Qualitatively mapping a research front through word-correspondence textual analysis: a case study

Stefano Castelli, Alessandro Pepe, Loredana Addimando

Dipartimento di Psicologia, Università di Milano-Bicocca, Italy

Abstract

This paper presents word-correspondence textual analysis as a useful tool to shed new light on the scientific production of a certain academic research group, university or other institution. The main rationale behind the study is that bibliometric assessment of scientific production is a precious and unique instrument to objectively detect regularities in the structure of a given scientific field. The *European Research Network About Parents in Education* (ERNAPE) is a rather informal (but thriving and productive) research group in the field of education, specifically focused on parents' role in schools. In this exploratory case study, word-correspondence analysis has been used as a mean to describe the process of building scientific production. To this end, authors investigate how emphasis on different aspects in the field of parents in education has evolved across different countries of the network. For this purpose, all abstracts stored in ERNAPE's electronic archive in the period from 2003 to 2007 (N=228) have been collected and analyzed. Evidences from correspondence analysis highlight how topics and keywords used by researchers reflect their cultural traditions and their representation of parental roles in education. For example, in Mediterranean countries researches are mainly focused on "family", while in Nordic Countries "civil" reasons for the home-school relationship are emphasized.

Keywords: word-correspondence analysis, bibliometric methods, mapping of research front

1. Introduction

The field of *parents in education* covers a wide range of topics concerning the area of relationships between parents and the school-systems. The field has evolved to cover multifaceted issues that shape school-family-community relationships. Within this perspective, the *European Network About Parents in Education (ERNAPE)* started in 1993 with the purposes of promoting: (1) awareness about *parents in education* themes, (2) interaction with other researchers and (3) quality researches in this field. ERNAPE has a very loosely organised structure, with no "legal" status, but with a remarkable productivity: it proved capable of organizing seven international conferences, involving researchers from many countries and from very different cultural backgrounds, and promoted the very first international journal on this topic (for details, see www.ernape.net). ERNAPE's perspective is a very complex one, in which pedagogical, psychological, sociological, anthropological, political and political spheres naturally intermingle. Then, the whole research group can be considered a network in which some overlaps between "content of research" and the "social aspects of researchers" appear.

The present paper uses the case of ERNAPE's written production to study the way in which word-correspondence textual analysis can be used in conceptual mapping. Our results try to identify how the emphasis put on different research topics into the field of parents in education vary across different "research traditions".

Bibliometrics helps us in assessing scientific research (Van Raan, 2003) but it is even more important when it comes to evaluating the evolution of universities, research group and public institution production performance. Are there "hidden" patterns in scientific production? Are there some links between the intrinsic nature of the researches and other variables? How can powerful statistical multivariate methods of data reduction be applied to mapping a research front?

As Small (2006: 595) stated: «By research area I mean a set of documents or other bibliometric units that define a research topic and an associated group of researchers who shared an interest in the topic. The definition of research area involves both content and social aspects. […] Thus, both cognitive and social forces work in tandem and reinforce each other: the cognitive generating the social and the social generating the cognitive». Moreover, a "research front" can be defined as «an emergent and transient grouping of concepts and underlying research issues» (Chen, 2006: 359).

The concept of *parents in education* represents a rather new and multidisciplinary "research front", and some kind of numerical analysis (i.e. scientometrics or bibliometric analysis) can be applied to analytically and synthetically evaluate changes in emerging themes and research areas in a given dataset of publications. Such studies, aimed at "measuring" the size of science or mapping research fronts, offer researchers, practitioners and policy makers crucial information about the state-of-the-art in a certain research field.

2. ERNAPE's landscape: bibliometric mapping of the research front

The main idea of the present work is based on the concept that word-occurrence networks can reveal the structure of publication contents in terms of the strength of linkages between pairs of selected targets (Bhattacharya and Basu, 1998).

