JCOM

BIG DATA AND DIGITAL METHODS IN SCIENCE COMMUNICATION RESEARCH: OPPORTUNITIES, CHALLENGES AND LIMITS

Computer-aided text analysis: an open-aired laboratory for social sciences

Yurij Castelfranchi

Abstract

Thanks, on the one hand, to the extraordinary availability of colossal textual archives and, on the other hand, to advances in computational possibilities, today the social scientist has at their disposal an extraordinary laboratory, made of millions of interacting subjects and billions of texts. An unprecedented, yet challenging, opportunity for science. How to test, corroborate models? How to control, interpret and validate Big Data? What is the role of theory in the universe of patterns and statistical correlations? In this article, we will show some general characteristics of the use of computational tools for the analysis of texts, and some applications in the areas of public communication of S&T and Science and Technology Studies (STS), also showing some of their limitations and pitfalls.

Keywords

Public understanding of science and technology; Representations of science and technology; Science communication: theory and models

The information deluge has flooded social sciences. Governments around the world digitalize and make available laws and projects, parliamentary debates and questions, social and economic indicators, party funding, government programs, polls and public opinion polls, maps, and models, due to new governance practices (and the new rhetoric) of transparency, accountability, participation and deliberation. Dynamics of globalized markets and customer demand force large and small media outlets (newspapers and magazines, TVs and radios, network on the Internet and on social networks) to make it possible to download huge databases, including archives of what was written, said, sung over a period of decades.

Publishers publish their books in digital format. Libraries and newspaper libraries, as well as corporations such as Google, file millions of publications containing a considerable amount of literature and world history. New forms of sociability, on and offline, new modes, effects and affections of contemporary individualism lead tens of millions of users to illustrate, narrate, edit, and caption on web 2.0 their tastes and intimacies, their daily life and their memories, the interactions, associations, conflicts, dreams, resentments and confessions. We are exposed to millions of gigabytes of information and, at the same time, we are tireless producers of further gigabytes of digital trails about our movements, our consumption, our inventions, our imagination. Far from being an Orwellian scenario, we are spontaneous watchers of the lives and traces of others, and accomplices of our electronic surveillance and self-surveillance.

Thanks to the information deluge, dreams and nightmares of social scientists are fulfilled at the same time. On the one hand, an unprecedented, extraordinary, global laboratory of social sciences, a colossal experiment in the open, became possible. Tens of millions of volunteer subjects can be studied, live or in their history, in their interactions, their conflicts, their sociability, their rituals, their deliberative processes. Models can be tested, corroborated, modified, rejected. New hypotheses can be formulated. On the other hand, this data (and the models derived from it) is difficult to control, analyse, interpret, validate, refute, and what is worse, according to some, it can make social sciences "irrelevant" and the scientific method itself "obsolete". Would theories and interpretations be useless, in this era of correlations and intelligent algorithms, of pattern recognition? With this article, I would like to present the general characteristics and some applications for the social studies of S&T and science communication about the use of computational tools for the analysis of texts, showing some of the promises, limits and dangers of the computational euphoria linked to big data.

Computer-aided text analysis

Although the idea of applying computational techniques to the production, translation, and comprehension of natural language was one of the first tasks and promises for the discipline of Artificial Intelligence [Castelfranchi and Stock, 2000], and although since the postwar period there have been programs for automated analysis of texts, only in recent years can we say that computing has become a crucial tool for the social scientist, especially thanks to two important developments [Wiedemann, 2013; Wiedemann, 2015]. On the one hand, automatic or semi-automatic analysis of texts has become an increasingly interesting option for the researcher thanks to the ease of access and the solid growth of the amount of texts in digital format. The ecosystem of the big textual data today available to social scientists is extraordinarily diverse: interviews, public speeches, questionnaire responses, online forums, blogs and microblogs, newspaper articles, user comments on topics or posts on social networks, laws and projects, parliamentary debates, patent applications, process transcripts, agreements and contracts, encyclicals, not to mention the gigantic bases filed in digital libraries and newspaper libraries.

