerging problems, to open the door to different questions, and to lead to new knowledge.

The main aim of this book is to uncover the opportunities of learning the evolution of the ideas of a discipline through a distant reading (Moretti 2013) of the contents conveyed by relevant scientific literature. The temporal evolution of ideas (theories, concepts, methods, and application domains) has been explored by means of the temporal evolution of the occurrences of "words" (with particular reference to "keywords", e.g. technical words, scientific terms, proper names) included in papers published by mainstream, leading, scientific journals. Quantitative methods, statistical techniques, and software packages are used to identify and study the main subject matters, both in the past and today, of a discipline from raw textual data. From a theoretical viewpoint, the book also aims at dealing with a concept of "quality of life" of words over time and at fostering a debate about the popularity of ideas rather than dealing merely with the problem of dating their birth (that represents one of the main concerns in the study of the history of languages).

The experiment is innovative because quantitative methods for text analysis have been used in areas of human and social sciences which are traditionally explored through qualitative approaches. The chapters show that from the point of view of different areas (philosophy, sociology, psychology, linguistics, and statistics), it is possible to obtain an effective (distant) reading of large amounts of scientific articles and that quantitative methods can work successfully alongside qualitative methods in the study of the history of a discipline. However, we are aware that these achievements represent only a first step in an immense, boundless field, and much work remains to be done.

This book reports a new development of a research study conducted by a small group of Italian scholars who worked together on an interdisciplinary research project funded by the University of Padova, *Tracing the History of Words. A Portrait of a Discipline Through Analyses of Keyword Counts in Large Corpora of Scientific*

Literature, that considered the analysis with quantitative methods of corpora of scientific literature. Following a preliminary study on the history of statistics (Trevisani and Tuzzi 2015), the project dealt with philosophy and sociology. In a second stage, the research group grew up and embarked in a new project: included further disciplines (social psychology and Italian linguistics), exploited new methods for the analysis of textual data, and it is still going on in the effort of joining new groups.

Those who study the history of ideas of a discipline (e.g. the history of philosophical ideas, history of sociological ideas) usually rely on a long tradition of research that brings together the most influential and authoritative readings of what was in the past and what is now the life cycle of ideas. In all areas of knowledge, a history of ideas and a narrative of what has been the evolution of the disciplines have been developed from a theoretical, epistemological, and methodological point of view. But what normally happens is that this history becomes, over time, a history told in the light of what is known ex-post and, as a result, today we have a "representation" of the history of the milestones experienced by a discipline that has been reworked, revised, and corrected in the light of the results that the proposed ideas have had in subsequent years. Furthermore, prominent scholars of the past have in particular been the subject of studies on the history of disciplines, who certainly have a following, but when the focus is on the ideas of the "Great Names", a great deal of the brainwork that has prepared the ground for great discoveries is overlooked.

At least for the major contributions of contemporary scientific and intellectual panorama, there is an alternative option of reading the history of a discipline through the works published in journals. Scientific journals are the new Agora for the exchange of ideas and for the dissemination of research results, and because they represent a written and documented legacy, it is possible to read the timings of scientific debate across the temporal sequence of the publications.

In this book, we propose a reading of the history of some disciplines through a distant reading of the contents conveyed by the articles published in mainstream journals. We are aware, in turn, that this is not "the" representation of the history of a discipline, but a narration from a particular viewpoint, which reflects the historical moment in which the journals under consideration were published. It is well-known that not everything that has been published has the same importance in the history of a discipline and the same influence in the instruction of future generations, but what is published leaves a trace and to some extent has been taken into account by the scientific community of that specific historical moment. We also know that scientific discoveries and innovative ideas are not published when they are brought to light by their creators but only with some delay, that is, after they have been accepted and evaluated by the scientific community of that historical moment and after considerable publication times. In addition, authors did not begin their work when it becomes published (especially when it comes to mainstream journals) and even the most famous scholars may have struggled to find time at the beginning of their careers or when they had proposed ideas too innovative, original, or outside the box.

This way of reading the history of a discipline allows us to bring out what at a specific time was deemed relevant, either because it was a trend at that time, or because it was lauded by the editor and/or board of the journal, or because it was deemed of a high level by referees, etc. What we present is, therefore, an identification of the main ideas (theories, concepts, methods, and fields of application) that, at the time of publication, were relevant to the most influential journals and the dominant scientific communities linked to them.

