
**M.W. Bauer and A. Suerdem**

# Developing science culture indicators through text mining and online media monitoring

## Conference Item

OECD Blue Sky meeting on Science and Innovation Indicators

Ghent, 19-21 September 2016

# Developing Science Culture Indicators through Text Mining and Online Media Monitoring

Martin W. Bauer (LSE) & Ahmet Suerdem (Bilgi University, Istanbul)

M.Bauer@lse.ac.uk

Ahmet.Suerdem@bilgi.edu.tr

This paper is relevant to the theme of the conference: *"The Internet and big data analytics as a source of data on STI: opportunities and challenges"*

List of contents

## 1. Introduction: background and project MACAS

In this paper we will present our ongoing research arising from the international project MACAS (mapping the cultural authority of science; http://www.macas-project.com/ ) and offer some guidelines for producing cultural indicators by means of media monitoring. The aim of the MACAS project is to develop a cultural indicator system through text mining and content analysing the changes in the intensity and contents of Science Technology and Innovation (STI) coverage in the online mass media. Analysing mass media is essential for understanding the nature and pace of the transformation occurring in the cultural sphere since they are the major source of mass-production of the common symbolic environment. Mass production and distribution of message systems have important cultivation functions as they provide the individuals with a common language and symbol system to make sense of the events surrounding them. Major transformations and trends occurring at societal, economic, political and cultural levels are sooner or later translated into linguistic patterns reflecting these changes. Representative abstraction from the collectively experienced total texture of messages can provide the social researchers with impeccable tools for developing cultural indicators to understand how public is cultivated during this process (Gerber, 1969). Hence, the systematic analysis of linguistic patterns residing in mass media texts offers a promising possibility for complementing the survey and poll data for measuring public opinion. High cost of the latter instruments does not permit to collect data at short intervals. Automatized media monitoring and text mining can provide data at monthly, weekly or daily regular intervals without extra costs. Furthermore, unlike surveys, media monitoring is not constrained for ex-post-facto data collection and can retrieve past information residing in digitalized archives. This provides opportunity to construct STI cultural indicators retrospectively for the times when data were not collected by surveys on a regular basis. While surveys and polls are the usual ways for the production of science culture indicators (such as Eurobarometers), complementing the structured information with unstructured textual information is essential for understanding how science is represented in the public mind. There have been some major efforts in the past to quantify the textual data for the historical analysis in terms of summary indicators measuring how ST&I related subjects are reflected in the mass media (e.g. Bauer et al., 1995). The text and data used during these analyses yield a tangible, openly available resource (the Media Monitor Archive) which can be accessed in British Science Museum. However, these projects are difficult to replicate and update as they require costly and time consuming intensive human coding of the documents. Web revolution offers a great opportunity to overcome these shortcomings. Media texts are now increasingly available in the Web environment and are accessible via digital archives. Accordingly, the research on how media map science over time has become widely available (see Bauer, 2002; Bauer et al, 1995; Bauer Petkova et al, 2006; Bucchi and Mazzolini, 2003).

Existing science in the media monitoring studies either focus on qualitative content analysis and interpretive inquiries or simple automatic counting of keywords in ST&I relevant online articles. The Brazilian SAPO (Vogt et al, 2012) and Italian SMM (Science in the Media Monitoring; Bucchi and Neresini 2012) projects are pioneers in automatic science monitoring. Although these efforts are important contributions, they only produce intensity measures and are still far from providing text mining features for transforming textual data into a comparative indicator system. They do not collect historical data but accumulate the texts as they become online and only provide a superficial reading of the textual data without extracting significant information. While digital tools for qualitative-interpretive analysis

based on human reading and understanding can help to overcome these shortcomings, such tools are not suitable for handling big data. Our ongoing research faces this challenge by combining techniques such as text mining, Natural Language Processing (NLP), and machine learning with qualitative interpretive methods such as coding and retrieving. Mixed methods in terms of applying qualitative and quantitative approaches to text analysis is in general an emerging topic and new to the Science in the Media Monitoring studies. Our project aims to contribute to the science culture indicator development efforts from a mixed method perspective.

At its present stage, MACAS project completed the collection of the historical data for several countries. At the rest of the paper, after presenting our preliminary findings for UK and Turkey, we will share the lessons drawn from the MACAS project to develop guidelines for producing indicators with the use of sophisticated NLP and text mining techniques. We will provide a step-by-step explanation of these guidelines through a case study of biotechnology.

## 2. Science news in the UK and Turkey

### 2.1 Science news in UK, 1990-2013

#### 2.1.1 The Text Corpus: The Collection of the Press Sample

Press sample was collected between 1990-2013.Two newspapers were selected: Times and Mirror. The TIMES is considered a 'Quality Paper' with a large traditional readership and the MIRROR is a 'Popular Paper' with a larger readership. The TIMES is owned by the Murdoch Group and traditionally centre-right on the political Left-Right spectrum, situated between the Guardian on the left and the Telegraph on the right. Its daily circulation runs between 500,000 and 1 Million declining. The MIRROR is traditionally a left-leaning popular paper with a larger circulation between 1-2 Million, initially over 5Million readers. The MIRROR was and remains the anti-voice to the Daily Mail on the Conservative side of the spectrum. For both papers, circulation figures and readership are declining since the early 1990s, in the case of the MIRROR more dramatically so than for the TIMES (see figure 4 above). The news sample was collected through systematic sampling of two artificial weeks of 14 days annually.

For every sample date and news outlet, the 'science news' items were selected from NEXIS/LEXIS, an on-line newspaper archive, with Boolean keyword (KW) queries. The relevant keywords were identified by means of taking out 10 days of newspapers and identifying manually the terms relevant to science. However, as this strategy did not return accurate selections, the news items were selected from NEXIS/LEXIS own classification of 'Science & Technology'. This still contained some noise, however as the sensitivity analysis produced more or less robust results, we concluded that this would be satisfactory for analysis purposes. Improvement of the selection method is left for future studies. We will address this problem in section *3: Thinking and working beyond MACAS.* Sensitivity analysis is conducted by generating two subsets of keywords and selecting two different samples of science news. Invariance or small variance of results across samples would suggest that the

results are internally robust. Complete corpus contains n=16,779 science related news.

### 2.1.2 Intensity and topic flow in UK, 1990-2013

**Figure 1. Intensity of science news in UK 1990-2013**

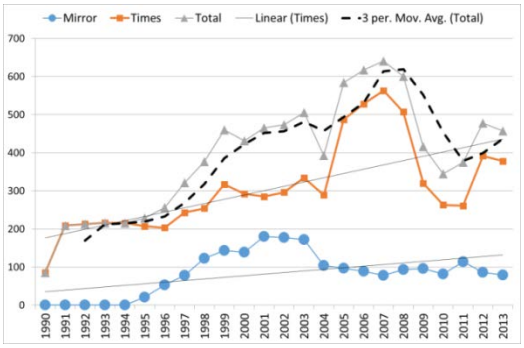


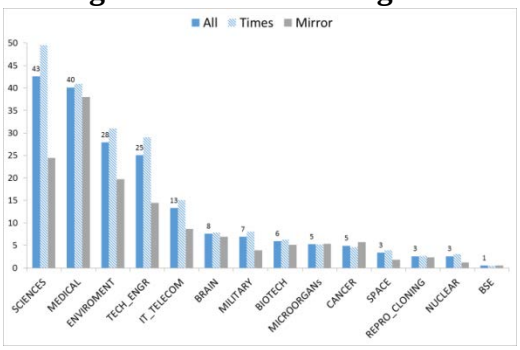

Figure 1 shows on the left, the intensity of science news flow over the period in terms of text units in the corpus. Science news increases from 100 units to over 600 units by 2006-2008, and it since declined to about 400 units. Science news is generally more intense in the quality press than in the popular press, which is what we expected from earlier research (Bauer et al, 1995); this quality-popular gap continued unabated at a ratio of about 2-3 articles in the TIMES on every articles in the MIRROR. However, quality and popular press have a different cycle: the popular press peaks in the early 2000s, while the quality press peaks in 2006-2008. The long-term trend is towards increasing science news over the period, despite the cycles of ups and downs.

### 2.1.3 Thematic classification: themes and special topics

News relevant to a particular theme is selected by means of automated content analysis dictionaries including the keywords representing the domain category. Accuracy of the representation of the keyword selected news to their relevant categories changes between 81 % and 83 %. Sensitivity analyses produced more or less robust results. As we will come back to selection strategies in section 3, we are not reporting detailed procedures.