On the other hand, computer-assisted text analysis (and, more generally, the use of software to aid social science research) has had an extraordinary boost due to recent advances in both the computational power of machines and the sophistication of mathematical models and algorithms for the collection, extraction, visualization, analysis and interpretation of data. In recent decades, new statistical approaches have provided the social scientist with tools that are particularly suited to the study of the complexity of qualitative variables and human language. At the same time, the increasing computing power and the decreasing cost of computers and software programs, which are more and more sophisticated, flexible and user-friendly, have made search-assistant machines ideal for many researchers.

Today, software for qualitative and quantitative analysis go far beyond helping us to mark interesting passages of texts, or count how many times words or concepts appear in a *corpus*. They allow us to explore not only the text but its context (geographic, temporal, meta-data), and investigate not only the explicit content of the message but also its latent dimensions and the semantic aspects of language. By recognizing patterns, correlations, identifying concepts, topics, parts of language, actors and their relationships, the software also assists with interpretive analyses.

Thus, since 1980s and 1990s, especially in the previously less-explored area of computer-aided qualitative data analysis (CAQDA — computer assisted qualitative data analysis), software packages (such as MAXQDA, ATLAS.ti, NVivo, ALCESTE, etc.) started to spread, and they rapidly were enriched with tools that were also mixed, qualitative and quantitative (such as QDAMiner and WordStat, Rapidminer, or modular projects, such as Knime, or very interesting open-science research efforts like CorText).

In less than twenty years, there was an exponential growth in the number of research group networks, specialized courses and manuals, collections and reference texts dedicated to the subject as a whole [Popping, 2000], to the use of computers in qualitative approaches [Kuckartz, 2014] or in more traditional areas such as content analysis [Riffe, Lacy and Fico, 2014; Neuendorf, 2016]. There was also an extensive academic production dedicated to the interdisciplinary territory ("E-humanities", "Digital humanities", etc.), which emerged at the intersection between humanities, social sciences and computing [Schreibman, Siemens and Unsworth, 2016].

Typologies and processes of computer-assisted text analysis

There are several ways to use computational aids, depending on the data available, the theoretical approach and objectives and research hypotheses. However, computer use begins before any analysis and methodological choice. Algorithms can work as inexpensive helpers for the collection and processing of texts. Scraping software from websites, portals and RSS feeds, or robots that download entire databases, or archive and organize the traces left by social networking users can be used in order to collect data (see for example Bucchi and Neresini [2011] and Neresini and Lorenzet [2016], and the media monitor "TIPS", described by Neresini in this issue of JCOM). In addition, Optical Character Recognition (OCR) programs are increasingly effective in transforming scanned images into digital texts that are editable and analysable, and provide the researcher with rich, almost ready-to-use empirical material.

Following this zero phase of the research, the computer is a powerful aid for data pre-processing. The software help to eliminate those prepositions, articles, adverbs, adjectives that are not relevant to the research, or to perform procedures of stemming and "lemmatisation", whereby different words are grouped together but refer to the same term (for example, because they are the singular, plural, feminine or masculine version of the same word) or to the same semantic field (for example, synonyms). In addition, in the analysis phase, the computer helps in measuring the reliability of the coding and in the validation of the data and results.

Finally, in the analysis of texts as such, the software helps in at least three different ways, depending on the deductive or inductive approach and on the type of operation of text encoding — automatic or manual (for a taxonomy of the types of computer-aided analyses see, for example, Wiedemann [2013] and Pollach [2012]).

1. Manual coding, categories created by researchers

This is the most traditional case whereby researchers are expected to read, categorise, analyse and interpret the text content, encoding and classifying the analysis units on the basis of categories they have previously chosen (for example, on the basis of previous hypothesis and theories, or with an inductive-deductive approach, such as in the Grounded Theory). Within this deductive approach, with categories that have been established *a priori*, the role of the researcher and the training of codifiers are crucial, however, the computer helps organise, visualise, retrieve the coded traces, and produce tables and data which are structured on the basis of work done by human beings. Both in quantitative and qualitative text analyses, as well as in interpretative approaches, the work carried out to categorise, write down, interpret texts and write about them can be assisted with CAQDA software (such as NVivo, Atlas.Ti, etc.) which creates intelligent, structured catalogues of the work done by the human researcher.