A *research area* is defined as a group of *research fronts*, and a single *research front* is composed of a set of *research themes*. A *research theme* can be identified by using results from word-occurrence analysis of a sample of articles (Cahlik and Jirina, 2006) because words used for describing a study (as in the case of abstracts) represent both the contents of the paper and the basic bricks of research themes. According to Van den Besselaar and Heimeriks's (2006) perspective, researchers acting in a given field share a common knowledge base which is reflected in the selection of specific terms. Through their lexical choices researchers reproduce their identity at an aggregated level. Thus, a systematic analysis of the most often used words and linkages among them in a set of published abstracts represents a means of reconstructing past trends and emerging topics across countries or over time. Consequently, we expect that researchers belonging to the same research paradigms, cultural contexts or historical traditions will exhibit similar terminological choices.

Considering these assumptions reasonable, we can try to find the empirical basis for Birte Ravn's (2002; 2003; 2005) idea that the way educational systems consider parents as partners has evolved with political, industrial, cultural and historical developments, as a result giving a sort of differentiation of the concept on a geographical basis. From Ravn's point of view, in Catholic dominated parts of Europe families are viewed as in charge of the development of their children. In Nordic countries of Europe the welfare state is responsible for education, administration and local decision-making. In these countries the home-school partnership is interpreted in terms of developing democratic citizens (Ravn, 2005). Finally, in the US and UK the parent-school partnership focuses on children's academic achievement as a result of a hierarchic school system based on the individual pupil. As far as the two maps (the map by

Ravn and the one independently generated by our dataset on the basis of word-counts) tend to overlap, we have good reason to believe in her results. The process is called "triangulation".

3. Dataset, methodology and limits

In order to answer the above research questions, we collected data from ERNAPE's archive which contains all the papers published in ERNAPE's conference proceedings from 2003 to 2007. The dataset uses textual data retrieved from all abstracts published by the ERNAPE network (N=228). The textual corpus results in 47.748 tokens, 5.090 words and 3.793 lemmas. Each abstract is composed of approximately 200 tokens (mean =209). Finally, we extracted the 635 most frequently occurring words (at least 8 occurrences as the threshold value) and used them to generate some contingency frequency tables, arranged according to different variables. With the chosen threshold value, the analysis covers over 70% of the whole number of occurrences found in texts.

Since we are interested in mapping changes in ERNAPE's scientific output at an aggregated level, we need to group similar contexts together to obtain a geographical perspective on published articles. A debatable, but not totally unreasonable way to do this without loss of information, is to group countries regionally, using the following criteria:

AREA 1: Northern European Countries (i.e. Denmark, Sweden, Norway, Finland);
AREA 2: Mediterranean Countries (i.e. Portugal, Spain, Italy, Greece, Cyprus);
AREA 3: Western European Countries (i.e. the Netherlands, France, Germany, Austria, Belgium);
AREA 4: Eastern European Countries (i.e. Poland, the Czech Republic, Serbia, Estonia);
AREA 5: Anglophone Countries (i.e. the United States, the United Kingdom, Canada);
AREA 6: Others (a residual area, not investigated further in this paper).

If analysts want to "test" hypotheses about the state-of-the-art of a certain research front, they need some tools to detect connections between a set of two or more (usually categorical) variables and map the distinctive features in a given unit of time or space. In this case, we chose to use multidimensional techniques of data reduction (i.e. correspondence analysis, CA) rather than simple graphical approaches to avoid loss of information during the analysis (Dore et al., 1996). As Anuradha and Urs (2007: 183) remark «Correspondence analysis has several advantages over other methods of analysis: it was specifically designed to compare profiles; it is a multidimensional method that achieves appropriate data reduction, filters out noise, and objectifies correlations among variables». As a consequence of these characteristics, CA has gained the fame of a necessary tool for data analysis in nearly all disciplines (Beh, 1999). It is worthwhile to note that CA mainly provides a visual graphic output (multidimensional maps) that is easier to grasp than series of numbers (Benzecri, 1992). More precisely, correspondence analysis typically generates numerical indexes (i.e. eigenvalues, coordinates on principal axes, squared cosine values, etc.) from which a set of low-dimensional maps can be built. Maps represent one form of visual communication that synthetically depicts the degree of association among objects.