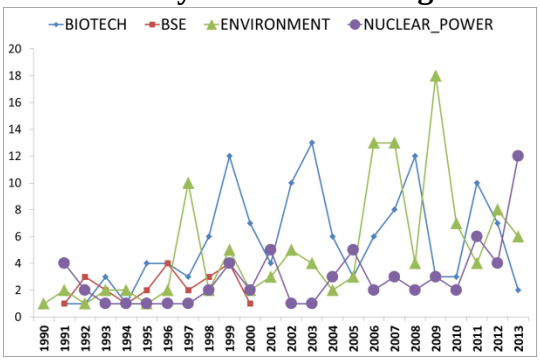
The thematic categories ‘medical’ and ‘general science’ are much more frequent in the corpus. With regard to the ‘medicalisation hypothesis’ (dominating trend in the intensity of medical news) we find, that in the quality press, here the Times, medical news does not involve the dominant news items, it remains superseded by ‘science in general’. However, as observed earlier studies, medical news dominates the popular press, here the Mirror (figure 2).

**Figure 2 Thematic categories**




Among the mid-range topics, which there are biotech, nuclear, environment and BSE, some themes have higher and others lower frequency all through the period. Biotech moves in the opposite trends, increasing to a peak in the mid-2000s and since on the retreat. We will make a more detailed analysis of this in section 3. Nuclear topic seems to have a constant flow with the occasional peak, but little long-term trend, which in the case of nuclear power is rather surprising, as we would have expected a continuous rise in the 2000s and certainly after Fukushima 2011, but not in the UK apparently (figure 3).

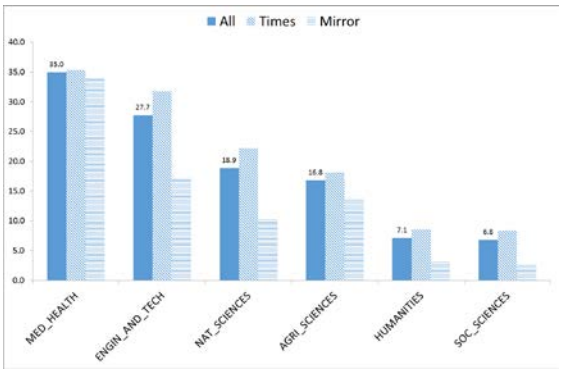
**Figure 3 Intensity of thematic categories in UK**




### 2.1.4 The Disciplines of Science in the News (Frascati categories)

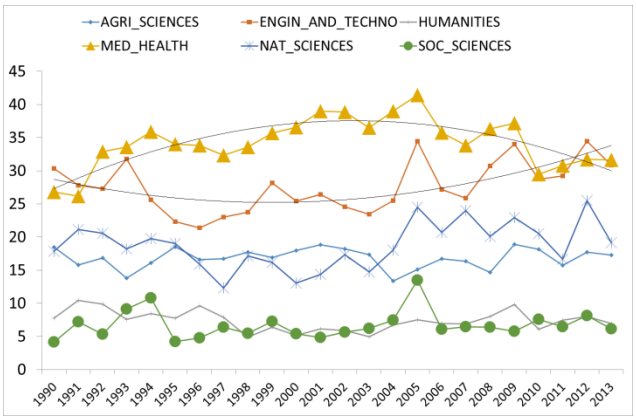
The accuracy of the selection by categorization dictionaries was lower compared to the thematic selection (between 68 and 73 %). We are considering six scientific disciplines following the categorisation suggested by the OECD in the Frascati Manual. These include: (1) agriculture, (2) engineering and technology, (3) medical and health, (4) natural sciences, (5) humanities and (6) social sciences. The latter two are treated as separate categories which is useful for our purposes (figure 4).

**Figure 4 Science disciplines**




Our results show that the differences between news outlets are not very striking. The overall ordering is more or less preserved, while the mid-field of national sciences and agricultural sciences is reversed. The popular press carries more agriculture than natural science news. Maybe the fact is striking that the popular press is much more focussed on Medicine & Health, than on other disciplines compared to the quality press (figure 5).

**Figure 5 Intensity of science disciplines**




Flow of disciplines over period is rather stable. Only Medicine & Health and Engineering and Technology display a certain trending in opposite directions.

## 2.2 Science news in Turkey, 1997-2014

### 2.2.1 The Text Corpus: The Collection of the Press Sample

For collecting the science and technology relevant news in Turkish media, we followed a different strategy. First, we built a web crawler that harvested all the news in Hurriyet newspaper since it came online in 1997 until August 2015 and put them in a database. We selected Hurriyet because it is considered to be the flagship and agenda setter of the Turkish media with 1.2 million readers of which 662 thousand are university or higher graduates. Its target audience is mostly from AB socio-economic status group who are car and home owners and consumers of high technology, financial and travel products.

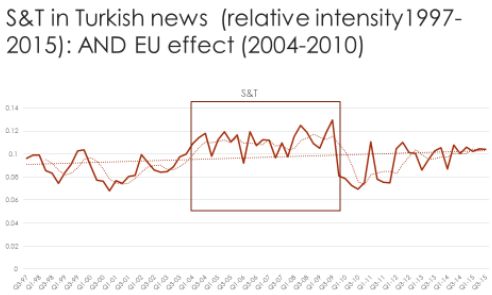
This news corpus contains around 5.000 news per month (60K p/a; 1 M whole corpus). It contains approximately 250 million tokens and 2.8 million unique types. Then we compiled a

dictionary containing science related keywords. When selecting the keywords, the idea was to consider S&T as a cultural phenomenon: an ecosystem of representations and concepts concerning to society hence mirrored in the mass media. We selected 417 Keywords reflecting cutting-edge research technologies, aerospace technology, astronomy, policy debates, life and health sciences, medicine and environmental issues. We did not include social sciences because of the difficulty of disambiguation of the keywords. We then used this dictionary to identify and keep STI news and filtering out unrelated news. To fine tune the queries, instead of delivering all mentions of a specific keyword, we used rule based Boolean logic in to deliver a specific subset of articles such as: Monsanto AND (biotech OR genetic OR seed OR frankenfood ) AND NOT herbicide within same paragraph. Altogether, science news make about 10 % of all the news, oscillating between 7 to 13 per cent between 1997 and 2015 (figure 6).

### 2.2.2 Intensity and topic flow in Turkey, 1997-2015

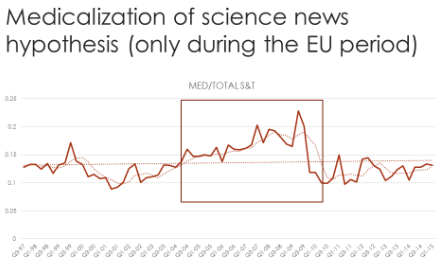
When we analyse the trend in intensity of science news, we can observe that moving average moves over the trend line in 2004 when Turkey's EU candidacy is accepted and moves under the trend line when both the Turkish government and EU have started to temporize on delivering the accession criteria after 2010. This is an interesting finding as it shows the effect of EU agenda on the salience of STI news in the mass media.

**Figure 6 Intensity of S&T news in Turkish media 1997-2015**



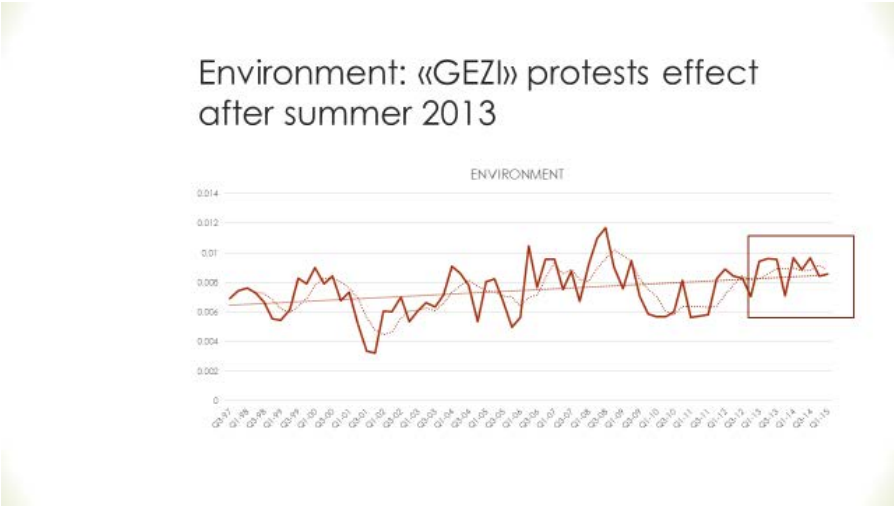

Another interesting finding is the medicalization of science news (increasing trend in the intensity of medical news). The trend line for medical news becomes much steeper during the EU agenda period compared to ST news in total. However, this effect reduces after the cooling off the EU agenda (figure 7).