2. Automatic coding, categories created by researchers

When categories for the analysis are established *a priori*, or created on the basis of a first interpretation of the content, algorithms can be used to help code the data, especially when working on huge *corpora*. For example, the common way to carry out text analysis is to produce a hierarchy of analysis categories included in a codebook, which describes in detail what to code, how and when. This procedure tends to be expensive and needs to be refined through subsequent rounds of coding, which are alternated by (annoying) inquiries of inter-coder reliability. When the *corpus* of text is huge, then the task becomes impossible. In Computational Content Analysis, attempts are made to resolve this problem by making sure that analysis categories are simple enough (or superficial) so that coding can be done automatically. The researcher can, for example, build a dictionary made of groups of keywords which refer to a certain category (for example, aggressive feelings, scientific jargon, etc.).

3. Automatic coding and computer-built categories

One problem occurring with automatic coding is that it is limited to the most superficial aspect of language (sets of key words), and largely ignores the context, the meaning of the texts. Computational aid can help the researcher give more thickness and robustness, even in an automatic analysis, in the case of an inductive approach, in which the categories of analysis are not chosen previously but from recurring patterns, correlations, themes or concepts, detected by computational means and that were not necessarily present in the hypothesis of the research.

Examples of this inductive construction of categories are approaches such as automatic speech analysis, or lexicometry, and the so-called *corpus* linguistics, in which the computational phase of the work precedes the interpretation [Pollach, 2012; Cheng et al., 2008]. More generally, text mining is precisely the attempt to make a semantic analysis of the texts and their structure [Wiedemann, 2013], extracting the "sense" by means of statistical methods that identify intrinsic characteristics of the *corpus* along with more interpretative aspects introduced by human coders. On the one hand, then, the software shows, in an inductive way, patterns and correlations (for example, between concepts, objects, or themes) present in the texts. On the other hand, researchers can mark important features of speech, or provide sets of examples to train the machine, and learning algorithms can infer rules and learn to encode texts.

The software identifies quantitative, statistical relationships between different parts of the text. This path, as opposed to classical content analysis (in which the researcher first encodes, and then the software calculates quantities), allows to identify structures, recurrences, patterns in the text that have not necessarily been imagined *a priori*, which makes this type of analysis interesting even for researchers traditionally linked to purely interpretive approaches, such as discourse analysts and post-structuralists. Software such as Alceste, Wordsmith and TextQuest were developed with this type of approaches in mind [Wiedemann, 2013] and are rather widespread.

Computer-aided analyses in Public Communication of Science and Social Studies of S&T

In sociology, anthropology, political science, history and, of course, in the area of communication, there has recently been a proliferation of research that has harnessed the potential of computational tools applied to text analysis. This has also been the case in social studies of S&T as well as in research about public communication of science and technology. The examples are already more than could be mentioned here, and we aim to show a few examples of the diversity of applications and approaches possible.

Semino et al. [2005] used semantic text analysis software for the study of the use of metaphors in scientific communication. They compared a *corpus* of internal communication of science between peers (extracted from the journal Nature Immunology), with a corpus of public communication (articles on the same subject matter taken from a major popular science magazine, the New Scientist). In fact, they found out different uses and functions for metaphors in the case of dissemination and in the case of specialized communication. Science and technology in the media are also being studied with the aid of software, including the case of research that was traditionally approached in an interpretive and qualitative way, like the analysis of frames. Tian and Stewart [2005], for example, addressed frame analysis in the case of coverage of the SARS crisis, whereas Bail [2016] studied frames about organ donation. Crawley [2007] used computer aid in a qualitative analysis about the coverage of biotechnologies in agriculture in community daily news. Thanks to the use of specifically built dictionaries and others available on the emotional and cognitive dimensions of language, Castelfranchi, Massarani and Ramalho [2014], identified that the discourse of scientific dissemination in important Brazilian TV programs was marked by metaphors of war, aggression, and heavily marked by gender inequalities.

However, the interest in studying large *corpora* with the aid of software packages is not confined to the material that circulates in the media. Some research has focused on the analysis of spontaneous responses to open-ended questions in opinion polls, which was empirical material often underutilized precisely because of the high cost of analysis. Stoneman, Sturgis and Allum [2013] applied statistical clustering techniques to the analysis of the spontaneous description of the meaning of the term "DNA" provided by subjects in a poll. Tvinnereim and Flø ttum [2015] have already identified recurrent themes in the opinions about climate change also in answers to open questions. They used a relatively recent and innovative technique, Structural Topic Modelling, which allows to detect latent themes in the statements. Topic modelling approaches were also used to analyse themes in large journalistic coverage *corpora* [Jacobi, Atteveldt and Welbers, 2016].

