P | U | S

Public Understanding of Science
1–15
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0963662515576864
pus.sagepub.com
SAGE

Ten years of science news: A longitudinal analysis of scientific culture in the Spanish digital press

Tamar Groves, Carlos G. Figuerola and Miguel Á. Quintanilla
University of Salamanca, Spain

Abstract

This article presents our study of science coverage in the digital Spanish press over the last decade. We employed automated information retrieval procedures to create a corpus of 50,763 text units dealing with science and technology, and used automated text-analysis procedures in order to provide a general picture of the structure, characteristics and evolution of science news in Spain. We found between 6% and 7% of science coverage, a clear high proportion of biomedicine and predominance of science over technology, although we also detected an increase in technological content during the second half of the decade. Analysing the extrinsic and intrinsic features of science culture, we found a predominance of intrinsic features that still need further analysis. Our attempt to use specialised software to examine big data was effective, and allowed us to reach these preliminary conclusions.

Keywords

automatic classification, media and science, public understanding of science, science and popular culture, science communication, science journalism, text clustering

1. Introduction

Science and the media constitute a thriving field of academic inquiry, which play host to contradictory interpretations regarding the presence and presentation of scientific content to the public. The media have often been criticised for their scant, unreliable and sometimes melodramatic treatment of science. These kinds of criticisms fit well with the scientific literacy paradigm, as the public's ignorance when it comes to science and technology can be attributed to inadequate media coverage. However, empirical studies have provided a much more complex picture. Although the presence of science in the media appears to obey journalistic news criteria such as *negativity, unexpectedness, unambiguity, personalisation, conflict, prominence, significance* and *human interest* (Major and Atwood, 2004), at the same time, there is evidence of an increase in science

Corresponding author:
Miguel Á. Quintanilla, Institute of Science and Technology Studies, University of Salamanca, Edificio I+D+i. C/ Espejo, n° 2, 37007 Salamanca, Spain.
Email: maquinta@gmail.com

coverage of worthy quality, which is contributing to the public's understanding of science (Dimopoulos and Koulaidis, 2002).

These attempts to analyse the presence of science in the media are part of a wider effort to expand the framework which determines our understanding of the relations between science and society. While in the past, the main source of information has been surveys, there is growing awareness of the limited nature of the information they are able to provide (López Cerezo and Cámara, 2007). Other indicators based on data relating to the education system, demography, legislation, media coverage, and so on have been proffered (Benoit and Gingras, 2000). Among these alternative cultural indicators, an analysis of mass media, and specifically of the press, has become relatively important (Bauer et al., 2007). The media are considered to reflect the level of social appropriation of science – that is, how the contents of scientific culture are incorporated into popular culture or general knowledge (López Cerezo and Cámara, 2007). The media are also active agents in the dissemination of scientific information, and help shape society's perception of science. As the media both reflect and influence public opinion, an analysis thereof enables us to gain an appreciation of the position of science in society (Bauer et al., 2006).

Science and technology appear in the press in many different ways. An interesting classification distinguishes between 'Science-popularization' – usually in special sections, dominated by biomedical topics and depicting science as a straightforward and consensual endeavour which improves people's lives – and 'Science as News', which more often deals with other fields, paying closer attention to controversy and risks (Bucchi and Mazzolini, 2003). Another study which also points to the plurality of types of science coverage in the press distinguishes between different kinds of triggers for science news: (1) Events within the scientific world (scientific papers, conferences, announcements by scientific institutions), (2) non-scientific events (epidemics, political decisions, etc.), and (3) scientific and non-scientific events (rocket launches, congressional hearings, etc.) which are partly related to scientific issues (Elmer et al., 2008). These distinctions show that science coverage serves multiple needs and occurs in response to a variety of triggers.

This article presents our study of science coverage in the Spanish digital press over the last decade. Our aim is to provide a general picture of the structure, characteristics and evolution of science news. Most publications on science in the press are dedicated to specific issues.[1] These publications characterise press coverage of different scientific fields and, in some cases, even show that specific scientific issues are covered differently by the media, depending on their epistemic culture (Schafer, 2009). Our research, on the other hand, builds on previous efforts to provide a general characterisation of science news in different national contexts such as the United Kingdom (Bauer et al., 1995), Greece (Dimopoulos and Koulaidis, 2002), Italy (Bucchi and Mazzolini, 2003), Bulgaria (Bauer et al., 2006), Germany (Elmer et al., 2008) and Croatia (Šuljok and Brajdić Vuković, 2013).

These studies demonstrate an increase in science coverage. The studies on the United Kingdom and Bulgaria show a similar pattern: a peak in science coverage in the 1960s, a decrease in the 1970s and an increase in the 1980s and 1990s (Bauer et al., 2006). In the case of Italy, there was a 4000% increase between 1946 and 1990 – the period on which Bucchi and Mazzolini's 2003 study focuses. In Germany, the researchers found a 48% increase between 2003–2004 and 2006–2007 (Elmer et al., 2008). In addition, generally speaking, these studies are largely in agreement with regard to the growing importance of medicine and biology, which constitute more than 50% of science news in Italy (Bucchi and Mazzolini, 2003), around 30% in Germany (Elmer et al., 2008), and 42.6% according to a study conducted in the United States (Clark and Illman, 2006).

Research on the presence of science and technology in the press varies in terms of what it defines as science. Some of these studies adopt a broad definition of science, including social science and even humanities (Šuljok and Brajdić Vuković, 2013). Others adopt a more restricted view, limited to pure sciences, biology and medicine (Bucchi and Mazzolini, 2003). For our purposes, we have used this more restrictive definition. However, while in most cases technology is included in science, one of our basic classifications distinguishes between ‘science’ and ‘technology’ news items. Technology is the application of scientific knowledge to respond to human needs and desires. Hence, even though it is connected to science, it is still epistemologically different. Science is mainly ‘know that’, whereas technology is ‘know how’. Thus, we believe that in characterising scientific culture, it is important to evaluate them separately. Nevertheless, it should be clarified that articles about gadgets were not included in our corpus, as we initially felt they were commercially oriented and did not fit in with our definition of science and technology coverage.