1.2 Tracks on the Ground: What Methods for What Purposes

Grounded in a field that is closed to the perspective of distant reading (Moretti 2013), the projects exploited computational methods for text analysis and created a shared theoretical and practical framework to achieve innovative data-driven findings across different disciplines. Since this research group operated in an interdisciplinary framework, the state of the art cannot be "just one" as it is essentially specific of each discipline and of each approach. As a consequence, the traits to draw a general background of a desirable link among quantitative methods, qualitative methods, and history of ideas will be tackled through the chapters of this book with specific reference to each discipline. From a methodological viewpoint, we can try to trace only a sort of general and brief background of the methods adopted, without any pretence of completeness.

Even though quantitative linguistics enjoys a long tradition, the "modern" idea of quantitative analysis of textual data (ATD) emerged in the 1980s (Beaudouin 2016; Bolasco 2005, 2013; Lebart et al. 1998). A number of scientific and cultural approaches as well as theoretical schools and research instruments have developed since then and today a sheer size of research fields are hard to distinguish and systematize (e.g. see Wang et al. 2018; Léon and Loiseau 2016; Kelih et al. 2016; Tuzzi et al. 2015; Mayaffre et al. 2016; Mikros and Mačutek 2015; Née et al. 2014; Obradović et al. 2013; Naumann et al. 2012; Dister et al. 2012; Köhler 2011, 2012; Popescu et al. 2009; Popescu 2009; Baayen 2001). Moreover, branches of research that today are recognized as separate disciplines (quantitative linguistics, computational linguistics, text mining, stylometry, digital methods for text analysis, etc.) have some common roots and over time they differentiated in terms of methods, aims, and objects of research. During the last decades, research activities in this field have experienced a rapid development, and this process has fostered a renewed interest for topics related to text analysis and new methods to achieve a distant reading of large amounts of texts. Many approaches for text mining (Aggarwal and Zhai 2012; Berry and Kogan 2010; Berry 2004; Sanger and Feldman 2007; Kao and Poteet 2007; Sahami and Srivastava 2009; Sullivan 2001; Weiss et al. 2005) combine linguistic concepts, computational methods, information technologies, statistical learning, and machine learning to analyse texts. The field is highly interdisciplinary and it is constantly growing.

A diachronic corpus is a collection of texts including information on their timings (e.g. the publication date of an article). Scientific journals represent a useful ground for studying the development of scientific language and topics since we can assume that the evolution of word occurrences reflects the evolution of the corresponding concepts (Trevisani and Tuzzi 2015, 2018; Popescu and Strapparava 2014; Chavalarias and Cointet 2008, 2013; Guérin-Pace et al. 2012; Hall et al. 2008; Salem 1988, 1991). In quantitative linguistics, a number of textual features can be observed as sequences of linguistic properties (Mikros and Mačutek 2015; Köhler and Galle 1993) and the problem of reading the evolution of a phenomenon over time is often tackled by resorting to linguistic laws (Köhler 2011; Tuzzi and Köhler 2015) or time series analysis (Pawłowski 2006, 2016; Pawłowski et al. 2010). From a statistical viewpoint, a word trajectory hardly shows a regular behaviour and requires special attention since in diachronic corpora data are typically sparse over time (an unavoidable feature of textual data known as the "large p, small n" problem; see, for example, Hastie et al. 2008; Tibshirani et al. 2015; Johnstone and Titterington 2009; Lebart et al. 1984).

When diachronic corpora are collections of scientific literature reference can be made to methods based on scientometrics (see, for example, Yin and Wang 2017; Cobo et al. 2011, 2012; Porter and Rafols 2009; Small 2006) and also methods for content mapping based on occurrences and co-occurrences of words (see, for example, Guérin-Pace et al. 2012; Tuzzi 2012; Maggioni et al. 2009; Cretchley et al. 2010; Michel et al. 2011; Van Den Besselaar and Heimeriks 2006; Cahlík and Jiřina 2006; Bhattacharya and Basu 1998) and clustering for assessing significant changes (Zhang et al. 2016, 2017; Koplenig 2017; Gries and Hilpert 2008, 2012; Hilpert and Gries 2009; Diwersy and Luxardo 2016) prove useful. A relevant research area is topic modelling, that starting with the seminal work of Blei and Jordan (2003), has been further developed by Griffiths and Steyvers (2004) that introduced a Latent Dirichlet Allocation (LDA) generative model to discover topics covered in the corpus (Hall et al. 2008). Topic modelling connects to scientometrics and an interesting overview, also from an epistemological perspective, has been provided by Chavalarias and Cointet (2008, 2013). In this volume also an alternative way for identifying topics provided by Reinert's method (Ratinaud and Marchand 2012; Reinert 1983, 1990, 1993) and mainly developed in social sciences is exploited for the analysis of scientific literature.