Also studying climate change, Farrell [2016] used textual data and analysis of social networks to investigate the influence that organizations receiving corporate funding have on the polarisation of views, while Veltri and Atanasova [2015] applied an automatic thematic analysis, together with an analysis of semantic networks, to study the sharing of information through twitter. Veltri and Suerdem [2013], used a mixed approach, which hybridises methods of automatic categorization of texts and codification by humans, in the study of the discourse on GMOs in Turkey. A combination of human coding and automatic classification for content analysis (through the use of DiscoverText software) was also attempted to study an online activism case against the controversial practice of "fracking" [Hopke and Simis, 2015].

Transcriptions of focus group discussions (a tool often used to explore attitudes and perceptions about science, technology and innovation) can also be analysed using computational tools as Miltgen and Peyrat-Guillard [2014] did to investigate generational and cultural influences perceptions about privacy.

Our own area of PCST studies was also studied reflexively with similar techniques, for example analyzing the production in one of the largest journals in the area — "Public Understanding of Science" [Bauer and Howard, 2012; Suerdem et al., 2013; Smallman, 2016].

Success and promises

The advantages of software aid for text analysis are obvious, not just for those interested in classical and quantitative approaches. Qualitative analysis is facilitated and enhanced by the use of CAQDA software. Machine learning tools, pattern detection and dictionary creation have already proved to be valuable both for interpretive studies and for discoveries of latent aspects of *corpora*, which are not always easy to identify by hermeneutics, nor by content analyses focussing on the explicit message. However, beyond the increasing possibilities of data collection and analysis, the arrival of algorithms in the social scientist's office can be a new *humus*, capable of stimulating theoretical and methodological innovations, and encouraging the creation of more ambitious and mixed methods. The computer comes, in a sense, as ambassador of a partial truce in the hardly fecund controversy between proponents of "quali" and "quanti" in the analysis of the texts. On the one hand, computing is making the boundary between the two more fuzzy, forcing researchers to better reflect on mixed-method searches, and investing in interdisciplinary teams. Although text mining is based on statistical and computational tools, it cannot be accused of being synonymous with a reductionist approach, or criticised, based on a caricatured *cliché*, of being the son of a "positivist epistemology", as it assists the "qualitative" researcher in the extraction of meaning and context, and, at the same time, in the validation and quantitative study of the relevance and reliability of the analysis performed [Wiedemann, 2015]. Many nowadays believe in the complementarity of qualitative (e.g. grounded theory) and quantitative approaches (such as computational content analysis), and in the usefulness of a theoretical and methodological deepening on mixed models and triangulation [Kuckartz, 2014]. The increasing, though still incipient, ability of topic modeling techniques, computational linguistics, lexicometry, etc., to investigate the semantic level, the context of enunciation, the place of speech and the meaning of the text, allows to reduce the distance between what a researcher interested in interpreting their object of study seeks to do and what a machine can contribute to.

However, the debate over the influence that software use may have on the research process itself is still open and fierce, for example because software is accused to make people incorporate forms, knowledge entities, analytic units, predetermined thinking styles such as thinking the speech and the communication in terms of coded units and hierarchies of codes, of predefined and formatted relations between entities. We run the risk of seeing only what we are already looking for, of knowing just what the software is programmed to detect [Wiedemann, 2013]. And, more importantly, we must not forget that having new tools and techniques is not the same as having invented new methodologies. And accessing more data is not synonymous with having solved theoretical problems.