The usual procedure employed in longitudinal studies of the press is to sample articles from printed versions. Some studies examine relatively extensive samples – for example, around 6000 articles from seven UK newspapers (Bauer et al., 1995) or 4077 from three German newspapers (Elmer et al., 2008); others use smaller ones, such as 1867 articles from four national Greek newspapers (Dimopoulos and Koulaidis, 2002), 1336 from a leading Italian newspaper (Bucchi and Mazzolini, 2003) or 60 issues of the *Science Times Section* (Clark and Illman, 2006).

In order to contribute to the effort to gather robust data, to enable international comparative studies to be carried out (Bauer et al., 2006), our aim was to analyse a substantially larger corpus. Therefore, we decided to use the digital editions of three Spanish national newspapers and employ automated text-analysis procedures. There can be no doubt that the digital revolution is changing the face of contemporary media. All major newspapers nowadays have digital editions, and recent surveys indicate that there is a growing tendency to use the Internet as a source for scientific information (FECYT, 2011). Hence, our corpus is based on digital editions of the press – that is, the articles published on the newspapers’ websites. As a result, our conclusions refer to Internet content rather than to the printed press.

Our study centres on the digital editions of two leading Spanish newspapers: *El País* and *El Mundo*, from 2002 to 2011. These are the most widely read dailies in Spain. We could not use the third most popular newspaper ABC, as technical factors prevented us from employing our automatic retrieval techniques. However, we did also include an analysis of a third newspaper: *El Público*. *El Público* was founded in 2007, and as its editors declared their intention to devote particular attention to science and technology, we felt it would be interesting to include it in our study.[2]

Analysing the digital press enabled us to use automated information retrieval procedures to find and catalogue science news. To perform the analysis, we used several kinds of specialised software, adapted to our needs. Recently, these kinds of software are being used with increasing frequency for text analysis, both in general and with regard to public understanding of science in particular. A programme that uses correspondence computation and cluster analysis to create semantic maps based on recurrent word clusters was employed to analyse 646 articles on nanotechnology in Spain (Veltri, 2013). The Word Space Model was employed to detect patterns and changes in vocabulary in 465 abstracts from the journal *Public Understanding of Science* over a 20-year period (Suerdem et al., 2013).

The best algorithm for automatic text categorisation is a hotly debated issue. We chose different kinds of software according to the specific questions that interested us. In this sense, we were following a typical content-analysis procedure of designing a framework for analysis, coding text units and drawing inferences (Krippendorff, 2004). Conversely to the usual practice of employing human coders, we used computer-based coding. The most obvious advantages of our attempt to offer an alternative to traditional manual content-analysis is that there is no need to draw a representative sample, as the computer can handle the whole corpus. Moreover, the software we used provides indicators with regard to the level of confidence of its classification. Notwithstanding, we also carried out manual checkups to confirm the degree of reliability of our results.

While we were interested in the salience of science in the press in general, and the specific topics that draw greatest attention in particular, we also wanted to characterise the way science and technology are presented. Most studies ask if the treatment of science and technology is positive or negative, or whether they are presented in a consensual or contested way; we, however, approach science-related coverage in a different way. In our research, we distinguish between two kinds of scientific and technological culture: intrinsic and extrinsic (Quintanilla, 2005, 2012). *Intrinsic scientific culture* refers to cultural components inherent to scientific activities. The scientific knowledge in each of the areas and fields of research, the scientific method and the values which are supposed to guide scientific research are included in this kind of scientific culture. *Extrinsic scientific culture* includes all the representational (beliefs), practical (norms) and evaluative (values) components which are related to scientific activities, institutions and people, but are not part of intrinsic scientific culture: the images of science, and the evaluations of science from cultural, legal, moral, political, religious and other points of view. We also distinguish between these two kinds of cultural components in the case of technological culture: ‘technological culture incorporated into technical systems’ (intrinsic) and ‘technological culture not incorporated into technical systems’ (extrinsic). Thus, we were interested to see whether scientific and technological information conveyed by the press is intrinsic – that is, whether it reflects the knowledge, methods and values of science and technology – or extrinsic, dealing with science or technology from other perspectives – for example legal, moral, practical, and so on.

Our main concern is to see the degree to which the digital press carries out traditional ‘science communication’ focusing mainly on scientific information, method and values, and the degree to which it is concerned with other external dimensions – that is, the political, economic, legal or ethical perspectives of scientific activities. Other studies make similar distinctions derived from literature about frames (Benford and Snow, 2000) and speak, on the one hand, of the scientific frame corresponding to our intrinsic science and, on the other hand, of the political, economic and ethical–legal–social frames referring to our category of extrinsic science (Schafer, 2009).

One of the characteristics of intrinsic scientific culture is its division along the lines of academic disciplines, which are the result of decades of evolution of science knowledge and the organisation of the scientific community. Science communication does not necessarily correspond to these disciplines, as it is influenced by other factors and dynamics, such as social concerns, political debates, journalistic values, and so on. The distinction between the esoteric and exoteric dimensions of science is indicative of these differences in the dynamics between ‘science’ and ‘science communication’ (Bauer, 2009; Vogt, 2011). In this study, we do not identify academic disciplines, but only look for the scientific and technological issues that dominate the digital press coverage of science in Spain.

Thus, the objectives of our study are (1) to evaluate the salience of science news in the context of Spain, (2) to analyse the ratio of intrinsic and extrinsic components, (3) to discern the most frequent themes, and (4) to test several methodologies capable of handling big data.

2. The corpus of news and methodological considerations

Our corpus (SCSC: Spanish Corpus of Science Culture) contains 50,763 text units dealing with science and technology, and includes all kinds of published materials (articles, interviews, editorials, etc.). Our first step was to download all the digital editions of the three newspapers from 2002 to 2011 (*El Público* only from September 2007). A web crawler (Olston and Najork, 2010), designed ad hoc, was trained to recover as many text units as possible. The articles (and other usable content) were identified and isolated from other elements such as adverts, banners, menus, and so on. The articles were first saved in HTML format and then converted into plain text.