In order to shape the history of individual words, a functional data analysis approach (Ramsay and Silverman 2005) is adopted and clustering methods for functional data are used to identify groups of keywords portraying similar temporal patterns. Two approaches to curve clustering are, in principle, viable: model-based and distance-based. The former is usually founded on finite mixture models and Gaussian processes for distributions (James and Sugar 2003; Jacques and Preda 2014a) although mixed effects models (Coffey et al. 2014; Giacofci et al. 2013; Trevisani and Tuzzi 2015) and non-Gaussian distributions (Lee and McLachlan 2013) or, within the Bayesian framework, Dirichlet processes (Angelini et al. 2012; Rodriguez et al. 2009; Ray and Mallick 2006) can be assumed for mixture components. In this volume, we opted for a distance-based approach as one of our

objectives was to set up an exploratory and mostly automated learning procedure to be integrated in a so-called knowledge-based system (Trevisani and Tuzzi 2018), that is a computer system capable of generating knowledge by a large-scale integration of data, information as well as knowledge from different sources (linguistic and specific subject matter expertise), and endowed with a user-friendly interface. Within distance-based methods k-means type clustering algorithms have been widely applied to functional data especially when combined with finite basis-expansion approaches. Further choices, which extend the classical k-means algorithm with functional data, are also available (see Jacques and Preda 2014b; Wang et al. 2016).

1.3 Tracing the History of Words: A Quantitative Way

The quantitative perspective adopted by this research is essentially based on words and word counts (i.e. it is lexical based and refers to a "bag of words" approach), and, in particular, on the presence, absence, and occurrence over time of keywords relevant to the study of a specific discipline. Occurrence is an imperfect measure of the relevance of a word, however, with regard to scientific journals, we know to handle a textual genre in which language tends to be precise and succinct. In particular, titles and abstracts of the articles are extremely short, thick, and concise: They include keywords, scientific terms, technical words, nouns (e.g. research objects), proper names (e.g. authors), and often nothing else. Consequently, the fact that certain words are present or absent, and that they occur more frequently in certain historical periods and rarely in others gives us important information on the evolution of ideas and also on how to represent them.

All of text corpora exploited in the contributions of this book are written in English or Italian but the proposed methods can be extended for applications to any other language. However, each language envisages specific technical measurements and precautions that are heavily language-dependent, particularly to fulfil the phases of text pre-processing and processing (tokenization, cleaning, identification of multiword expressions, part of speech tagging, etc.).

An important assumption of this research is that the temporal course of word occurrences is viewed as a proxy of word diffusion and vitality, i.e. a word's life cycle. We assume, therefore, that the individual trajectories of words reflect the relevance through time of the corresponding ideas in the scientific discourse. Moreover, the research projects aimed at achieving interpretations of these findings. The fact of wanting to observe the trajectories drawn over time by occurrences of words also opens an interesting theoretical perspective that concerns the study of the difference between the first occurrence and the "settlement" of a given word. The research objective is not only to date the birth of subject matters but also to study their "fortunes" and fates. Moreover, to introduce the unprecedented concepts of "quality of life" of words and "life cycle" of ideas. The idea of "shaping the history of words" (Trevisani and Tuzzi 2015, p. 1288) is markedly unusual in linguistics and the study of the history of a language. Research in these areas focuses

on the problem of dating the birth (first appearance) of a word and to study the possible semantic shift. Rarely do they care about the fate, or the eventual disappearance of a word.

Research has faced partially unexplored territory and has been shown to have great potential both from a theoretical point of view and from that of the application fields. Textual data retrieved from large corpora pose interesting challenges for any data analysis method and today represent a growing area of research in many fields. New problems emerge from the growing availability of large databases and new methods are needed to retrieve significant information from those large information sources.

1.4 Objectives and Procedures

As previously stated, the quantitative analysis adopted by these research studies is essentially based on words and word counts as part of the "bag of words" approach and, in particular, is based on the occurrences over time of the most significant keywords for the study of a specific discipline.

By *occurrence*, reference is being made to the number of repetitions of a word in a corpus of texts, usually expressed as the relative frequency (or rate) as compared to the size of the texts.