**Figure 7 Medicalization of science news**




We can also observe how larger political events can effect and change a science relevant topic trend in the news. At the end of May 2013, protesters gathered to protest government's decision to construct a shopping mall to one of the few left green areas in central Istanbul. Violent police interventions attracted hundreds of thousands of people to protest the government and turned into large scale demonstrations. The protests lasted for months increasing awareness about environmental issues. Moving average for the intensity of environment related news moves over the trend line after the second quarter of 2013 and remains so until 2015 (figure 8).

**Figure 8 Environment news in Turkish media**




## 3. Thinking and working beyond project MACAS

Indicators developed by means of MACAS analyses are far from perfect, and mostly depend on ad hoc experimentation rather than a systematic approach. In the rest of the paper, we will highlight what we have learned from project MACAS for developing a more systematic, and therefore more accurate approach to media monitoring for developing science culture indicators.

### 3.1 Existing methods of media monitoring

Media monitoring is the process of observing media news flow on a continuing basis to identify, capture and analyse contents containing specific topic keywords. While historically a matter of content analysis methodology that was developed during the 1940-60s Cold War effort for purposes of intelligence and propaganda analysis (Krippendorff, 2012), it recently

revived as a tool for Public Relations, Marketing and Corporate Intelligence .While traditionally it had been dependent on human reading and coding of press clippings of the relevant articles, the process is now semi-automatized thanks to digital media content and advances in computational linguistics and text mining techniques. However, the traditional criteria of transparency, reliability and validity of content classification remain the sticky points of all analysis.

The traditional bottleneck of data sourcing has been overcome as the worldwide web provides a fertile ground for harvesting a wide range of global media contents as most press and broadcast outputs are also published online. Moreover, availability of online search engines such as Google News or LEXIS/NEXIS make media monitoring accessible to wider communities with up-to-the-minute news coverage, keyword search facilities and well organized archives ordered by date and topic categories. While not directly intended for social research but rather for corporate intelligence, these tools nevertheless offer impeccable opportunities for social researchers for constructing and analysing corpora of text data to indicate the flow and trends of published opinion.

However, free online services come with a cost for the social researchers. First, they are not suitable for complex queries since they allow to search for a limited number of keywords at one go. Second, the relevance of the search results reflect the logic of Google algorithms rather than the topic of scientific research. Third, the retrieved texts in search results contain redundant content requiring to filter out duplicate ones manually. Last but not the least, they don't include a way to store retrieved content, the text content needs to be manually separated from other content such as photos or ads by copying and pasting into a database or spreadsheet. Since these drawbacks require substantial time and staff investment, they make high quality media monitoring a tedious and inefficient task for the solo social researcher.

Online news monitoring services such as Cyberalert (http://www.cyberalert.com/) are offered as a solution to these drawbacks. These services offer sophisticated features such as automated search queries; advanced Boolean queries to increase the relevance of the retrieved texts; online archives allowing to save, search and organize the text excerpts; automatic translation services; and analysis tools such as graphs and charts. However, these services are mostly designed for business intelligence gathering and their subscription is expensive. They do not meet the requirements for a system designed for public surveillance and research service accessible by all citizens and researchers. Considering this gap, in the rest of the paper we will delineate some guidelines for constructing such a system allowing to monitor issues concerning public engagement to science and responsible research and innovation.

### 3.2 Need for an open access science in the media monitoring system:

Continuous monitoring of public STI-related information sources and early detection of adverse events may provide feed-back and information for the policy makers, science communicators and citizens to act upon challenging issues concerning public. Important STI related news is increasingly available on the World Wide Web in electronic form, and has been shown to be a useful data source for monitoring trend in public opinion towards science. Building upon our experiences in the MACAS project, will offer practical guidelines for a Science in the Media Monitoring System (SMM) that automatically collates, annotates and categorizes news articles from multiple open news sources. This automatic approach is more convenient than the traditional labour-intensive and time-consuming ways of science news

content analysis still applied by many researchers in the field of science communication. A system relying the state of the art artificial intelligence technologies such as data mining, machine learning and natural language processing will help researchers to construct indicators to monitor public attitudes towards science and technology through media data. This system will focus on automatically harvesting STI related news from the web; organizing the information obtained into different categories; topic detection and sentiment analysis.

### 3.3 The guidelines for generating indicators by means of Media monitoring (MM)

In this section, building upon our experiments in MACAS study we will offer some methodological guidelines for future media monitoring studies. These guidelines are not definitive rules but aim to provide an informative point of departure for continued and open research for a science in the media monitoring system providing accurate and timely information to policy makers and public.

Box: Methodological steps are as follows:

a. Crawling the web for harvesting the online news to collect a domain specific corpus.
b. Calculating and visualizing the quantity and the quality of press coverage of the domain of the interest measured as cultural indicator.
c. Developing a sentiment analysis system enabling the monitoring of the attitudes towards the evaluation of these topics.
d. Providing a basis to explore systematically the relationship between press coverage of ST&I and the political, cultural and social contexts through topic detection techniques such as *Latent Dirichlet Analysis*.

#### 3.3.1 Crawling the web for harvesting the online news to collect a domain specific corpus.

The first step in media monitoring is collecting the texts representing the language and jargon of the interest domain. This is commonly done by making a dictionary of keywords and search the web for the texts containing these keywords. While the common practice for social researchers has been to make these queries manually, automatic web crawlers are increasingly being used to alleviate the burden. Crawlers start from a seed web link and follow the hyperlinks within it to go to other pages. This process snowballs with new hyperlinks to follow until a saturation point is reached or a predetermined objective is met. Once the links are collected, they are kept in a database and text units containing the relevant keywords are retrieved by means of queries. Finally, an HTML scraper identifies and collects the textual article content from each of these web links cleaning out the unnecessary parts such as pictures.

Although automatically harvesting the texts and using the web as a corpus “surrogate” seems to be an attractive idea, this might be scant because of the low level of control on data. The texts collected this way typically do not convene with sampling design principles and often represent opportunistic samples of the data. The balance and representativeness of the web language is questionable (Ide et al., 2002). While some treat web as the largest linguistic corpus in the world, it is hard to assert that it represents any focused language because of the

information overload. A corpus is made of *complete* and *thematically unified collection of texts with some research objectives in mind* (Atkins et. al, 1992). Without a sound corpus design strategy, the signal-to-noise ratio can be excessive. A careful selection of the media type, text sources, type of articles, time period, as well as identification of keywords is essential for constructing a well-balanced and representative corpus. Without an intelligent and actionable corpus construction strategy, just automatizing the keyword search would follow the naïve approach of manual search strategy, albeit in an unmanageable manner. Unfocused Web crawlers will harvest all texts on their way through the hyperlinks and produce massive document collections on a wide range of topics representing different languages and jargons. A focused crawling requires a meticulous corpus construction strategy where the corpus collector controls the topical and temporal restrictions from an "expert" point of view about the subject domain of the interest.

Therefore, an ad hoc "keyword" based query strategy by itself would not retrieve efficient and relevant collection of texts. From a content analysis perspective, keyword dictionaries can scan and classify the documents according to some predefined categories. However, classical content analysis does not treat the texts at discursive level concerning how the meaning is embedded in the context. Texts are not just meaning containers but are complex discursive constructions embedded within a social context. A corpus does not only contain information about dependent variables such as the contents of the messages, topics, themes, keywords but also contain insight on complex contextual information reflecting the organization of a social system (Biber et al., 2007). External independent variables such as type of media channels (i.e. newspapers, TV broadcasts, scientific articles, blogs, social media etc....); properties of the broadcaster and audience; time of broadcast; genre (i.e. news, comment, technology, culture etc.) and functions (i.e. persuade, express or inform) as well as internal independent variables such as linguistic regularities about social practices, institutional practices, framing of events; and agenda setting strategies operate together to align the content with context. Content analysis is concerned with linguistic features which as such have little interest to the social scientist. The linguistic information discovered by means of content analysis only provides researcher the tools for investigating patterns in social practices; values; representations or structures based on naturally produced textual data. On the other hand, the contexts where the texts are embedded have their own discursive formations, and it is reasonable to hypothesize that all texts from a common domain will tend to share similar patterns of discourse organization. Therefore, studying the text at discursive level requires focused corpora with a purpose in mind rather than taking web for granted as a linguistic corpus. Designing a balanced corpus representative of and focused to a particular domain requires meticulous interpretive touch of a domain expert for investigating how society functions and comprehends itself through its texts (Bauer & Aarts, 2008). Texts in a corpus should be collected in a way to reflect only one thematic focus in representative and balanced way.