At the beginning, we encountered a problem with *El Mundo*, as the number of items it contained was much higher than in the case of the other two newspapers, and also increased dramatically from year to year. Fuzzy hashing techniques (Kornblum, 2006, 2010) revealed that many items were 'near duplicates', with almost identical information but a different URL. We excluded these items from the final corpus.

In the final analysis, we downloaded close to 900,000 articles. Due to changes in the structure of the websites, it is impossible to guarantee that we gathered all the articles that have appeared on the digital versions. Even if improvements are made to the technical procedures, automatic retrieval procedures will never be perfect, because to a certain extent, the corpus of digital news is a fuzzy entity. However, although we are not dealing with exact numbers, we are able to process big data.

The total quantity of news items published on the online version of *El Público* is relatively stable, at around 26,000 articles per year, from 2008 to 2011. In the case of *El País*, in 2001, there were around 52,000; this figure remained stable until 2006, when it started to decrease slightly, reaching 37,000 in 2009. From 2009 to 2011, there was no substantial change. The newspaper which has experienced the most spectacular change is *El Mundo*. It started with a very low quantity of news items in 2002 (5400) and actually grew to be the newspaper with the highest number of items, with almost 59,000 in 2011.[3]

We used two kinds of procedure to analyse this immense quantity of articles. First, we used an automatic classifier to find all the items related to science and technology and to distinguish between intrinsic and extrinsic components of science and technology. Second, we employed a text-clustering programme to identify the main themes of science and technology coverage.

Automatic classification

We used an automatic supervised classifier, based on the Support Vector Machine (SVM) (Joachims, 2002; Vapnik Vladimir, 1995), and employed LibSVM for Python.[4] Usually, thousands of articles are classified manually to begin with, in order to provide the automatic classifier with a training dataset. In our case, we used a 'Self Training' procedure (Abney, 2002; Sebastiani, 2002). A relatively small sample of 999 articles classified manually (as science, technology and their intrinsic and extrinsic features) was used to catalogue all the rest of the articles. The human coders received a table that briefly described the characteristics of each category (Science, technology, intrinsic science, extrinsic science, intrinsic technology, extrinsic technology) and participated in two training sessions to clarify any misunderstandings. The general instruction to the human coders was that an article should be classified as science and technology if it contains information that provides the reader with any kind of knowledge regarding science or technological artefacts. We did not require a minimum quantity of information, because we consider that science and technology may also be present in articles that deal with other topics. As a result, our corpus includes articles published in specialised science and technology sections, as well as articles which appeared in other sections, as long as they contain data informing the reader about science and technology. An article about the Fukushima nuclear disaster (March, 2011), for example, could be classified as both science and technology if it treated radiations tests and their implications, in addition to a description of the systems failure to resist the earthquake. We ran Krippendorff's Alpha coefficient test (2004) to check the reliability of the manual coding and obtained the following results: Science = .701525, Intrinsic Science = .788416, Extrinsic Science = .724790, Technology = .769894, Intrinsic Technology = .651114 and Extrinsic Technology = .771417, which are reasonable.

In the first stage, we used the manually coded sample to train the classifier to identify science and technology articles. In the second stage, we used this sample to train it to identify the types of science and technology articles. In this first sample, we had 529 articles classified as Intrinsic

Science, 105 as Extrinsic Science, 101 as Intrinsic Technology and 224 as Extrinsic Technology. The classifier was run on the whole corpus, and the articles which were classified with the highest confidence level were added to the training sets. Even at this second stage, we had already obtained 733 articles classified as Intrinsic Science, 172 as Extrinsic Science, 158 as Intrinsic Technology and 246 as Extrinsic Technology. This new set was used to re-classify all the articles. This process was repeated until the results were deemed reliable. We finally ended up with 1000 articles classified as science, 1000 as technology, 1000 as Intrinsic Science, 457 as Extrinsic Science, 158 as Intrinsic Technology and 802 as Extrinsic Technology. This process is known as ‘active learning’ (Novak et al., 2006), and it includes manual controls (classifying random samples). We carried out a *k*-fold cross-validation ($k = 5$) (Arlot and Celisse, 2010). The mean of correct classifications for all categories was 89.2%.

When we compared the overall results of the manual and automatic classifications, we found clear similarities. According to the manual classification, 76.4% of articles were tagged as science and 29.4% as technology. The automatic classifier gave 74.8% and 27.8%, respectively. The same is true for intrinsic and extrinsic science: 91.1% and 18.1%, respectively, in the manual classification, and 65.5% and 11.0% with the automatic classification, respectively. The results in the case of technology were less convincing. With regard to Intrinsic and Extrinsic components, we identified 45.1% and 100%, respectively, in the manual classification, and 6.7% and 75.6% according to the automatic classifier, respectively. It is clear that, generally speaking, the automatic classifier gave reliable results, except in the case of Intrinsic Technology, where the percentage of articles is substantially lower in the automatic classification than in the manual one. As we saw in the manual classification, Krippendorff’s Alpha coefficient was significantly lower in this category, and in the automatic learning process, we were able to find no more than 158 articles for our intrinsic technology training set. This may have reduced the ability of the automatic classifier to identify this kind of articles. Another explanation which might explain the problem of classifying this category in both the manual and the automatic analyses is related to the fact that technology articles are dominated by extrinsic components, and the presence of intrinsic information is scarce and always accompanied by extrinsic elements. Thus, the classifier finds it difficult to recognise intrinsic dimensions.