By *keywords*, reference is being made to a set of words (e.g. *theory*) and of word sequences (e.g. *theory of knowledge*) that have been identified by means of specific automatic (or semi-automatic) recognition procedures relevant to the study of a specific discipline. The keywords represent theories, concepts, scientific terms, technical terms, proper names, etc.

The work carried out for each of the disciplines involved in the projects pursued the following objectives:

1. To select relevant journals and compile suitable text corpora,
2. To obtain a first overview of the relationship between time and contents in order to verify the existence of a latent temporal pattern (correspondence analysis),
3. To identify relevant keywords,
4. To cluster keywords portraying similar temporal patterns and to identify latent dynamics of cluster keywords (curve clustering),
5. To identify relevant topics as groups of related words (topic detection).

1.4.1 Selection of Journals and Corpus Description

First of all, the experts of the disciplines selected journals to work on taking into account their reputation and centrality to the discipline. When possible, it was decided to go back to the year of the journal's founding. The texts were collected in a corpus (see Chap. 7) through a phase of text harvesting, which consists of downloading information (authors, title, year, volume, issue, number of pages and,

if available, abstract) from public websites of these journals, through repositories and also resorting to printed versions. It was necessary to merge several sources because those available are not always complete and accurate as you would expect and it was necessary to look for other sources and resort to printed versions of journals to be able to fill the gaps and to collect all items. For some insights, it was decided to work with selected articles in full text (see Chaps. 2 and 5), also retrieved from the printed version of the journals.

These first raw data were processed to obtain a detailed overview of the available material, for information about the period of observation and to get a description in terms of the number, frequency, regularity and size of volumes, and issues of the journal. Similarly, the number, frequency, regularity and size of the articles, titles, and abstracts were examined.

In this phase, depending on the discipline being studied, decisions on the possible selection of items were also taken. It should be kept in mind that the journals not only publish scientific articles in the strict sense, some of the items retrieved from archives are not articles (e.g. *List of publications, News, Reviews*); some of them do not include content words in the title or in the abstract (e.g. *Comment, Rejoinder*) and, since many of them are works from the past, often they do not have abstracts.

At the end of the text harvesting, there is a diachronic corpus, i.e. a collection of texts including information on their time period, e.g. the publication date of an article. These texts might be arranged into groups (subcorpora) that refer to the same time interval, thus generating a temporal sequence of text sets. In our case, for each different journal we have diachronic corpora of texts (titles, abstracts, full texts) that are grouped by volume, which usually corresponds to groups by year of publication (with some exceptions in which volumes and years do not coincide).

A first assessment of the size of the corpora was made on the basis of words count "as they appear in the texts" (in technical terms this is called *forms* or graphical forms), that is, words are defined simply as sequences of characters of the alphabet isolated in the text by means of separators (spaces and punctuation). The recognition phase of words in the texts is technically called "tokenization" and is followed by a phase of "cleaning" that, in our specific case, essentially consists in removing the upper case and in the recognition of proper names. Other forms of tagging are used in the later phases, for example, the "part-of-speech" (POS), which serves to assign to each word, the lemma, and the grammatical category, or the "stemming", which serves to attribute more words to the same root (or "stem").

Once recognized and counted, the words can be divided into occurrences, or word-tokens, and in distinct words, or word-types. The frequency of a word-type is the number of corresponding word-tokens, the number N of word-tokens is the size of the corpus in terms of occurrences, the number V of word-types is the size of the corpus in terms of different words and the set of word-types represents the vocabulary (or word list) of the corpus. The observation of the word list and frequencies leads to first considerations of the most frequent words of the corpus. In addition, at the end of this phase it is possible to make some initial assessments on the length of available titles, abstracts, and articles.

1.4.2 Correspondence Analysis (CA)

Correspondence analysis (CA) is an explorative data analysis (EDA) that can be used to create content mapping and that, in this research, has served to reconstruct the general system of relationships among years, among words and between years and words (Greenacre 1984, 2007; Murtagh 2005, 2010, 2017; Lebart et al. 1984, 1998). The CA is based on the word list (vocabulary) that for each word comprises occurrences in the different volumes/years, representing our reference time-points. It recognizes similarities and differences through the lexical profiles, that is, through the frequencies of words over time: two words are similar (and close on the graph) because they have been used with similar frequency in the same time-points (volumes); two time-points are similar (and close on the graph) because the volumes of the journal at that time used the same words with similar frequency.