Challenges and limits

The herald of the idea that the "data flood" came to us, bringing a new era, was the editor of "Wired" magazine, Chris Anderson. And the new era, said Anderson euphorically, is that of the "end of theory" [Anderson, 2008]. The journalist said that children of the "Petabybte Age", companies like Google, no longer need sociologists, nor models or hypotheses: indexing, classifying, archiving data, and detecting their patterns, regularities, dynamics, that's all. Understanding and interpreting is no longer needed. Forget the theory, whatever the theory is. It does not matter why humans and human groups behave the way they behave. It only matters to track, record, measure what they do, and *voilá*, "with enough data, the numbers will speak for themselves." The scientific method itself, the formulation of causal hypotheses, the construction of models, the tests, the experiments would then be obsolete. Knowing what facts, phenomena, and behaviours are correlated would be more than sufficient: algorithms will be responsible for finding patterns and predictions for human behavior, where theories and models have never succeeded. May science open wings for the arrival of Google.

Predictably, these and other statements from Big Data enthusiasts have generated a huge amount of controversies, which we do not intend to deal with here. However, it is worth mentioning briefly the dangers and pitfalls of the idea that the data and the patterns *per se* are knowledge. This problem has at least one epistemological dimension, one technical and one markedly political.

Because, if the data is important, models, hypotheses and interpretations are even more important. For science and for decision-making. Considering patterns and predictions as the new synonym of the word truth means to make a data driven policy (i.e. a policy without policy) no longer the place of choice and conflict about living well in common. In response to Anderson, the writer James Bridle [2016] states, for example, that “the belief in the power of data [. . .] leads [. . .] to a belief in the truth of data-derived assertions. And if data contains truth, then it will, without moral intervention, produce better outcomes.” The second dimension of the problem is technical. By relying on data mining, we run the risk of forgetting, or disregarding, that we do not always know how to estimate the errors in the new statistical models applied to the analysis of texts and, in more general terms, to social processes. According to Grimmer and Stewart [2013], the automatic methods can give relevant results only when complemented by the interpretation and reading of the material by the researchers. And not only: when comparing different methods, the same authors conclude that such methods, in the current state of the art, suffer from serious problems of validation. Errors and pitfalls in the use of computer-aided analysis are striking, and we do not yet have adequate methodologies to deal with automation.

The sociologist Neal Caren [2015] has a similar view: the euphoria about Big Data is followed by the enthusiasm to incorporate new statistical and computational methods, but it is early to assess to what extent such tools are scientifically robust and heuristically fertile. The ability of a machine learning model to provide predictions does not necessarily mean to have found meaningful causal variables to explain a social phenomenon, or to actually understand its functioning.

Finally, outside the problems of validity, robustness and reliability, there is something deeper in the controversy. The role of method, theory, and what we mean by explanation and science.

Massimo Pigliucci, a philosopher of science, also criticises Anderson’s simplistic triumphalism, and asks himself: “If we stop looking for models and hypotheses, are we still really doing science?” [Pigliucci, 2009]. Finding patterns is only part of the scientific practice, which is completed by seeking explanations for the patterns found. Without models, be they conceptual or mathematical ones, data, according to the philosopher, are nothing but noise, science advances only when it can provide explanations.

We agree. We do not need to be apocalyptic about the “dangers” of the Big Data, we may welcome its arrival as the largest laboratory the social sciences can dream of, but we should not be euphoric to the point of imagining that data can provide the solution to the foundational dilemmas of the social sciences, or for the lack of good theories and good models. There is no doubt: the availability of huge textual *corpora* represents an unprecedented opportunity for scientists, an indispensable resource. And computer aid provides extraordinary, valuable news. However, neither the data nor the algorithms are in themselves a new science or an answer to the scientific questions. A barometer and a faithful helper who notes down numbers in a notebook do not do physics and do not figure out how the climate works, although they can identify patterns and make predictions. In the same way, statistical regularities and petabytes are maps — good maps — of dynamic phenomena. But having beautiful colored maps does not mean knowing a country. Phenomena need not only to be described and portrayed, but also explained and, as Max Weber said a long time ago, understood and interpreted in their sense.

Acknowledgments

The author would like to express his gratitude to the CNPq for its support through the Scholarship of Productivity in Research, and to Alisson Soares, Brunah Schall, Gabriela Reznik from the InCiTe Observatory for their suggestions and productive discussions.

Translated by Sabina Brusemini

References

Anderson, C. (2008). 'The end of theory: The data deluge makes the scientific method obsolete'. *Wired* 16 (7). URL: https://www.wired.com/2008/06/pb-theory/.