Text-clustering

The aim of text-clustering is to group similar documents together. Documents which discuss the same topic should be clustered together. However, clustering of documents is computationally expensive: most of the classic algorithms (*k*-means, etc.) require a number of desired clusters to be established in advance, and are afflicted by noise (ambiguous documents, marginal or minority topics). Thus, we used social network analysis to track and detect themes in our corpus. The idea was to treat our corpus of news on science and technology as if it was a graph or a grid. Each text unit was considered as a node and the links between them (the edges) reflect their semantic correlation. The links are undirected, and their weight reflects the degree of semantic similarity between two text units. In order to compute the similarity between text units, we used the Vector Space Model (Salton and McGill, 1983). Every text unit is a vector of terms, and every term in every text unit has a specific weight. The resemblance between every pair of text units is the similarity between their vectors of terms. The value of the similarity ranges from 0 to 1. In order to detect communities, we chose to use the Infomap algorithm (Rosvall et al., 2009). We have also tried other algorithms such as Walktrap (Pons and Latapy, 2005), Multilevel (Vincent et al., 2008) and Label Propagation (Raghavan et al., 2007). However, based on our own experience as well as on comparative studies (Lancichinetti and

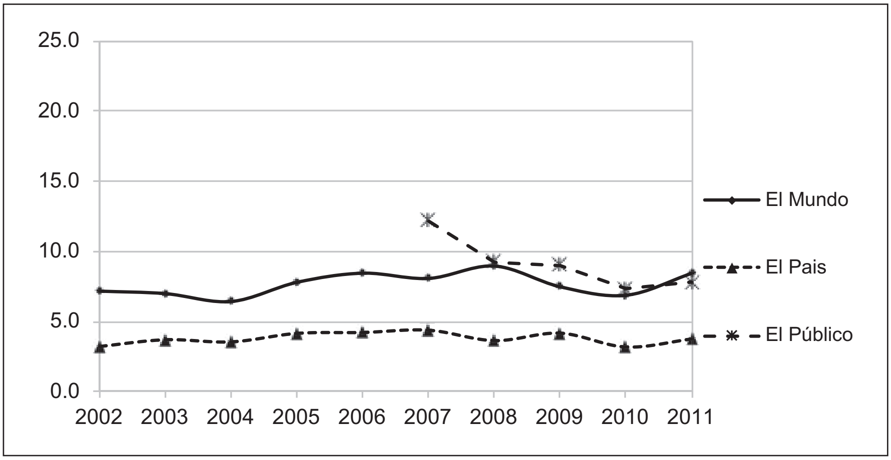

Figure 1. Percentage of S & T news.

Fortunato, 2009; Lee and Cunningham, 2014; Plantié and Crampes, 2013) we opted for Infomap, as it is very fast and memory efficient (we have 50,000 nodes and more than 3 million links!), it is effective at handling weights or strengths of links and produces a hierarchical tree of communities, with only three sublevels.

3. Results

According to our study, the number of science and technology articles published each year increased from 2082 in 2002 to 8543 in 2011 – a 400% increase. The growth rate was always positive, except in 2010 (−13%). We saw substantial growth in 2006, and even greater in 2007, with *El Público* first being published in September that year.

We found 50,753 articles containing science and technology. These articles accounted for 7.6% of news items in *El Mundo*, 3.7% of news items in *El País* and 9.1% in *El Público* (Figure 1). These percentages are relatively stable during the period upon which our analysis focused. Generally speaking, during the years 2002–2011, science-related coverage in these three newspapers accounted for 6.9% of total news items.

While in many studies, there is no distinction between science and technology, we classified the text units into four categories. The most populous category is Science, including 70% of the articles. It is followed by Technology, with 23%. There are also 5% of the articles which are classified as both Science and Technology, and 2% of articles are unclassified. These articles were identified as science and technology during the first stage of classification, designed to isolate science and technology articles from the rest of the text units. However, in the second stage, when we wanted to distinguish between science and technology, the classifier could not determine to which of the two they belonged. Evidently, 2% is very low, and generally speaking, the classifier is able to differentiate between the two.

From 2002, when science appeared in 64% of the analysed news items and technology in 24%, there was a continuous increase in the relative weight of science, which reached its maximum value in 2006 (75%). From that point on, the weight of science decreased and that of technology steadily increased, up until 2011 (63% science, 31% technology). The differentiation between science and technology increased (the percentage of articles classified as both decreased from 9% in

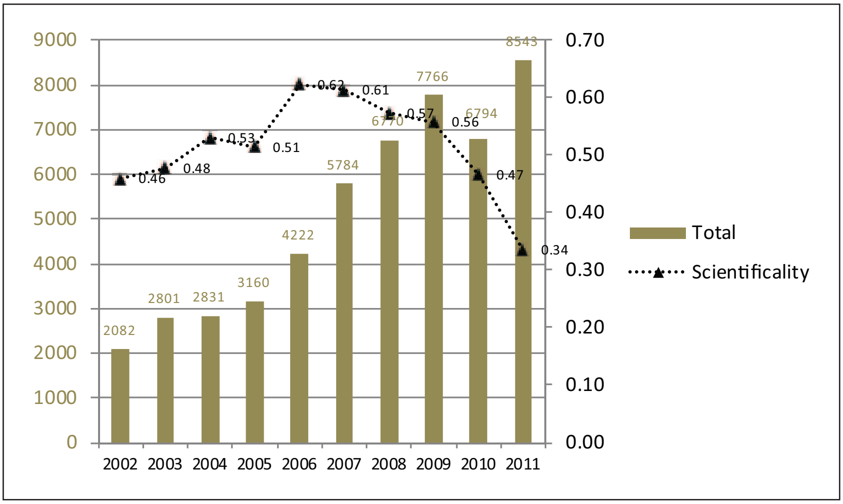

Figure 2. The total number of science and technology articles, and the 'Scientificality' indicator.

2002 to 5% in 2011) and the number of unclassified articles also decreased (from 3% to 4% in the earliest years to less than 2% in the most recent).

To summarise all this information, we constructed an indicator of 'Scientificality', which reflects the ratio of scientific to technological contents (S − T/S+T, which varies from −1 to 1). As we can see, in the SCSC corpus, scientific culture is more present than technological culture, although it seems that the situation is evolving towards a relative balance. The highest presence of scientific components in comparison to technological components can be seen in 2006, while the lowest is in 2011 (Figure 2).