Although the CA is not able to definitively describe all relevant features of a large corpus, exploring relationships between words and years contributes to obtain an effective and immediate (distant) reading of the main contents and to distinguish features otherwise hardly perceptible with the sequential reading (close-reading) of texts. In the graph generated by the CA, it is possible to immediately verify whether the volumes of the journal have experienced an evolution of the contents over time. In fact, if the journal had a clear chronological development its volumes can be seen as a line on the plane that respects the order in time (see, for example, Fig. 2.1, Chap. 2). Only if a journal displays a chronological development of its contents does it make sense to use it for the study of the history of ideas of a discipline. For this reason, we considered this first EDA as consistent with our aims. For our processing, the words that exceeded a frequency threshold in the vocabulary of the corpora were selected. The threshold chosen was based on the coverage rate of the corpus and the coverage rate of the vocabulary.

Correspondence analysis is still widely used in the analysis of textual data. Since CA offers a way to achieve a Euclidean embedding of different information spaces based on cross-tabulation counts, it handles multivariate numerical and symbolic data with ease and proves useful to analyse great masses of textual data. From this perspective, it can be exploited in information semantics, and particularly in "big data" settings, to collect relationships. Murtagh (2010) showed how to work with CA as a Semantic Analysis Platform and through further experiments (Murtagh 2017) that involved data analysis in very high-dimensional spaces showed the benefits of CA as a tool suitable for carrying out latent semantic or principal axes mapping in big data scaling.

Since CA is a well-known and established method, this book does not include a specific chapter on this topic (see Greenacre 1984, 2007; Murtagh 2005; Lebart et al. 1984, 1998). However, the Appendix of this chapter provides a brief introduction to understand the rationale of CA.

1.4.3 Identification of Keywords

> "Within the perspective of analysis of textual data and, more in general, in all cases of data collection based on text harvesting, the construction phases, first of the corpora and then of textual data, are essential moments: choices made before statistical analysis are crucial to guarantee the quality of data" (Trevisani and Tuzzi 2015, p. 1289).

There are many examples of statistical analysis of textual data in literature that simply take into account the most frequent words in the corpus (or the most frequent n-grams, n-word-grams, etc.) and grammatical words are usually excluded from the word sets to be analysed (often referred to as "stop words"). We preferred not to follow these procedures systematically because they do not sufficiently take into account redundancies, compounds, and ambiguities. We preferred to retrieve all the relevant keywords in semi-automatic mode and identify content words and sequences of words (compounds, multiword expressions, segments) that are relevant to the study of a discipline. To this end, we also adopted a procedure for the automatic recognition of multiword expressions (MWEs, see Chap. 8) as well as the intersection of the corpus vocabulary with discipline-specific glossaries, and, when it was possible, also the index of keywords for the retrieved papers available in the online databases. It is worth mentioning that in any text we have sequences of words that have different meanings if they are considered alongside adjacent words, i.e. when we read them from a keyword in context (KWIC) perspective. The observation of the occurrences of MWEs increases the amount of information conveyed by keywords because a sequence of words partly reduces the noise and disambiguate the meaning. Moreover, semantic changes and semantic shifts of a word over time should envisage also the appearance of new MWEs and, when these new objects become relevant in a scientific language, their occurrences in publications should increase as well (and they should be retrieved by our procedures). We select nouns, names, MWEs, and through a matching with the most appropriate glossaries of each discipline we verified whether we collected all relevant keywords.

1.4.4 Curve Clustering

The purpose of curve clustering is to cluster keywords portraying similar temporal patterns and to identify latent dynamics of cluster keywords. This approach assumes that the "shape" of each keyword's trajectory in the volumes of mainstream journals, as it has been drawn by the keyword's occurrences over time, reflects the relevance of the corresponding subject matters in the scientific discourse.

In the frame of functional data analysis (FDA) approach, the proposed method consists of a two-stage functional clustering approach for statistical learning: first, a filtering step in which functional data are represented as smooth functions by a basis-expansion method (B-splines), second, a distance-based curve clustering in which the k-means algorithm is used combined with a metric to measure the distance between curves. Before filtering, a crucial choice concerns how to properly

normalize word raw frequencies. Lastly, interpretation by expert opinion to decipher detected dynamics and thus compose a narrative of the evolution of the discipline is the conclusive step of the learning process (see Chaps. 6 and 9).