This procedure is akin to qualitative purposive sampling than quantitative statistical sampling. Statistical sampling designs by themselves usually are not convenient for collecting a corpus. Traditional sampling theory stands on the assumption that the properties of sampled units are distributed in the same way as in the population. The paradigm behind this assumes that units are independent from each other and their probabilistic selection from a population would guarantee the representativeness of the sample. According to Krippendorf (2012), assumptions of traditional sampling theory does not fit to textual analysis as the collection of

texts need to consider more than one populations: the population of texts containing the content relevant to the study and information answering the objectives set by the thematic focus of the study. Hence, textual analysts are rarely interested in probabilistic representativeness of the population. Their concern is to construct a corpus providing information relevant to the thematic focus of the study. Standard approaches to statistical sampling are hardly applicable to corpus construction for linguistic studies (Atkins et al., 1992: 4).

Representativeness in corpus design is maintained across two important axes: first the collection of texts in a corpus need to fulfil the representativeness according to the external criteria such as date, genre, registers or functions reflecting the contextual strata where the language is used. Second, the corpus should include the linguistic variations reflecting the internal distributional patterns concerning the language of the study (Biber, 2007). According to Biber, corpus linguistic studies need to be discourse analytical as well since they investigate systematic patterns of language use across discourse contexts. Discourse analysis is the study of linguistic patterns beyond syntactical sentence level focusing on how the extended sequences of utterances are constructed as a text to fulfil a social function and how those texts are organized in systematic ways. This second dimension is mostly concerned about how language reflects social representations covering a repository of cultural values, world views, symbols, beliefs and practices reflecting the life world of a community rather than the grammatical structures as such(Bauer & Aarts, 2000).

Maintaining the first dimension is relatively straight forward. The usual procedures for statistical sampling can apply without much problem. Strata, functions and registers include the common sense criteria external to the thematic focus of the study which are more or less transparent. However, maintaining the second dimension cannot be accomplished in a straight forward way as the representational variety at the discursive level is unknown in advance. On the one hand, a corpus needs to be representative of linguistic patterns used by a community, on the other hand we need to determine these patterns from a representative corpus. This paradox, called 'corpus-theoretical paradox' (Schmied, 1996: 192) can only be resolved through an iterative-cyclical process. A representative corpus cannot be identified in advance and requires a stepwise procedure starting with a preliminary selection of texts, identifying the variety in them, and extending the corpus until no additional variety is detected. In other words, corpora emerge in a controlled manner during the use rather than being constructed for once ready to be used.

*A working example: biotechnology corpus (2000-2015)*

The ultimate objective of the MACAS project is to construct a corpus covering all the news pertaining to STI. However, identifying the boundaries of the domain of interest is extremely hard, if impossible since the discourse on STI is highly complex, dynamic and fuzzy. For example, an innovative object such as a mobile phone which once had been considered as a high-tech concept can then become an everyday artefact in considerably short time. A highly scientific concept such as stem-cells can also be the subject of political or moral debates. Therefore, in line with our definition of corpus construction as an open and dynamic process, thematically focusing on the domain would emerge as an iterative-cyclical process. As in corpus linguistics, we can renounce any hopes for a general-purpose corpus fully representing a domain of interest. A multitude of topical corpora may emerge from a flourishing practice

of corpus construction. As a principle, we suggest to start from narrower sub-domains that are relatively easy to control and focus, and enrich the corpus by identifying the inner sub-domains and outer larger domains of these sub-domains. In this paper, we start with the sub-domain of "biotechnology".

*i. Identifying the external criteria.*

In order to overcome the initial paradox of corpus construction, we started with external strata and functions. First step for identifying the external criteria is to determine the units of analysis. Generally, units are the wholes that analysts distinguish and treat as independent elements. Determination of the units is a statistical procedure where the usual sampling practices apply. In statistics, a "sampling unit is one of the units into which an aggregate is divided for the purpose of sampling, each unit being regarded as individual and indivisible when the selection is made" (The International Statistical Institute, 2003). We identified our sampling units as "press news" items on the premises that agenda setting press leads public opinion. Press have long been a major organ of the public opinion in terms of reflecting and shaping the aggregate expression of individual worldviews, attitudes, and beliefs about a particular issue. Traditional media in general has an agenda-setting function by highlighting the salience of some issues and concealing the others. Compared to recently emerging social media, information arising from traditional media is relatively more transparent and accountable as they are more subject to regulations. Within traditional media, press plays the major role in agenda setting and comparative studies show that science and technology content of press, TV, and radio do not significantly differ (Hansen & Dickinson, 1992), only magazines have a very different content structure. Hence, for the media monitoring efforts of most organizations, news monitoring is the core service.

Second step for identifying the external criteria is the determination of the criteria for selecting the strata from which the sampling units should be collected. During this step, the entire population is divided into mutually exclusive subgroups or strata where every element should be assigned to one stratum. The criteria for determining the subgroups can be multi-dimensional to achieve the maximum level of diversity. We identified the population where the news will be collected as the newspapers. Obviously, diversity in the selection of newspapers is very important since every newspaper has a different worldview, language, genre and audience. For example, some newspapers express their views through a more factual language while others can concentrate on more subjective languages. To determine the strata, we identified four criteria: if the newspaper belongs to popular or quality press; represents left or conservative worldviews; is in the leader or follower category, and have high or low readerships. For the purposes of this study we only considered opinion leaders with high readership numbers. The selected newspapers are (UK) Telegraph (quality, conservative); Independent (quality, centre); Guardian (quality, left); Mirror (popular, left) and Daily Mail (popular, conservative). At the time of the research we had the full corpus of only Telegraph and Daily mail and our analyses will be built upon this corpus.

Instead of selecting a sample, we crawled all the links in the newspaper archives referring to every single news since the newspaper came online. We then scraped the metadata and news content from the html structures and saved them in a database.

### *ii. Identifying the internal criteria*

Determining the linguistic patterns reflecting how values, world views, symbols, beliefs and practices concerning the social representation of STI is much less straight forward than identifying the sampling units according to external criteria. Classifying specific chunks of content into predefined categories covering these patterns is essential for constructing a thematically focused corpus. These chunks are typically contained in the news, at most coinciding with them, but never exceeding them. In other words, they are the parts of the news item relevant to the subject of study. A news item can cover not at all, partially or totally the categories reflecting thematic focus of the study. Thematic focus of a corpus can be defined as the combination of the categories making its storyline (Baker, 2006). For manual content analysis, these categories can be determined within a coding frame and human readers read the text, highlight the text segment comprehending any of these categories and code the highlighted segment. For automatized content analysis on the other hand, identifying any text segment reflecting a thematic category is accomplished through categorization dictionaries covering the relevant keywords. Machine identifies if a text segment belongs to a certain category according to the occurrence or frequency of the keywords in the dictionary in this segment. Determining the categories and their corresponding keywords is a challenging task as these are not only grammatical categories but also social categories representing the thematic focus of the study. As linguistic patterns represent social categories corresponding to abstract domains such as worldviews, values and the like, it is very difficult to determine the linguistic variety in these domains in a way external to the text. Although it is possible to construct dictionaries representing these categories in a top-down, deductive manner, applying these dictionaries outside the domain for which they were developed can lead to serious errors. For example, General Inquirer project created categories to represent social-science concepts of several grand theories that were prominent at the time the system was first developed, including those of Harold Lasswell, R.F. Bales, Talcott Parsons and Charles Osgood (http://www.wjh.harvard.edu/~inquirer). Such top-down content analysis dictionaries are criticised on the grounds that complexity of language does not permit to produce accurate accounts of text selection. '[T]he linkages between content analysis and linguistics have been generally tenuous' (Markoff, Shapiro and Weitman, 1974:8), since words in discourse may only get their meanings in the context in which they occur. Automatic pattern detection algorithms such as cluster analysis can overcome this problem since they determine the categories and their relevant keywords in a bottom-up, inductive manner from within the corpus representing the discourse of the domain. These algorithms simultaneously estimate the categories and then classify keywords into those categories. However, there is no guarantee that these algorithms will return theoretically interesting and valid categories. Furthermore, inductive techniques are prone to "corpus theoretical paradox" since on the one hand, the corpus needs to be representative of the keywords, on the other hand we need to determine these keywords from a representative corpus.