Extrinsic and intrinsic features of science culture

We attempted to automatically distinguish between intrinsic and extrinsic scientific content. Our automated procedures were able to classify 69% of the texts identified as science. There is a relatively high percentage (31%) of unclassified articles. These articles were classified as science, but the classifier found it difficult to decide whether they were intrinsic or extrinsic. While on the theoretical level, intrinsic and extrinsic science can be defined as distinct, the automatic classifier is not yet sophisticated enough to always classify them with a high level of certainty, and therefore leaves them as unclassified. A total of 58% of the articles were categorised as intrinsic scientific content – that is, scientific information coherent with traditional science dissemination – 4% were classified as extrinsic content; and in 7%, we encountered both kinds of scientific features – intrinsic and extrinsic.

Extrinsic and intrinsic features of technology

We applied a similar procedure for technology. In this case, our automatic method was unable to classify 24.4% of the articles – slightly less than in the case of science. This means that among the articles originally classified as technology, the classifier was unable to tag nearly a quarter as intrinsic or extrinsic technology. A total of 68.9% were classified as extrinsic technology and 6.7% as both intrinsic and extrinsic technological culture, while texts with only intrinsic technological information are practically absent (.03%).

Main themes

By using social network analysis for the study of semantic nets, we were able to detect the main themes of science news. We found 23 first-level themes. They all have between 1 and 190 subgroups. We limited our analysis to those thematic groups which contained more than 50 articles. Thus, the number of first-level groups was reduced to 13, and that of the second-level groups to 126, covering 87% of the corpus.

The labels given to the clusters are an abstraction of their contents (we used the same letter for clusters that deal with similar themes). *A1* was called *Biomedicine*, as it includes sub-clusters referring to a wide range of medical issues such as cancer, sexual health, Alzheimer's, stem cells and so on. *A2* was called *Public Health*, as its sub-clusters deal with different kinds of epidemics. *B* was called *Energy*, as it deals with issues such as nuclear technology, climate change, natural gas, solar energy and so on. *C1* is *Natural Resources*, as its sub-clusters refer to food, water, woods, and so on. *C2*, called *Biodiversity,* includes sub-clusters of articles on coasts, natural parks, ecological initiatives and so on. *C3, Development*, has sub-clusters on issues such as hunger, poverty and demography. *C4, Contamination*, has sub-clusters dealing with different kinds of toxic waste. *C5, Protected Species*, has sub-clusters on red tuna fish, whales, and so on. *D, Aerospace*, has sub-clusters relating to National Aeronautics and Space Administration (NASA), asteroids, the planet Mars and so on. *E, IT (Information Technology),* has sub-clusters such as Internet and video games. *F1*, called *Astronomy and Cosmology*, has sub-clusters dealing with astrophysical discoveries. *F2, Human Evolution*, has sub-clusters on palaeontology, apes, Neanderthals and so on. *G, Science Policy*, has sub-clusters dealing with science policy, Nobel prizes and universities. *H* (null) has the texts which are not included in the 13 most numerous first-level clusters – around 13% of the corpus.

The dominance of medical themes is clear: biomedicine (19%) and Public Health (13%) together count for 32% of the total number of articles. Energy is the second most important theme (16%). The third one can be dubbed 'Environment' (16%), as it includes Development (5%), Natural resources (5%), Biodiversity (4%), Contamination (1%) and Protected Species (1%). The next themes are Aerospace (8%) and IT (7%). Themes related to basic science can be found in the Astronomy and Cosmology (3%) and Human Evolution (2%) clusters. The relative prevalence of Scientific Policy (3%) is also noteworthy.

Figure 3 illustrates the changes in the relative presence of each group of topics. The most pronounced increase is in IT, whose cumulative annual growth rate reached 35%. We can also see an increase in articles dedicated to Protected Species (31%), Energy (21%) and Human Evolution (23%) and Astronomy and Cosmology (20%). There is a decrease in absolute terms in the presence of Contamination (−3%), and in relative terms in the topics of Development and Environment (except biodiversity). Biomedicine, Public Health and Biodiversity are relatively stable, and there is a slight decrease, with some fluctuation, in Aerospace and Scientific Policy.

When we checked the presence of intrinsic and extrinsic features of the different themes, we found that intrinsic scientific components characterise the news items on Human Evolution (90%), Biomedicine (81%), Astronomy and Cosmology (78%), Biodiversity (70%) and Protected Species (65%). Extrinsic components are especially pronounced in Science Policy (81%). The ambiguous articles (not classified by our automatic procedure) have a significant presence in Natural Resources (89%) and also in Contamination (70%), Development (69%), and Energy (60%). These themes are related to the environment, but not necessarily from a scientific or technological perspective. One possible explanation for the high presence of unclassified items which is characteristic of these themes might be that their articles deal mainly with economic, political and social issues, and