Of course, researchers must be aware of potentials and limitations of bibliometric indexes. To this end, it is relevant to remember that the papers analysed in our study represent *only* the output presented during the ERNAPE conferences. Obviously, these data *do not* capture *all* contributions published in the field of *parents in education*, and not even the papers presented

elsewhere by researchers who sometimes prefer to present their research during an ERNAPE conference and sometimes elsewhere. However, a bibliometric study of ERNAPE's archive allows us to explore the state-of-the-art of research activity of the network and emerging trends among *Ernape researchers*.

4. Results in mapping ERNAPE's landscape

Standard algorithms for correspondence analysis have been used in order to reduce the dimensional space in which data can be represented. Analysis has been performed with T-Lab 5.1, a software for quantitative analysis of textual data. Numerical indexes have been computed using the "lemma-X-variables table", a kind of data-organization that allows researchers to explore the number of sub-occurrences for each modality of the variable considered.

The first principal dimension accounts for 33.72% of total inertia (or variance), the second one for 29.54% and, together, they generate a 2-dimensional space that explains 63.26% of the total inertia (Tab. 1).

Dimension Princ.	*Inertia*	*Percentages*	*Cumul. Percentages*
1	0.1399	33.72	33.72
2	0.1226	29.54	63.26
3	0.0912	21.97	85.23
4	0.0613	14.77	100.00

Table 1: Inertia, percentages and cumulative percentages of principal dimensions

The explained inertia is high and, as Greenacre (1993) suggests, each principal inertia accounted by a CA solution should be decomposed with respect to both columns (*AREA*) and rows (*LEMMA*) contribution to improve legibility of the results. The 2-dimensional map thus generated is a fairly good representation of the structure of the data (Fig. 1).

The first principal axis is the straight line that runs closest to the profile point (in the sense of least squares) and passes through the zero point (Greenacre, 1993), thus the major decomposition of inertia is along Dimension 1 or X-axis, and Dimension 2 or Y-axis is the second in order of importance.

Var_AREA	*Mass*	*Coordinates*		*CRT Inertia**		*Quality (squared cosine)*		*Total quality*
		X-Axis	**Y-Axis**	**X-Axis**	**Y-Axis**	**X-Axis**	**Y-Axis**	**%**
Area 1	0.1510	-0.5353	-0.5515	310.8	376.5	0.4365	0.4633	89.98
Area 2	0.4961	0.3557	-0.0649	448.7	17.1	0.9209	0.0307	95.16
Area 3	0.0930	-0.4963	0.6132	163.7	285.3	0.2302	0.3514	58.16
Area 4	0.0700	-0.3118	-0.2919	48.8	48.8	0.1071	0.0939	20.10
Area 5	0.1880	-0.1440	0.4206	28.0	272.3	0.0467	0.3982	44.49
				1000.0	1000.0			

*Table 2: Coordinates and relative contribution to dimensional inertia (*values in per mill)*

Looking at Fig. 1, we find that the variable *Area* (in Fig. 1, the "big", darker dots) moves along the first main axis with the following two poles: the extreme "positive" point to the right is represented by Mediterranean countries (*Area 2*) and the extreme "negative" to the left is the Nordic European area, i.e. *Area 1* (of course, "negative" and "positive" are just artefacts of the calculus, and they could be easily inverted). *Area 2* (448.7 ‰) and *Area 1* (310.8 ‰)

contribute 76.0% of the inertia of the first axis and the quality of *Area 2* is .92, meaning that its contribution to the generation of X-axis is very intense. On the contrary, *Area 3* (163 ‰), *Area 4* (48 ‰) and *Area 5* (28 ‰) do not account much for the inertia of X-axis.