Therefore, a representative corpus cannot be identified in advance and requires a stepwise procedure starting with a preliminary selection of texts, identifying the variety in them, and extending the corpus until no additional variety is detected. For determining the preliminary keywords covered by the biotechnology domain, we used Wikipedia "Biotechnology" entry as a seed site and followed selected hyperlinks to other pages such as "genetic engineering",

"bioculture" and the like. We then built a mini-corpus of biotechnology and selected the relevant keywords from this corpus. The steps for selecting the keywords is as follows: we pre-processed the corpus for the most common and basic Natural Language Processing (NLP) tasks, such as tokenization, sentence segmentation, part-of-speech tagging and named entity extraction. We filtered the nouns, adjectives and named entities to select most frequent 150 keywords which we think are relevant to biotechnology. We then classified these keywords into different facets making the ontology of the biotechnology domain. An ontology is "an explicit conceptualization of a domain of discourse, and thus provides a shared and common understanding of the domain."(Gruber, 2008). A terminology is first developed to provide a controlled vocabulary for the subject area or domain of interest, then it is organized into a taxonomy where key concepts are identified, and finally these concepts are defined and related to create an ontology. The arranged groups of elemental classes making up the scheme are called facets. Facets can be construed as perspectives, viewpoints, or dimensions of a particular domain. A faceted scheme, therefore provides a controlled vocabulary in the form of keywords arranged systematically by facets and a set of rules on how to combine such keywords to define conceptual categories (Ranganathan, 1967). We have organized the ontology into the following core conceptual facets: Generic (i.e. genome), organisms (i.e. enviropig), applications (i.e. cloning), and peripheral facets related disciplines (i.e biomedicine), biology (i.e. genetics), tools (ie Bioreactor), companies (Celera), and organizations (i.e. OECD).

Rule based keyword queries are much more accurate and efficient in terms of determining the relevance of a text segment to a certain category. Boolean logic queries such as: Monsanto AND (biotech OR genetic OR seed OR frankenfood OR "round-up ready") AND NOT herbicide… can deliver relevant subset of texts more accurately compared to delivering "all mentions" of a specific keyword. While the repetition of a certain keyword many times in a text increases its frequency, this information can be redundant. For example, a single co-occurrence of (Monsanto AND genetic) can give much more information about the text than the occurrence of Monsanto twenty times. However, constructing deductive rule based models are extremely difficult as there can be an infinite number of combinatory possibilities for relating the keywords. Ontological schemes can overcome this shortcoming as they accomplish queries on a semantical basis comparing conceptual facets rather than single keywords. Building rules relating different facets is much more simpler than generating different combination of single keywords, as these facets already include the keywords. In our case, our rule was: if (Generic OR organisms OR applications) AND (Generic OR organisms OR applications) => biotechnology. We increased the relevancy score of the retrieved texts if any of the keywords in the peripheral facets is hit. After collecting the news as HTML pages by the crawler, then we used this ontological rule to filter out unrelated news and retrieved 1158 biotechnology related news. After building the biotechnology corpus, we made quality check through human reading which returned highly relevant results.

### 3.3.2 Detecting linguistic patterns and developing cultural indicators

Mass-produced texts reside linguistic patterns reflecting how the elements of collective cultivation are represented in public message systems. They highlight how information about existence, priorities, values, and issues are integrated into frameworks of cognition. A high

level of abstraction and selection is necessary to encompass particular classes of statements. This level can be obtained through detecting patterns in the images, practices, and language of the most widely shared (i.e. mass-produced and rapidly distributed) message systems of a culture. Hence developing cultural indicators is basically a pattern detection exercise for discovering the representation of the issues in the mass-media. According to Gerbner(1969), general analysis of public message systems and developing cultural indicators is concerned with the following analytical measures: 1) attention, 2) emphasis, and 3) tendency and 4) structure of the cultivation of collective notions. While the first three describe the composition of the system from singular elements and how they are distributed in it, the last one describes how they are assembled or associated to one another. Terms of analysis for these measures are: 1) distribution, frequency of attention; 2) ordering, scaling for emphasis 3) measures differential tendency, how things are presented, and 4) contingencies, cluster, structure. In the following paragraphs we will operationalize these measures by means of different natural language processing and text mining methods.

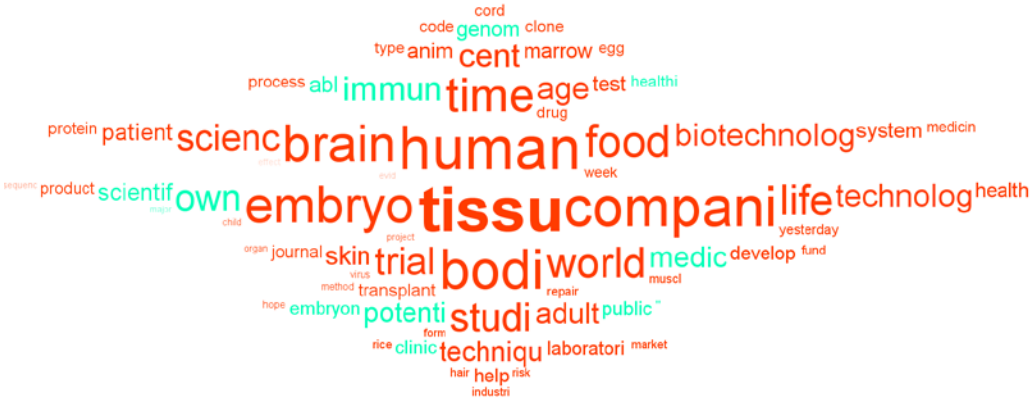
#### 3.3.2.1 Distribution, frequency of attention: What is represented?

Some issues within a domain may attract more media attention than the others. The domination of an issue as a focus of attention during the process of representing a domain cultivates assumptions about the "common sense" understanding of this domain. Bringing forward some symbols over the others during the representation process provides the members of a community with a common code to discuss, understand and hence construct social realities (Moscovici, 2000). Reformulating the unfamiliar into the words of everyday use helps the public to turn something abstract into concrete (objectification) and ascribing meaning to new phenomena by means of grounding them to existing vocabulary (anchoring). Indicators of such media attention can be calculated as the frequencies of what kinds of subject elements are called to the attention of a community. Distribution of these elements within a corpus provides information about what exists as an item of public knowledge and assumptions about the nature of existence in a domain of knowledge. (Gerbner, 1969) The most obvious indicator for understanding the distinguishing features of a subject domain is looking at the distribution of the words and expressions in its vocabulary. There are several ways for doing this: looking at frequencies for single words, n-grams, collocations and named entities and looking at the concordance of the significant keywords to see how they make sense in the context. To understand the distribution of keywords making the domain of biotechnology we calculated their frequencies and present them as tag clouds. We annotated the corpus with part-of the-speech tags and pre-processed the corpus (removed stop words, stemmed the words and filtered all other words except nouns and adjectives) before proceeding to calculations. We used KNIME text processing modules for major natural language processing (NLP) and text mining texts. We applied a cut-off to exclude word types with little information value such as word types occurring in many or few documents. The cut-off value is determined according to the tf*idf value which is a weighting factor intended to reflect the importance of a word to a document within a corpus. The tf*idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

***Unigram tag clouds***

Media attention between 2000 and 2015 in the biotechnology domain revolves around the tissue, embryo and companies concepts. Human organs such as brain, skin, and marrow are the minor concepts complementing these major concepts. From these findings we can premise that these concepts anchor and objectify abstract biotechnology domain to common sense framing the public understanding of science. Public imagination is mostly concerned about the opportunities and challenges offered by biotechnology for existential problems such as regenerative medicine and aging.

**Figure 9 Unigram tag cloud**




### *Bigrams*

Consecutive word patterns such as phrases provide more information to capture precise or complex meanings. N-grams are tuples of *n* successive words which can be used to get more detailed information about the distribution of a concept attracting media attention. Observing the bi-gram tag cloud confirms our findings in unigram tag clouds. Media attention revolves around phrases such as bone marrow, gene therapy, immune system which are indicative of regenerative medicine. We can also detect human genome research and genetic code as a secondary issue. From this, we can premise that biotechnology represents a long lasting cultural archetype, "decoding the mystery of life" in the public mind.

**Figure 10 Bigram tag cloud**


### *Named entities*

For a finer grained examination of the distribution of what are the semantic elements making up the domain of biotechnology, we checked the named entities. Named entities are the sequences of words in a text such as persons, organizations, locations or specific terminologies such as gene or protein names. We used Apache OpenNLP named entity recognizer which offers a number of pre-trained name finder models to tag the named entities for organizations (https://opennlp.apache.org/documentation/manual/opennlp.html#tools.namefind.recognition.