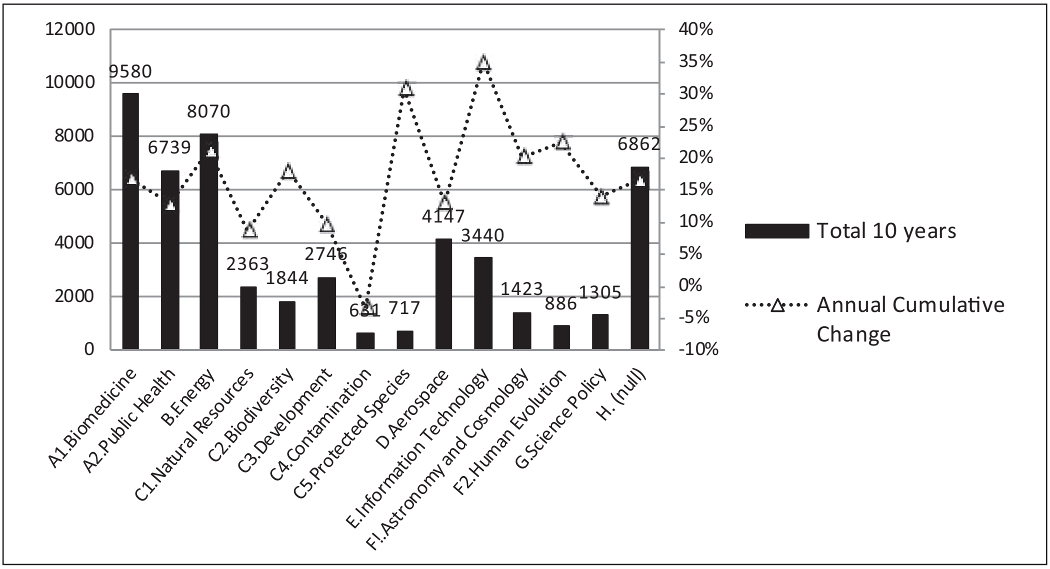

Figure 3. Number of articles and annual cumulative change by theme.

the presence of science and technology content is very low. Thus, when the classifier compares them with its models for intrinsic and extrinsic science, it cannot reach a decision.

We also carried out an analysis of the length of the articles in order to obtain some indication with regard to the complexity of their content. The average text has 540 words. The average text length of texts classified as science is a little longer: 558 words, followed by those classified as technology: 485 words. The longest texts are those classified as intrinsic science (565), and the shortest, intrinsic technology (444). When we look at different topics, there are hardly any differences. Biomedicine (594) and scientific policy are the longest, with respective averages of 594 and 613, while aerospace and IT are the shortest (460 and 481). We detect a constant growth rate of almost 4%, starting with an average of 400 words in the first years (2002–2004) and reaching almost 600 in the last year 2011. However, it is very difficult to interpret these findings, as the standard deviation for the whole corpus is 373.

4. Discussion

We found between 6% and 7% of science-related coverage in our corpus. This figure is similar to (though slightly higher than) international findings about the presence of science and technology in the printed press. However, our initial definition of science and technology excluded social science and humanities, which are included in other research protocols. However, due to the fact that we did not require a minimum quantity of scientific or technological information, our corpus also includes articles dealing mainly with other issues. This may explain why we might have detected more science and technology than is usual. An establishment of a standard definition of what is being sought where science and technology in the press are concerned would facilitate international comparisons. In addition, it must be remembered that our analysis is based on the digital press which is different from the printed editions. We were not able to locate information about science coverage in the digital press in other countries, and generally speaking, in most research on the sources of scientific information of the public, we find data about the daily press on the one hand and the Internet on the other hand, without specific reference to the digital press. As a result we can only compare our findings with the printed press.

When one looks at the data with regard to the main sources of scientific knowledge, there is evidence that television is the first on the list, followed by the Internet, the daily press and the radio (Moreno Castro, 2010). A research project on the presence of science and technology on the little screen in different European countries found that Spain came second after France, leaving behind Italy, Germany and the United Kingdom (León, 2008). While not much research has been done on this issue, this publication does at least indicate a similar pattern with regard to the salience of science coverage found by us, but in a different kind of media.

The percentage of science and technology articles across the three digital versions of the newspapers we observed was relatively stable. This phenomenon in the context of a continuous growth of digital content means that science and technology are part of this general expansion.

In our corpus, we detected a predominance of science over technology. However, from 2006 onwards especially, there is an increase in technological content. Technology is the application of scientific knowledge to the solution of practical problems. The fact that it substitutes scientific content might imply the replacement of the traditional academic view of science with a more practical technological approach. Given that content about gadgets was excluded from our corpus, this phenomenon may actually be even more widespread than is reflected by our data.

Analysing the extrinsic and intrinsic features of science culture, we found that intrinsic components were more present than extrinsic components. These results are not surprising. Intrinsic science includes scientific knowledge and methods. It naturally appears in traditional science communication articles as well as in articles dealing with other topics referring to related scientific facts. Extrinsic science on the other hand includes the social appropriation of science that is how society perceives and manages science. Dealing with science as a social or political endeavour is relatively rare. It usually happens when moral issues are concerned or with regard to scientific policies related to different issues. This means that in our corpus, there was a predominance of scientific information which, generally speaking, is presented as isolated from its social, political, legal and economic implications. However, it must be mentioned that our classifier was unable to categorise 31% of science articles as intrinsic or extrinsic. This could be related to the fact that with articles which have a relatively small amount of scientific content alongside a totally different issue, the classifier’s task of distinguishing between intrinsic and extrinsic content is complicated still further. There is a need for additional analysis in order to test this interpretation. Being able to analyse these 31% of ambiguous articles could give a different picture with regard to the relations between extrinsic and intrinsic features of science news.

The technological content in our corpus is almost exclusively extrinsic. There is very little intrinsic information, and when it appears, it is almost always accompanied by extrinsic content. Intrinsic technological information refers to the characteristics of technological systems. It is unsurprising that it does not appear in the press, and when it does, it appears alongside discussions of the use and purposes of technology, which are classified as extrinsic features.

Our study confirms the international findings about the dominance of biomedicine when it comes to science coverage in the press. Also, in the case of Spain, we detected an undeniable ‘medicalisation’ of science news. When analysing the kind of scientific culture (intrinsic or extrinsic) of news referring to biomedicine, we found a dominance of intrinsic components. This might imply the predominance of the academic model of public communication of science in Spain – that is, that the press tends to reproduce scientific information without entering into debates about its social, political or moral implications. This conclusion is also in accordance with international findings, which indicates a relatively consensual depiction of science in the news. We also detected a clear increase in the coverage of IT, which confirms the previous finding of the increase of technology. The only issue that has a negative rate of growth is Contamination. This can be explained by the fact that the first 2 years of the period we are analysing were dominated by the ecological disaster of the Prestige oil spill on the Galician coast in 2002.