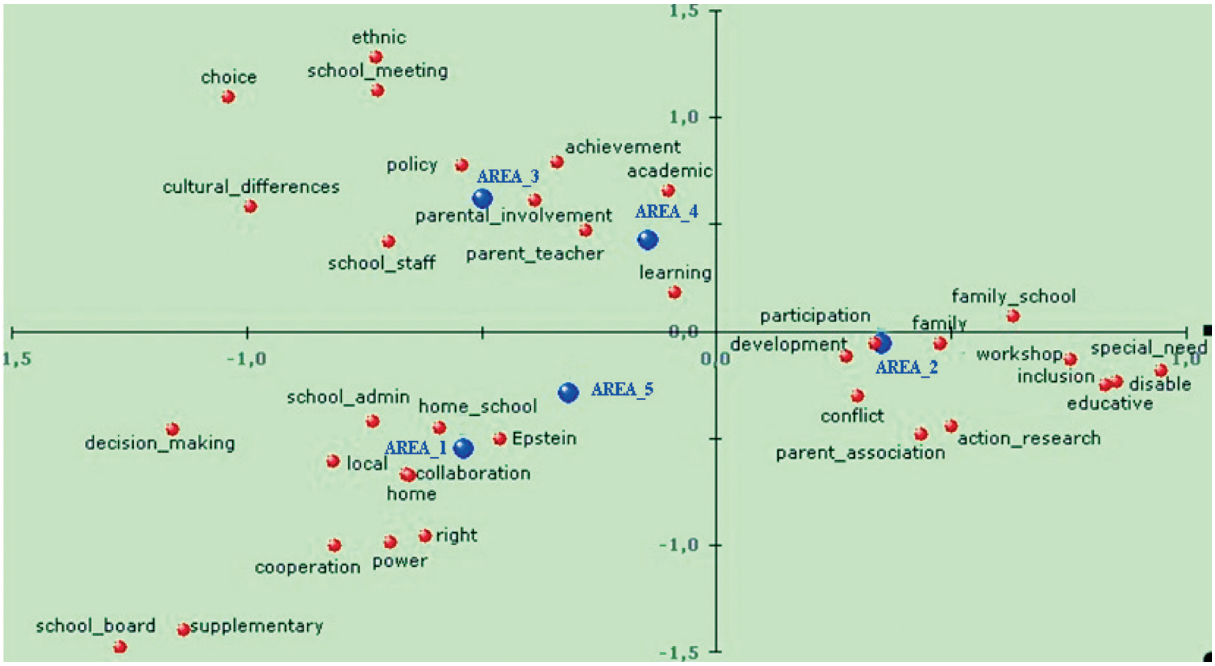

Figure 1: Two-dimensional plot with lemmas and variables

In a similar fashion, the contribution of Y-axis to the total inertia can be computed. In this case, Y-axis segregates Nordic countries (*Area 1*), in the lower side, from Western European countries (*Area 3*) at the upper, "positive" side. Additionally, Anglophone countries (*Area 5*) lie on the "positive" side of the axis (near *Area 3*) while Eastern Europe countries (*Area 4*) lie on the "negative" side (near *Area 1*). More than 90% of inertia on the Y-axis comes from *Area 1* (376 ‰), *Area 5* (272‰) and *Area 3* (285‰). On the contrary, the contributions of *Area 2* (17‰) and *Area 4* (48‰) are fairly small and they do not contribute to explain Y-axis.

To give a more specific numerical support to our interpretation of axes, the next step of analysis involves the decomposition of inertia for "word-points" projected on the same dimensional space. A further inspection of Fig. 1 reveals three major "clouds" of points (i.e. lemmas) roughly corresponding to the three different vertexes of a triangle where *Area 2* is vertex 1, *Area 1* is vertex 2 and *Area 3* is vertex 3.

The decomposition of "lexical" inertia on X-axis suggests that a group of words strongly contributes to its overall inertia: "family" (36‰), "school" (31‰), "school-board" (20‰), "home" (20‰), "parent" (about 10‰) and "disable" (9‰): taken together, these words account for 13% of factorial inertia. Decomposition of inertia of Y-axis reveals that "cooperation" (35‰), "school_board" (31‰), "ethnic" (20‰), "pupil" (15‰), "parental_involvement" (13‰) and "school_meeting" (13‰) account for 13% of its total inertia.