Examination of the tag cloud shows that media attention on biotechnology revolves around organizations such as health and medicine authorities (i.e. Embryology Authority), science societies and councils (i.e. National Academy of Sciences), universities, government authorities (i.e. NHS) and companies (i.e. Wellcome trust). This observation may point to a controversy about the regulation and deregulation of the biotechnology research. Since biotechnology research involves important issues of human life how and by whom it would be conducted poses an important challenge.

**Figure 11 Organizations tag cloud**


## Concordances

Until now, we have excluded the keywords used for constructing the corpus such as stem cell since they were off the tf*idf cut-off limits (they appeared in almost all documents). Obviously, we did not do this because they were not relevant to our research but because we assumed that the information they contain is self-evident. To analyse what these keywords explain, we conducted a concordance analysis. A concordance analysis lets you search through the corpus for all the instances of a given keyword, and then see the surrounding context for each instance. Concordances can help us to understand the significant keywords in the context. When we check the concordances of the keyword “stem cell” we can observe three important patterns: opportunities offered by stem cell research for regenerative medicine; concerns about hybrid organisms and chimaera; and regulating stem cell research. As for the concordances of the keyword GM; the patterns are: Economic prospects offered by GM crops, controversy around the environmental and health risks of GM and regulation of GM. We can also detect interesting statements such as “GM would ease world hunger”; “environmental groups turn their opposition to a religious crusade”, and regulation of GM is called mostly by EU (table 1).

**Table 1 Concordances for “stem cell” and GM(genetically modified)**

| **Regenerative medicine** | | |
|---|---|---|
| the creation of | stem | cells by therapeutic cloning |
| based on human embryonic | stem | cells . |
| develop new sources of parent | stem | cells , to treat serious diseases |
| therapeutically useful human | stem | cells |
| be cloned for research on the use of | stem | cells for therapies. |
| huge potential of embryonic | stem | cells to alleviate suffering. |

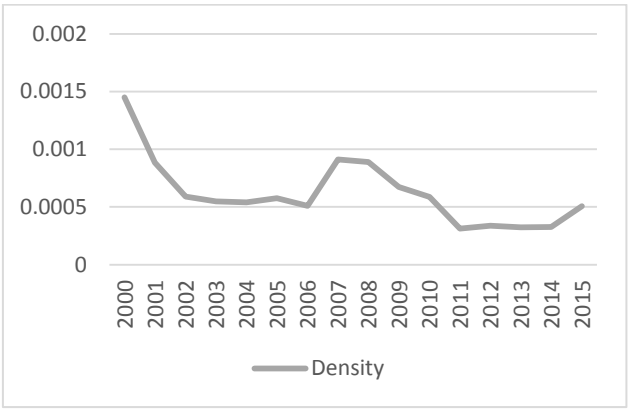
| **Economic prospects** | | |
|---|---|---|
| new technique offers the prospect of | GM | crops tailor-made for specific situations |
| BASF included the nptII gene in their | GM | potato. The gene will help protect |
| the science shows overwhelmingly that | GM | crops will help to ease world hunger |
| before concluding that | GM | crops would not benefit the Third World |
| however , that | GM | technology will lead to more efficien |
| has urged government to embrace | GM | food as the only way to feed the |

| | | | | | |
|---|---|---|---|---|---|
| help us direct the evolution of | stem | cells to repair cellular damage | commercial growing of Amflora , their | GM | potato , early in 2008. |
| **hybrid organisms** | | | **GM controversy** | | |
| | Stem | cells , hybrids , cybrids and chimerea | as they were at the height of the | GM | food controversy of the 1990s. |
| human genes into chicken embryonic | stem | cells , then blended those cells wit | fears of environmental groups ? .The | GM | food controversy |
| Chimeras could be made with embryonic | stem | cells that we have had available for | the environmental impact that their | GM | crops could have. |
| shocking : a hamburger grown from | stem | cells in a laboratory | groups of turning their opposition to | GM | plants into a 'religious crusade ' , |
| **Regulation** | | | studies to push the case against | GM | crops. He also argues that the green |
| all countries to introduce relevant | stem | cell legislation | Feeding or Fooling the World - Can | GM | crops really feed the hungry ? |
| whether cells are to be safe to use. If a | stem | cell runs amok | Lords select committee on | GM | crops , which reported in 1999. |
| legislation to conduct research on | stem | cell isolation and therapeutic cloning | prove the long-term safety of | GM | crops. 'Science has its limits . |
| research permitted by the Act to allow | stem | cell isolation and therapeutic cloning | The consequences of | GM | are more about politics than science |
| | | | to control precise properties of | GM | crop , they have conducted too little research |
| | | | **Regulation** | | |
| | | | imposed a five-year moratorium on | GM | crops - from 1999 until last year |
| | | | organisations calling for a ban on | GM | crops , applies 'a similar lack of |
| | | | The European Commission encourages | GM | crops , which it has declared to be safe |
| | | | Member States have consistently rejected | GM | applications [ but ] they tend to be |

### 3.3.2.2 Ordering, scaling for emphasis: How salient is a topic?

Emphasis measures are indicators of saliency or relevance of a topic or an issue relative to the others. These indicators are useful tools particularly in terms of revealing trends and how the agenda is set for public conception and discourse as they are cultivated by the message systems. They are based on the comparison of such measures as frequency, intensity, or emphasis on the representation of the subject domain in news items across time or some other external variable. An important indicator of emphasis in media monitoring studies is the density indicator. **It refers to the** relative space dedicated to news about a specific topic measured as the percentage of words of within these news over the total number of words in the newspaper. This indicator can help us to highlight moments of “media epidemics” on a specific topic, seasonality related to certain periods (i.e. days or months) and specific events of commemorative dates. Calculating this indicator as an average over intervals, it is possible to evaluate how the salience of a topic changes over time.

**Figure 12 Density of biotechnology topic between 2000-2015**

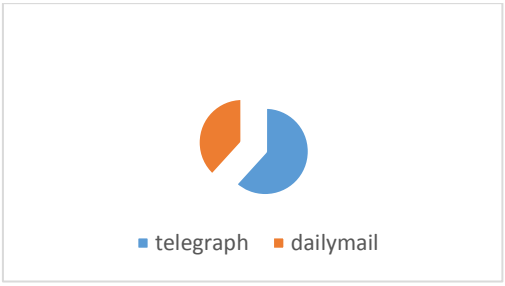



We did not detect any significant seasonality effects relevant to the density of the biotechnology news. The trend in reporting biotechnology issues does not change according to any month of the year. Therefore we do not report the seasonality charts for not overcrowding. As for the media epidemics, we can observe two important peaks in terms of

saliency of biotechnology news between the period 2000 and 2015. First one is a fall from peak occurred in the end of 1990s. This peak represents the media epidemics owing to the excitement caused by the cloning of the first mammal from an adult somatic cell using the process of nuclear cell transfer. After occupying the public imagination and media attention for a while, the saliency of the issue fades gradually reaching to a trough in 2002. Biotechnology becomes salient in the news again during the period between 2007 and 2009. In 2006, researchers at Advanced Cell Technology of Worcester, Massachusetts, succeeded in obtaining stem cells from mouse embryos without destroying the embryos. Another technique announced in 2007 also contributed to the debate and controversy. Research teams in the United States and Japan have developed a simple and cost-effective method of reprogramming human skin cells to function much like embryonic stem cells by introducing artificial viruses. However, these new technologies created controversy regarding the ethics of research involving the development, use, and destruction of human embryos. The controversy reached its peak in 2007 when Bush vetoed the Stem Cell Research Enhancement Bill, which would have amended the Public Health Service Act to provide for human embryonic stem cell research. Another significant event keeping the debate alive was Obama's removal of the restriction on federal funding for newer stem cell lines in 2009. After that date, saliency of biotechnology in the news falls again.

Density indicator can also be used to compare different newspapers' editorial and cultural policy in terms of emphasising a topic. We can observe that Telegraph tends the emphasize biotechnology news much more than the Daily Mail.