Bail, C. A. (2016). 'Cultural carrying capacity: Organ donation advocacy, discursive framing, and social media engagement'. *Social Science & Medicine* 165, pp. 280–288. DOI: 10.1016/j.socscimed.2016.01.049. PMID: 26879407.

Bauer, M. W. and Howard, S. (2012). 'Public Understanding of Science — a peer-review journal for turbulent times'. *Public Understanding of Science* 21 (3), pp. 258–267. DOI: 10.1177/0963662512443407.

Bridle, J. (1st November 2016). 'What's wrong with big data?' *New Humanist*. URL: https://newhumanist.org.uk/articles/5104/whats-wrong-with-big-data.

Bucchi, M. and Neresini, F. (2011). 'Monitoring Science in the Public Sphere: The Case of Italy'. In: The Culture of Science. Ed. by M. W. Bauer, R. Shukla and N. Allum. New York, U.S.A.: Routledge.

Caren, N. (2nd April 2015). 'The Path to Big Data Sociology Isn't Obvious'. *Mobilizing Ideas*. URL: https://mobilizingideas.wordpress.com/2015/04/02/the-path-to-big-data-sociology-isnt-obvious/.

Castelfranchi, Y., Massarani, L. and Ramalho, M. (2014). 'War, anxiety, optimism and triumph: a study on science in the main Brazilian TV news'. *JCOM* 13 (3), A01. URL: https://jcom.sissa.it/archive/13/03/JCOM_1303_2014_A01.

Castelfranchi, Y. and Stock, O. (2000). Macchine come noi. La scommessa dell'intelligenza artificiale. Bari, Italy: Laterza.

Cheng, A. S., Fleischmann, K. R., Wang, P. and Oard, D. W. (2008). 'Advancing social science research by applying computational linguistics'. *Proceedings of the American Society for Information Science and Technology* 45 (1), pp. 1–12.

Crawley, C. E. (2007). 'Localized Debates of Agricultural Biotechnology in Community Newspapers: A Quantitative Content Analysis of Media Frames and Sources'. *Science Communication* 28 (3), pp. 314–346. DOI: 10.1177/1075547006298253.

Farrell, J. (2016). 'Corporate funding and ideological polarization about climate change'. *Proceedings of the National Academy of Sciences* 113 (1), pp. 92–97. DOI: 10.1073/pnas.1509433112. PMID: 26598653.

Grimmer, J. and Stewart, B. M. (2013). 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts'. *Political Analysis* 21 (3), pp. 267–297.

Hopke, J. E. and Simis, M. (2015). 'Discourse over a contested technology on Twitter: A case study of hydraulic fracturing'. *Public Understanding of Science* 26 (1), pp. 105–120. DOI: 10.1177/0963662515607725.

Jacobi, C., Atteveldt, W. van and Welbers, K. (2016). 'Quantitative analysis of large amounts of journalistic texts using topic modelling'. *Digital Journalism* 4 (1), pp. 89–106. DOI: 10.1080/21670811.2015.1093271.

Kuckartz, U. (2014). Qualitative Text Analysis: A Guide to Methods. Los Angeles, London, New Delhi, Singapore and Washington: SAGE Publications Inc.

Miltgen, C. L. and Peyrat-Guillard, D. (2014). 'Cultural and generational influences on privacy concerns: a qualitative study in seven European countries'. *European Journal of Information Systems* 23 (2), pp. 103–125. DOI: 10.1057/ejis.2013.17.

Neresini, F. and Lorenzet, A. (2016). 'Can media monitoring be a proxy for public opinion about technoscientific controversies? The case of the Italian public debate on nuclear power'. *Public Understanding of Science* 25 (2), pp. 171–185. DOI: 10.1177/0963662514551506.

Neuendorf, K. A. (2016). The Content Analysis Guidebook. Los Angeles, London, New Delhi, Singapore and Washington: SAGE Publications Inc.

Pigliucci, M. (2009). 'The end of theory in science?' *EMBO Reports* 10 (6), p. 534. DOI: 10.1038/embor.2009.111. PMID: 19488038.

Pollach, I. (2012). 'Taming Textual Data: The Contribution of Corpus Linguistics to Computer-Aided Text Analysis'. *Organizational Research Methods* 15 (2), pp. 263–287. DOI: 10.1177/1094428111417451.