A finer look at word coordinates using test-value of axis poles, allows to evaluate the relationship between lemma and axis and between variable and lemma in term of proximity/distance or, more precisely, in term of similarity/dissimilarity between considered objects. In this cases test-value helps in understanding which lemma contributes to the polarity of the axis. The

main rationale under the use of t-test values is to test whether the words distribution on the 2-dimensional space (or more precisely on X-axis or Y-axis) contributes to the meaning of the axis in a statistically significant way. The measure has a threshold value (1.96) corresponding to a statistical significance of $p < .05$. Results are summarized in Tab. 3.

First dimension (X)				*Second dimension (Y)*			
Positive side		*Negative side*		*Positive side*		*Negative side*	
Lemma	**T-Value**	Lemma	**T-value**	Lemma	**T-Value**	Lemma	**T-value**
family	9.23	school	8.59	ethnic	6.38	cooperation	8.49
disabile	4.56	cooperation	6.85	parental_involv	5.16	school_board	7.97
promote	4.45	home	6.83	school_meeting	5.13	home	7.10
integration	4.30	school_board	6.83	student	4.75	power	4.85
inclusion	4.28	parent	4.84	policy	4.48	right	4.71
family_school	3.54	choice	4.76	practise	4.43	collaboration	4.04
intervention	3.52	local	4.60	behavior	4.40	supplementary	3.96
parent_training	3.34	collaboration	3.95	quality	4.17	sphere	3.88
		school_admin	3.86	mathematic	3.98	local	3.45
		cultural_differ	3.70	academic	3.96	interest	3.30
		democracy	3.64	SES	3.88		
		teachers	3.40	achievement	3.85		
		power	3.39	externalize	3.79		
		home_school	3.37	differ	3.69		
		school_meeting	3.29	homework	3.58		
		decision_making	3.27				
		parental_involv	3.26				

Table 3: Positive and negative poles of Cartesian space (t-test value)

On the positive pole, the horizontal dimension segregates words like "family", "disable", "promote", "integration", "inclusion" and "family_school" on the right side of the graph, from "school", "cooperation", "home", "school_board", "school_admin" and "home_school" which lie on the left. The vertical dimension separates "ethnic", "parental_involvement", "school_meeting", "student", "policy", "practice" and "quality", at the top of Fig. 1, from "cooperation", "school_board", "home", "power", "collaboration" and "local", at the bottom.

The last step of word-correspondence analysis involves the attempt to label the two axes. To this end the three vertices (*Area 2*, *Area 1* and *Area 3*) should be projected onto both principal and secondary dimensions. Using Greenacre and Hastie (1987: 439) words, «Interpretation consists of looking for grouping and contrast in the configuration of projected vertices and in the configuration of projected profiles». In our map, the geometric closeness between X-projected points of *Area 1* and *Area 3* contrasts with their distance from the projected point of *Area 2*, suggesting a difference between issues studied in Mediterranean Countries and the above mentioned areas. In other words, the diversification of ERNAPE's research front is explained by the opposition between studies mainly focused on family sphere in *Area 2* and studies in which the word "family" is much less used, as a consequence of some shift in the focus of researchers.

The analysis of Y-axis reveals that the geometric distances of projected *Area 3*, *Area 2* and *Area 1* points are nearly equivalent. In other words, Y dimension opposes Western European countries to Nordic countries, while Mediterranean Countries lie between them. In this case, the interpretation can be found in the different traditions researchers refer to in exploring the fields of *parents in education*. According to Ravn (2005), the concept of *parents in education*

is shaped by three major rationales that reflects the ways in which the partnership is intended: (1) economic reasons, (2) educational reasons and (3) humanistic reasons. Economic reasons are grounded in a consumer-oriented philosophy when considering educational issues. From this point of view, parents and children are both clients, and the school system is a product, in which customers want to choose from different offers. The consumerist perspective results in discourses about school efficiency, scholastic achievement and quality of services offered to parents. From the educational perspective parents are expected to support school learning, providing ideas, methods, and materials to enhance children's academic success. Parental responsibility in education is regarded as a remedy for increasing school productivity, countering the failure of disadvantaged groups and achieving both the individual's and the nation's economic success. Finally, the humanistic perspective is based on a comprehensive understanding of democracy. Parents are expected to be active partners of a democratic society and their involvement is framed into tasks ranging from decision-making to cooperation with the local community. Parents take care of children's personal and social development in order to help them in becoming *citizens*, rather than *workers*.