This observation with regard to Contamination fits well with a more general characterisation of the themes of science-related coverage in our corpus and their sub-clusters. They tend to be related to current events, such as epidemics (in the case of Public Health, this is very clear) or natural disasters (as mentioned with regard to Contamination).

These results lead us to distinguish between three types of treatments of science and technology in the Spanish press: (1) *The dissemination of science (news of science)*: articles which include scientific information related to a specific phenomenon. This can be called 'intrinsic scientific culture'. (2) *Science as a social phenomenon (news on science)*: here we refer to articles about the image of science or technology as something positive or negative, efficient or dangerous, important or marginal. These are components referring to science but which do not necessarily have any intrinsic scientific information. We name them 'extrinsic scientific culture'. (3) *The scientific dimension of news (news with science)*: this is content related to current affairs and events, which also includes scientific information. These components can include both intrinsic and extrinsic features, but are mainly triggered by contemporary events and current affairs. Usually, they include intrinsic scientific elements. There is a need for further analysis of our corpus to verify these three categories. However, they seem to coincide with the conclusion about the case of Germany, in which it is argued that previously, science news was transmitted in a factual manner, and that recently a more 'everyday' coverage of science has emerged, triggered by disasters or the impact of political decisions on issues such as health and the environment (Elmer et al., 2008: 879).

In this study, we also tested the use of automated information retrieval procedures to recuperate science news and several kinds of specialised software to classify and analyse it. As discussed above, their usage was efficient in analysing our vast corpus and allowed us to reach some preliminary conclusions. Nevertheless, we are left with the challenge of explaining the high number of unclassified articles related to our model of intrinsic and extrinsic dimensions of science and technology. In addition, while the clustering yielded very interesting results, there is a need to analyse the sub-clusters and their significance more carefully. Still, we feel that the ability to handle big data can contribute to the analysis of science in the media and to the standardisation of science culture indicators based on the analysis of the digital press.

Acknowledgements

We would like to thank Bárbara Estevez Ribeiro, Marcelo Arancibia, Fernando Pita, Ana Victoria Pérez, Cristina Gónzalez, Óscar Montañes and Antonio Montero for their contribution to the manual classification.

Funding

This research has been funded by the Spanish Ministry of Science and Innovation **FFI2011-2776**.

Notes

1. Recent publications, using different methodological and theoretical approaches, include issues such as genetically modified organisms (Castro and Gomes, 2005), climate change (Carvalho, 2007), Nanotechnology (Donk et al., 2012) and fracking (Jaspal and Nerlich, 2013).
2. *El Público* did not last for long and was closed after 4 years. It had the most evident leftist leanings. *El País* is considered to have a slight leaning towards the left, while *El Mundo* is considered to be politically centrist.

3. A possible explanation for this state of affairs is that our automatic procedures were not able to retrieve all of the items from earlier years, due to technical changes in the website. We navigated to random dates to check whether there had been any items left out of our collection, and found none. We believe that this spectacular growth in news items is the result of the newspaper's policy to expand its online version. On many occasions, the newspaper's directors declared their great interest in augmenting the paper's presence on the Internet, and according to its previous Editor-in-Chief, it now has more digital subscribers than all the other Spanish newspapers put together. In fact, if you visit the website of the newspaper today, you can clearly see it has more content than *El País,* for example. See PEDRO J. RAMÍREZ, Cambia el director, sigue la orquestra, El Mundo. Retrieved 30 April 2014: http://www.elmundo.es/opinion/2014/02/01/52ed53f122601de37a8b4575.html
4. On SVM, see http://www.csie.ntu.edu.tw/~cjlin/libsvm/ Guide to SVM, see http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. Information and Software: http://svmlight.joachims.org/

References

Abney S (2002) Bootstrapping. In: *Proceedings of ACL 2002*, Philadelphia, USA, pp. 360–367. USA: Association for Computational Linguistics.

Arlot S and Celisse A (2010) A survey of cross-validation procedures for model selection. *Statistics Surveys* 4: 40–79. Available at: http://projecteuclid.org/download/pdfview_1/euclid.ssu/1268143839

Bauer MW (2009) The evolution of public understanding of science – Discourse and comparative evidence. *Science, Technology and Society* 14(2): 221–240.

Bauer MW, Allum N and Miller S (2007) What can we learn from 25 years of PUS survey research? Liberating and expanding the agenda. *Public Understanding of Science* 16: 79–95.

Bauer MW, Petkova K, Boyadjieva P and Gornev G (2006) Long-term trends in the public representation of science across the 'iron curtain': Britain and Bulgaria, 1946–95. *Social Studies of Science* 36(1): 99–131.

Bauer MW, Ragnarsdottir A, Rudolfsdottir A and Durant J (1995) *Science and Technology in the British Press, 1946–1990: A Systematic Content Analysis of the Press*. London: The Science Museum.

Benford RD and Snow DA (2000) Framing processes and social movements: An overview and assessment. *Annual Review of Sociology* 26: 611–639.

Benoit G and Gingras Y (2000) What is scientific and technological culture and how is it measured? A multidimensional model. *Public Understanding of Science* 9: 43–58.

Bucchi M and Mazzolini RG (2003) Big science, little news: Science in the Italian daily press, 1946–1997. *Public Understanding of Science* 12(1): 7–24.

Carvalho A (2007) Ideological cultures and media discourses on scientific knowledge: Re-reading news on climate change. *Public Understanding of Science* 16: 223–242.

Castro P and Gomes I (2005) Genetically modified organisms in the Portuguese press: Thematisation and anchoring. *Journal for the Theory of Social Behaviour* 35: 1–18.

Clark F and Illman D (2006) A longitudinal study of the New York Times Science Times Section. *Science Communication* 27: 496–513.

Dimopoulos K and Koulaidis V (2002) The socio-epistemic constitution of science and technology in the Greek press: An analysis of its presentation. *Public Understanding of Science* 11(3): 225–241.

Donk A, Metag J, Kohring M et al. (2012) Framing emerging technologies risk perceptions of nanotechnology in the German press. *Science Communication* 34(1): 5–29.