**Figure 13 Relative saliency of biotechnology compared to all news**




### 3.3.2.3 Tendency: From what point of view are things presented?

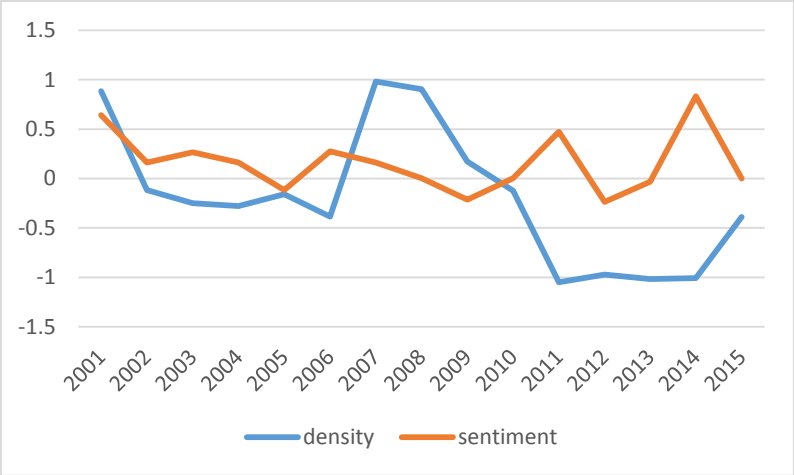
Measures of tendency are about the directionality of the evaluation of the things, issues, events and people as represented in a given piece of text, such as a document, paragraph, or sentence. They are about the explicit or contextual judgment of qualities of phenomena expressed in the media presentation and based on identifying if a text about an issue appears in a supportive or critical context. The broadest overall dimension of judgment is a summary evaluation of the goodness or badness, rightness or wrongness of things (Gerbner, 1969). Tendency indicators address the aspect of subjective opinion such as identifying different points of view, identifying different emotive dimensions, and classifying text by opinion. Various computational linguistics models and methods are developed for calculating the

degree of subjectivity in the text. Gathered under the rubrics of "sentiment analysis" or "opinion mining", these methods aim to identify positive and negative opinions, emotions, and evaluations; to distinguish attitudes from factual statements; to detect if the author reports his/her point of view or someone else's; or to distinguish between explicitly and implicitly stated attitudes (Shanahan et. al, 2006).

There are two basic approaches for scoring a piece of text according to the directionality of the sentiment as expressed in its content: The first one considers sentiment analysis as a text classification task and employs supervised machine learning algorithms to identify the terms demarcating between the different poles or levels of sentiment. After training the algorithm, it builds a classifier containing word vectors for scoring the degree of subjectivity in a text unit. These algorithms require manually tagged set of texts as a training set which is a quite labour and time extensive process. The second one manages the task first building a sentiment lexicon containing a list of keywords denoting positive or negative sentiments, then verifies the degree of subjectivity of a text unit through some function based on the positive and negative categories as defined by the lexicon. While there is no need for tagged data for this method, it requires robust linguistic resources (a meticulously prepared and well established dictionary) which are rarely available. Moreover, such dictionaries are criticized on the grounds that they may not function accurately as the context they were designed for might be very different than the other contexts. As both approaches have advantages and disadvantages, we opted for the second one because of resource limitations.

Lexicons are widely used in sentiment analysis and opinion mining and there is a variety of them prepared for different purposes. For the purposes of this study, we used the subjectivity lexicon developed by the Multi-perspective Question Answering (MPQA) Opinion Corpus project. The MPQA Opinion Corpus contains news articles from a wide variety of news sources manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.). (http://mpqa.cs.pitt.edu/). Subjectivity lexicon is constructed employing manual and automatic identification of keywords in negative and positive pole categories. It covers keywords and information about their polarities, degrees of subjectivity, and their part-of-speech tags. It scores the direction and degree of the subjective expressions in documents by means of detecting the existence of the keywords in text units. A subjective expression is any word or phrase used to express an opinion, emotion, evaluation, stance, speculation, etc. MPQA subjectivity lexicon focuses on *sentiment expressions* – positive and negative expressions of emotions, evaluations, and stances. The algorithm scores the text unit as -1 if a negative keyword is detected and as +1 if a positive keyword is detected. The sum of scores in a document gives the overall sentiment.

**Figure 14: The sentiment changes in cycles oscillating (2001-2005)**




We can observe an interesting trend in sentiment about biotechnology. The sentiment changes in cycles oscillating between higher positive to lower positive between 2001 and 2015. What is more interesting is this oscillation follows a counter trend with the highs and lows of media attention. This observation brings to mind if we can hypothesize sentiment about biotechnology follows a kind of "fashion cycle". Fashion cycles take off with an enthusiasm about an emerging style usually after an endorsement by a celebrity. Emulation, the second phase, begins when the style becomes a hype because of the intense media attention. Saturation, the last phase starts when the style becomes a cliché, a buzzword that everyone can afford. A quick qualitative reading of the articles revealed that media attention starts with an optimistic debate about the potentials of a new biotechnological innovation and when the attention reaches its peak, sentiment falls with news concentrating on criticisms and risks. This observation is obviously a tentative inference which needs to be investigated with more systematic and meticulous analytical strategies for future studies.

### 3.3.2.4 Structure of the cultivation of collective notions.

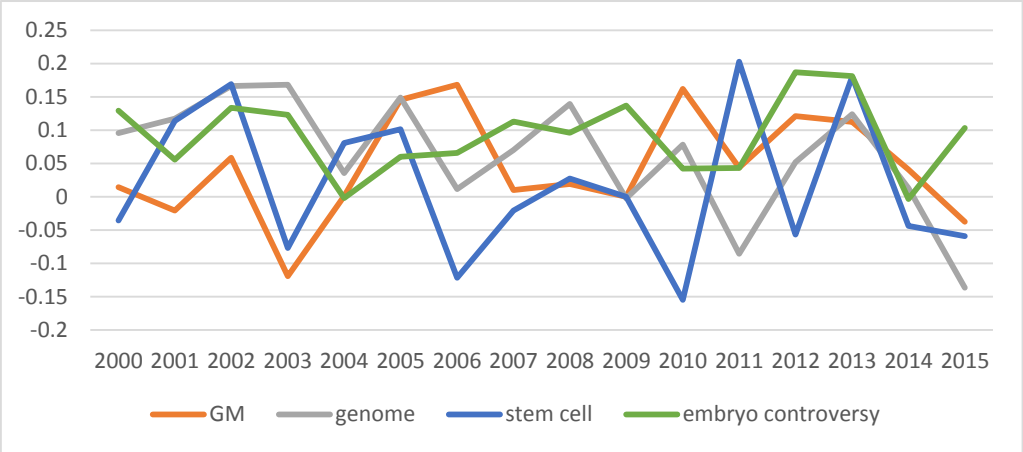
Coverage, accounts, explanations residing in singular news items deal with particular life events or concepts extracted from total situations. Nonfictional mode of presentation is primarily analytical in the sense that it focuses its attention mainly to one major theme which makes a part of a knowledge domain organized as classes of topics. While other three indicators are concerned about the composition of the individual elements making a message system, their distributions, and their directionality, structure indicators are concerned about how these elements come together to reflect the constitution of a whole system. These elements can come together in terms of proximity, causality or some other logical relation to constitute the overall structure. In this paper, we will mainly concentrate on the mechanisms behind proximal structuring or clustering since they are based on more established methods and the latter require higher level computational linguistics algorithms that are still under development. The mechanism behind proximity based methods are based on the assumption that semantic proximity arises if two words co-occur with similar words within a context. For example, if the words cat, dog, hamster, parrot always co-occur with words such as pet shop, toys, food, then they belong to the same category (i.e. pet). If these words frequently co-occur in predetermined contexts such as documents, paragraphs or sentences, it is then possible to

calculate a proximity measure such as the correlation between the common co-occurrence of these words within these units. Proximity matrix obtained as a result of these operations then enters as an input to an automatic pattern detection algorithm that performs the task of grouping a set of words in such a way that words in the same group are more proximate to each other than to those in other groups.

As the digital revolution created an increasing need for information retrieval, information filtering and document organization, a variety of topic modelling methods based on automatic identification of semantic content are becoming available. Topic modelling refers to a suite of algorithms for discovering latent thematic structure in corpora containing large amounts of documents. These algorithms identify latent patterns pervading a corpus by grouping words into 'topics' and annotate the documents according to those topics. Among these methods, Latent Dirichlet Allocation (LDA) is a well-established and highly effective unsupervised learning technique that model each document as a combination of topics that generate words with certain probabilities. The document-topic and topic-word distributions learned by LDA describe the best topics for documents and the most descriptive words for each topic (Blei et. al., 2003).