A closer look at Y-axis clearly reflects such theoretical approaches: on the upper side of Figure 1, we find Western European countries (*Area 3*) and Anglophone countries (*Area 1*) where the main foci are linked to economic aspects of partnership ("policy", "practice", "quality", "achievement"). On the lower side, on the contrary, we find Nordic countries, where the interest of research is focused on the democratic aspects of partnership ("cooperation", "school board", "power", "right", "local" and so on). Finally, Mediterranean Countries are characterized by pedagogical intents and lie between *Area 3* and *1* on Y-axis, with words like "family", "promote", "disable", "inclusion" and "parent training".

5. Concluding remarks

Mapping science allows us to explore the activity and the structure of scientific landscapes in a given unit of analysis (research fields, research topics or institutions). We are well aware that mapping a research field starting from just a limited number of published paper is far from being optimal. But our aim was much more limited in scope: we only wanted present an empirical application of word-correspondence analysis to analyze the scientific production of a research network. In this case the method has been targeted on ERNAPE's activity, but it can be used for the evaluation of other *units of analysis* (universities, research groups or public institutions). The results demonstrates the suitability of word-correspondence analysis in analyzing the scientific production of a research network. Results from a rather simple correspondence analysis highlight how topics explored by researchers reflect their cultural traditions and their representations of the same object. In an expected and foreseen way, our results fit into Ravn's theoretical framework and somewhat support her point of view of current shifts in the concept of *parents in education*. Moreover, the paper highlights how research data can be shown in a readable format, quite useful to communicate with policy makers.

References

Anuradha K.T. and Urs S.R. (2007). Bibliometric indicators of Indian research collaboration patterns: A correspondence analysis. *Scientometrics*, Vol. 71, 2: 169-179.

Beh E.J. (1999). Correspondence analysis of ranked data. *Communication in statistics. Theory and method*, Vol. 28, 7: 1511-1533.

Benzecri J.P. (1992). *Correspondence Analysis Handbook.* New York: Marcel Dekker.

Bhattacharya S. and Basu P.K. (1998). Mapping research area at the micro-level using co-word analysis. *Scientometrics*, Vol. 43, 3: 359-372.

Cahlik T. and Jirina M. (2006). Law of cumulative advantages in the evolution of scientific fields. *Scientometric*, Vol. 66, 3: 441-449.

Chen C. (2006). Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, Vol. 57, 3: 359-377.

Dore J.C., Ojasoo T., Okubo Y, Durand T., Dudognon G. and Miquel J.F. (1996). Correspondence factor analysis of the publication patterns of 48 countries over the period 1981–1992. *Journal of American Society for Information Science*, Vol. 47, 3: 588-602.

Greenacre M.J. (1993). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre M.J. and Hastie T. (1987). The geometric interpretation of correspondence analysis. *Journal of the American statistical association*, Vol. 82, 398: 437-447.

Ravn B. (2002). The cultural context of learning and education in England, France and Denmark as a basis for understanding educational change. *Journal of Educational Change*, Vol. 3, 3: 241-263.

Ravn B. (2003). Cultural and political divergences in approaches to cooperation between home, school and local society in Europe. In Castelli, S., Mendel, M. and Ravn, B., editors, *School, family and community partnership in a world of differences and changes*, Gdansk: Wydawnietwo Universyletu Gdanskiego, pp. 9-18.

Ravn B. (2005). An Ambiguous Relationship: Challenges and Controversies in the field of Family-School-Community Partnership. Questioning the discourse of partnership. Paper presented at first *"School-family partnership: past, present and future" meeting*, Department of Psychology, Universty of Milano-Bicocca, Milano, Italy.

Small H. (2006). Tracking and predicting growth areas in science. *Scientometrics*, Vol. 68, 3: 595-610.

Van den Besselaar P. and Heimeriks G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, Vol.68, 3: 377-393.

Van Raan A.F.J. (2003). The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. *Technikfolgenabschätzung*, Vol. 1, 12: 20-29.