Popping, R. (2000). Computer-Assisted Text Analysis. Los Angeles, London, New Delhi, Singapore and Washington: SAGE Publications Inc.

Riffe, D., Lacy, S. and Fico, F. (2014). Analyzing Media Messages: Using Quantitative Content Analysis in Research. New York, U.S.A. and London, U.K.: Routledge.

Schreibman, S., Siemens, R. and Unsworth, J. (2016). A New Companion to Digital Humanities. Chichester, West Sussex, U.K.: John Wiley & Sons.

Semino, E., Hardie, A., Koller, V. and Rayson, P. (2005). 'A computer-assisted approach to the analysis of metaphor variation across genres'. In: Corpus-based Approaches to Figurative Language. Ed. by J. Barnden, M. Lee, J. Littlemore, R. Moon, G. Philip and A. Wallington. Birmingham, U.K.: University of Birmingham School of Computer Science, pp. 145–153.

Smallman, M. (2016). '*Public Understanding of Science* in turbulent times III: Deficit to dialogue, champions to critics'. *Public Understanding of Science* 25 (2), pp. 186–197. DOI: 10.1177/0963662514549141.

Stoneman, P., Sturgis, P. and Allum, N. (2013). 'Exploring public discourses about emerging technologies through statistical clustering of open-ended survey questions'. *Public Understanding of Science* 22 (7), pp. 850–868. DOI: 10.1177/0963662512441569. PMID: 23825238.

Suerdem, A., Bauer, M. W., Howard, S. and Ruby, L. (2013). 'PUS in turbulent times II — A shifting vocabulary that brokers inter-disciplinary knowledge'. *Public Understanding of Science* 22, pp. 2–15. DOI: 10.1177/0963662512471911.

Tian, Y. and Stewart, C. M. (2005). 'Framing the SARS Crisis: A Computer-Assisted Text Analysis of CNN and BBC Online News Reports of SARS'. *Asian Journal of Communication* 15 (3), pp. 289–301. DOI: 10.1080/01292980500261605.

Tvinnereim, E. and Flø ttum, K. (2015). 'Explaining topic prevalence in answers to open-ended survey questions about climate change'. *Nature Climate Change* 5 (8), pp. 744–747. DOI: 10.1038/nclimate2663.

Veltri, G. A. and Suerdem, A. K. (2013). 'Worldviews and discursive construction of GMO-related risk perceptions in Turkey'. *Public Understanding of Science (Bristol, England)* 22 (2), pp. 137–154. DOI: 10.1177/0963662511423334. PMID: 23833021.

Wiedemann, G. (2013). ‘Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences’. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 14 (2).

— (16th November 2015). ‘Text Mining #1: Extending the method toolbox: text mining for social science and humanities research’. *Europeana Research*. URL: http://research.europeana.eu/blogpost/extending-the-method-toolbox-text-mining-for-social-science-and-humanities-research.

Author

Yurij Castelfranchi graduated in quantum physics at the University of Rome, becoming later a professional science writer and journalist for 15 years. In 2002, he moved to Brazil, where he became a scholar in science communication and earned his PhD in sociology of S&T. As a writer, he collaborated with newspapers, radios, TVs, and magazines, and authored 6 books, as well as educational and multimedia products. As a scholar, he has been a researcher at Labjor (Laboratory of Advanced Studies in Journalism, Unicamp, Brazil) and a collaborator of OEI (the Organization of Iberoamerican States for Science and Culture). Today, he is associate professor of Sociology at the Federal University of Minas Gerais (Brazil), where he coordinates “InCiTe” (Innovation, Citizenship, and Technoscience Research Group). He is a member of the National Institute of Science and Technology for Science Communication. In 2015, he has been awarded, as one member of the winning team in science communication, the Mercosur Prize for Science and Technology. In 2016, he has been awarded a prize, by the State Foundation of S&T (Fapemig), for his contributions in promoting the advancement of science and public participation. E-mail: ycastelfranchi@gmail.com.

How to cite

Castelfranchi, Y. (2017). ‘Computer-aided text analysis: an open-aired laboratory for social sciences’. *JCOM* 16 (02), C04_en.

ISSN 1824-2049. Published by SISSA Medialab. jcom.sissa.it