The main difference between this approach and the one represented by topic detection methods is that this analysis seeks to study the individual micro-histories and identify groups of keywords that, regardless of whether they belong to similar or completely different topics, have experienced the same temporal evolution over time. The main advantage is being able to effectively visualize trajectories that may represent interesting developments: words that, over time, have shown a growing trend, words that were in vogue in the past and have shown a decreasing trend, words which enjoyed a golden age in a period of great popularity and then disappeared, and words that live alternating phases of presence and absence, etc. (see, for example, Fig. 6.3, Chap. 6; Fig. 9.6, Chap. 9).

For our computations, all the keywords with a high enough frequency to produce a distinguishable trajectory and in a large number (but limited, usually to around 1000) were chosen to represent all the most relevant subject matters of a journal.

1.4.5 Topic Detection

Topic detection methods have objectives and assumptions very different from those of curve clustering. In this approach, the topics are identified as lists of related words that have in common the fact of co-occurring in texts. From a diachronic perspective, the temporal evolution of a topic does not depend on the trajectories of the words that compose it (i.e. taken individually) but by the "weight" that the frequencies of these words (as a whole) have comprehensively in different periods of time. The main advantage of topic detection with respect to curve clustering is that this approach makes it possible to automatically identify which are the main topics that emerge in the corpus without having to select a priori a list of keywords upon which to concentrate the analysis.

For our purposes, we used two different methods: Reinert's method and Latent Dirichlet Allocation (LDA). The two methods are both valid and, in many cases, have been preferred only on the basis of better readability of the results (see Chap. 10). Reinert's method is based on occurrences of words in texts and a similarity measure (chi-square distance). A descending hierarchical cluster analysis is performed on a distance table, which generates classes of units that best differentiate the vocabulary: it extracts classes of words that co-occur and that are best differentiated from other classes. Latent Dirichlet Allocation (LDA) depends on a topic-modelling algorithm, i.e. it bases itself on a generative statistical model that assumes the existence of a generative process: documents are generated by first picking a distribution over topics; words are generated by picking a topic from this distribution and then picking a word from that topic. For modelling the contributions of different topics to a document, LDA treats each topic as a probability distribution over words, viewing a document as a probabilistic mixture of these topics.

1.5 Chapters Outline

The volume is divided into two parts: there are five chapters in the first part for each of the disciplines involved (Chap. 2 philosophy, Chap. 3 sociology, Chap. 4 psychology, Chap. 5 linguistics, and Chap. 6 statistics) with the results of the analyses of corpora; the second part is dedicated to four methodological insights, which describe the methods used for processing the data (Chap. 7 compiling and pre-processing corpora, Chap. 8 MWE identification, Chap. 9 curve clustering, Chap. 10 topic detection). A concluding chapter summarizes the main findings and, perhaps first and foremost, the challenges of this interdisciplinary research work (see Chap. 11). Although the chapters of the first part of the book discuss the history of different disciplines and the contents of several journals, they are very similar to each other from the point of view of the approach and methods adopted because the work within the research group had been coordinated.

All contributions were primarily focused on verifying the existence of temporal patterns in the chronological textual data sets (diachronic corpora) and demonstrating by statistical analysis that the evolution of the contents of journals actually follows a chronological pattern. In fact, although it may seem a reasonable a priori condition, the existence of a temporal pattern in the evolution of the contents of a journal cannot be taken for granted and should be verified through processing of the data. In fact, a corpus which is diachronic in the structure does not always show a clear temporal pattern of its contents (Cortelazzo and Tuzzi 2007) and even a few exceptions were found in this research. For example, when we dealt with journals that arrange all their publications into special issues (that are focused on specific topics).

For the study of the temporal evolution of the contents of journals, methods to study the temporal trajectories of individual words were used as well as methods for topic detection. Depending on the discipline, the results obtained were compared with the ("content-metric") analysis of several journals. The possibility to work with titles, with abstracts, or with full-text articles was evaluated on a case-by-case basis with a focus on habits and scientific writing traditions of the relevant discipline and with reference to practices and traditions of the specific journal subject to analysis.

1.6 About This Book

In our research activities, we often have to take into account a large number of scientific papers in order to trace the history of a discipline and the temporal development of ideas in a specific field. There are definitely too many texts for one scholar to read in a lifetime. Instead of close-reading a limited number of texts, we have the opportunity to work with thousands of texts, uploading them into the memory of a computer and ask a software package to produce analyses and results. A software package (and a statistical model behind it) cannot "close read" a text. On the contrary, by means of mathematical and statistical tools, it might be smart enough to