Elmer C, Badenschier F and Wormer H (2008) Science for everybody? How the coverage of research issues in German newspapers has increased dramatically. *Journalism & Mass Communication Quarterly* 85(4): 878–893.

FECYT (2011) Percepción social de la ciencia y la tecnología 2010 [Social Perception of Science and Technology in Spain 2010]. Madrid: FECYT.

Jaspal R and Nerlich B (2013) Fracking in the UK press: Threat dynamics in an unfolding debate. *Public Understanding of Science* 23(3): 348–363.

Joachims T (2002) *Learning to Classify Text Using Support Vector Machines – Methods, Theory and Algorithms*. Boston, MA: Kluwer Academic Publishers.

Kornblum J (2006) Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation* 3: 91–97.

Kornblum J (2010) Fuzzy hashing and ssdeep. Available at: http://ssdeep.sourceforge.net/

Krippendorff K (2004) *Content Analysis: An Introduction to Its Methodology*, 2nd edn. Thousand Oaks, CA: SAGE.

Lancichinetti A and Fortunato S (2009) Community detection algorithms: A comparative analysis. *Physical Review E* 80(5). Available at: http://arxiv.org/pdf/0908.1062v2.pdf

Lee C and Cunningham P (2014) Community detection: Effective on large social networks. *Journal of Complex Networks* 2(1): 19–37. Available at: http://comnet.oxfordjournals.org/content/2/1/19.full.pdf+html

León B (2008) Science related information in European television: A study of prime-time news. *Public Understanding of Science* (17): 443–460.

López Cerezo JA and Cámara M (2007) Scientific culture and social appropriation of the science. *Social Epistemology* 21(1): 69–81.

Major A and Atwood E (2004) Environmental risks in the news: Issues, sources, problems, and values. *Public Understanding of Science* 13: 295–308.

Moreno Castro C (2010) Los medios, el público y la ciencia. Una relación que no progresa adecuadamente [The media, the public and science: A relationship which is not advancing adequately]. In: *Percepción Social de la Ciencia y la Tecnología en España, 2008* [Social Perception of Science and Technology in Spain, 2008]. Madrid: FECYT.

Novak B, Mladenič D and Grobelnik M (2006) Text classification with active learning. In: *From data and information analysis to knowledge engineering: Proceedings of the 29th annual conference of the Gesellschaft für Klassifikation eV*, University of Magdeburg, Magdeburg, 9–11 March 2005, pp. 398–405. Berlin, Heidelberg: Springer.

Olston C and Najork M (2010) Web crawling. Invited survey article. *Journal of Foundations and Trends in Information Retrieval* 4(3): 175–246.

Plantié M and Crampes M (2013) Survey on social community detection. In: *Social media retrieval*, pp. 65–85. Available at: http://hal.archives-ouvertes.fr/docs/00/80/42/34/PDF/Survey-on-Social-Community-Detection-V2.pdf

Pons P and Latapy M (2005) Computing communities in large networks using random walks. In: *Computer and information sciences (ISCIS)*, pp. 284–293. Available at: http://arxiv.org/abs/physics/0512106

Quintanilla MA (2005) *Tecnología: Un enfoque filosófico y otros ensayos de filosofía de la tecnología* [Technology: A philosophical approach and other essays on the philosophy of technology]. México, D.F., México: Fondo de Cultura Económica.

Quintanilla MA (2012) Cultura, Tecnología e innovación [Culture, Technology and Innovation]. In: Aibar E and Quintanilla MA (eds) *Ciencia, tecnología y sociedad. Enciclopedia Iberoamericana de Filosofía* [Science, Technology and Innovation. Iberoamerican Encyclopaedia of Philosophy]. Madrid: Trotta, pp. 103–136.

Raghavan UN, Albert R and Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* 76. Available at: http://arxiv.org/abs/0709.2938

Rosvall M, Axelsson D and Bergstrom C (2009) The map equation. *European Physical Journal Special Topics* 178: 13–23.

Salton G and McGill MJ (1983) *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill.

Schafer MS (2009) From public understanding of science to public engagement: An empirical assessment of changes in science coverage. *Science Communication* 30(4): 475–505.

Sebastiani F (2002) Machine learning in automated text categorization. *ACM Computing Surveys* 34(1): 1–47.

Suerdem AK, Bauer M, Howard S and Ruby L (2013) PUS in turbulent times II: A shifting vocabulary that brokers inter-disciplinary knowledge. *Public Understanding of Science* 22(1): 2–15.

Šuljok A and Brajdić Vuković M (2013) How the Croatian daily press presents science. *News Science & Technology Studies* 26(1): 92–112.

Vapnik Vladimir N (1995) *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Veltri G (2013) Viva la Nano-Revolución! A semantic analysis of the Spanish national press. *Science Communication* 35: 143–167.

Vincent D, Blondel VD, Guillaume JL, Lambiotte R and Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10. Available at: http://arxiv.org/abs/arXiv:0803.0476

Vogt C (2011) The spiral of scientific culture and cultural well-being: Brazil and Ibero-America. *Public Understanding of Science* 21(1): 4–16.

Author biographies

Tamar Groves is a lecturer at the Education Sciences Department, Extremadura University. She is a former researcher of the Institute of Science and Technology Studies, Salamanca University, where she has been working on projects related to scientific culture, higher education, teacher education and the history of science and education.

Carlos G. Figuerola is a lecturer at the Computer Science Department and member of the Institute of Science and Technology Studies, Salamanca University. He specialises in Information Retrieval, Text Mining and Analysis Techniques of Social Networks. His research activities lie currently in applying information technologies to Humanities and Social Sciences.

Miguel Á. Quintanilla is Chair of Logic and Philosophy of Science, Salamanca University, and the director of the Institute of Science and Technology Studies of the same university. He published extensively on Philosophy of Science and Technology, Scientific Culture, Science and Society, Science Policy, Science Communication and Higher Education.