We recovered six topics represented by ten words each: **regenerative medicine** (*stem, cell, research, patient, human, body, heart, blood, tissue, disease*); **gene therapy** (*gene, therapy, genetic, disease, immune, patient, treatment, virus,* ); **stem cell controversy** (*stem, cell, research, scientist, human, embryo, egg, genetic disease, treatment*); **human genome research** (*human, genome, research, scientist, gene, sequence, genetic, DNA, code, life*); **commercial biotechnology** (*drug, companies, market, fund, biotechnology, science, research, medical, treatment, world*) and **GM controversy** (*GM, food, crops, plant, agriculture, product, research, technology, animals, gene*). We summarized the content of the topics as follows after a deep reading of sample news with highest scores for each topic: The sample news in topic 1 mostly mention about issues such as regenerative medicine, spare organs, organ reconstruction and neurogenesis. Topic 2 is about gene therapy, stem cell transplantation and xenotransplantation. Topic 3 is mentions mostly parthenogenesis, embryonic stem cell research, cloning human embryos, ethics, moral concerns, banning reproductive cloning, baby cloning, and IVF. Topic 4 is about issues such as artificial life forms, decoding genes, genetic engineering, transforming of human existence, tracing the origin of species, designer species, designer babies and hybrid creatures. Topic 5 is about issues such as biotech companies, market, biotech products, share prices, drugs, and medicine. Topic 6 is covers issues such as Frankenfood, GM crops, food, GMO, GM controversy, corporates, ecology, genetic engineering and biotech companies

**Figure 15 Change in intensity of biotechnology news (2000-2015)**




Because of the space limitations, we are not going into a detailed explanations of each topic, but we can say they follow an oscillating trend as we have observed for the biotechnology topic in general.

## 5 Conclusions

Recent developments in natural language processing, computational linguistics and text mining technologies offer great potentials for online media monitoring. These potentials can help to overcome the shortcomings of classical automated content analysis techniques which usually follow a naïve keyword approach to category detection. This means increasing validity in the categorization of text units. However, online media monitoring are far from perfect and still face some challenges. First, should the researchers follow a deductive or and inductive approach to define the categories in their domain of interest? Should we proceed with etic categories built upon a theory, assumptions and method before starting to collect and analyse the data (King et. al, 1994)? Or should the categories emerge after the data collection and through analysis in an emic way (Saldana, 2009; Herring & Hunsinger, 2009)? While the former keeps the deductive paradigm of classical content analysis, the latter converges the qualitative text analysis paradigm with text mining paradigm to an inductive logic. Second, while the second one brings some interpretive intelligence into text analysis by means of pattern detection and interpreting them in the context, they raise some validity concerns. Deductive categorization has well established reliability and validity instruments which is usually neglected by the inductive researchers. This raises concerns about the validity of categories obtained through inductive automatic pattern detection techniques. We hardly see any validation exercise for the categories obtained by the popular pattern detection methods such as cluster analysis or topic modelling. As deductive validation instruments would not be applicable, the researchers using these techniques might want to consider qualitative criteria such as triangulation, credibility, transferability for evaluating the soundness of their categories. This is usually neglected by the text miners. Third, as most of the time the boundaries of the subject domain for textual data is unknown, fuzzy and dynamically changing it will be very difficult to select the units of analysis. We have pointed to the shortcomings of using web as a corpus since the documents collected by the search engines

(Google) and online companies (Twitter) are given to the researcher rather than being the result of a meticulous sampling design. In this paper, we tried to address some of these challenges and offer some guidelines for using these techniques for developing cultural indicators by means of online media monitoring. Our aim is not to suggest definitive canons for such a task, but is to point to the opportunities and challenges offered by the emerging field of online media monitoring for developing cultural indicators. Present online media monitoring tools are usually used for business intelligence gathering purposes and do not meet the requirements for a system designed for public surveillance and research service accessible by all citizens and researchers. This paper tried to address this gap by delineating the guidelines for constructing such a system allowing to monitor issues concerning public engagement to science and responsible research and innovation.

## 6 References


Atkins S., Clear J., Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing* 7(1): 1–16.

Baker P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum

Bauer M.W. (1998). The medicalization of science news: from the 'rocket-scalpel' to the 'gene-meteorite' complex, *Social Science Information*, 37, 731-751.

Bauer M.W., Durant J., Ragnarsdottir A., Rudolfsdottir A. (1995). Science and technology in the British press, 1946-1992, The Media Monitor Project, Vol 1-4, *Technical Report*, London, Science Museum and Welcome Trust for the History of Medicine.

Bauer M.W., Falade F.,Suerdem A, (2016). Science News in the British Press, 1990-2013 MACAS module 2 Mass Media Mapping, - *Rolling Progress Report*, London, LSE.

Bauer M.W., Gaskell G. (2008). Social representations theory: a progressive research programme for Social Psychology, *Journal for the Theory of Social Behaviour*, 38, 4, 335-354.

Bauer M.W., Petkova K., Boyadjieva P., Gornev G. (2006). Long-term trends in the representations of science across the iron curtain: Britain and Bulgaria, 1946-95, *Social Studies of Science*, 36, 1, 97-129

Bauer M.W., Suerdem A., Bicquelet A. (2014). Text Analysis – an Introductory Manifesto, in: Bauer, MW, Suerdem A., Bicquelet A. (eds) Textual Analysis – *Sage Benchmarks in Social Research Methods*, London, SAGE, Vol 1, pp xxi-xlvii (of 4 volumes) [ISBN 978-1-4462-4689-4]

Bauer, M. W. (2005). Distinguishing Red and Green Biotechnology: Cultivation Effects of the Elite Press. *International Journal of Public Opinion Research*, 17(1), 63-89.

Bauer, M. W. (2005). Public Perceptions and Mass Media in the Biotechnology Controversy. *International Journal of Public Opinion Research*, 17(1), 5-22.

Bauer, M. W. (2007). The Public Career of the 'gene'—trends in Public Sentiments from 1946 to 2002.*" New Genetics and Society* 26.1, 29-45.

Bauer, M; Aarts, B; (2000). Corpus construction: a principle for qualitative data collection. In: Bauer, M and Gaskell, G, (eds.) *Qualitative researching: with text, image and sound, London: Sage*

Biber D., Connor U., & Upton T. A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins Pub, 209

Blei D.M., Andrew, A.Y., Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 993–1022

Dodge Y. (2003).The International Statistical Institute, "The Oxford Dictionary of Statistical Terms", *Oxford University Press*

Falade B., Bauer M.W., P. Pansegrau et al. (2017). Comparing science in the news across countries: British, German and Indian press, in**:** Bauer M.W, Pansegrau P.& Shukla R.(eds) Mapping the Cultural Authority of Science, London, *Routledge Studies on Science, Technology and Society.*
Gerbner G. (1969). Toward Cultural Indicators: The analysis of mass mediated public message systems, *AV communication review* 17(2), 137-148

Gruber T. (2008). "*Ontology"*. Liu, Ling; Özsu, M. Tamer, eds. *Encyclopedia of Database Systems.* Springer-Verlag.

Hansen A., & Dickinson R. (1992). Science coverage in the British mass media: Media output and source input. Communications: *The European Journal of Communication*, 17(3), 365-377

Herring, S. (2009). Hunsinger, Jeremy, ed. *Web Content Analysis: Expanding the Paradigm*. Springer Netherlands. pp. 233–249

Ide, N., Reppen, R., Suderman, K. (2002). *The American National Corpus: More Than the Web Can Provide. Proceedings of the Third Language Resources and Evaluation Conference* (LREC), Las Palmas, Canary Islands, Spain, 839-44

Illia, L., Sonpar, K., & Bauer, M. W. (2012). Applying Co-occurrence Text Analysis with ALCESTE to Studies of Impression Management. *Brit J Manage British Journal of Management*, 25(2), 352-372

King, G, Robert O. Keohane, & Sidney Verba. (1994). Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton: Prince University Press.

Krippendorf K. (2012). Content Analysis; An Introduction to its Methodology, (3rd ad.) Thousand Oaks, CA: Sage. *Organizational Research Methods*, 441

Markoff J., Shapiro G., Weitman S.R, (1974). Toward the Integration of Content Analysis and General Methodology. *Sociological Methodology 1975*. San Francisco: Jossey-Bass.

McEnery T., Hardie A. (2012). Corpus Linguistics: Method, theory and practice. *Cambridge: Cambridge University Press*.

Moscovici S. (2000). *Social Representations. Explorations in Social Psychology*. Cambridge, UK: Polity Press

Ranganathan, S.R. (1967). *Prolegomena to Library Classification*. Asian Publishing House, Bombay, India.

Saldana J. (2009). The Coding Manual for Qualitative Research. London: SAGE Publication Ltd.

Schmied J. (1996). Second Language Corpora, in S. Greenbaum (ed.) *Comparing English worldwide: the international corpus of English, Oxford: Clarendon*

Shanahan J. G., Yan Q, and Wiebe J., (2006). Introduction, in James G. Shanahan, Yan Qu, and Janyce Wiebe (eds), *Computing Attitude and Affect in Text: Theory and Applications Dordrecht: Springer.*