“ MAPPING THE CULTURAL AUTHORITY OF SCIENCE
across Europe and India ”

MACAS - Mapping the Cultural Authority of Science

Module 2 – Media Mapping

MACAS-British Council Joint Workshop

Istanbul 10-11 January 2014

Pushing the agenda for an International platform for standardizing and harmonizing Indicators of Cultural Authority of Science indicators

Short Workshop report

The workshop lasted two full days and took place at Bilgi University, Istanbul with 15 participants from Turkey, UK, Germany, Italy and Spain.

Day1: Friday 10th January 2014

Ahmet Suerdem (Istanbul) welcomed colleagues to Bilgi University and wishes the workshop a good success.

Martin Bauer (LSE) introduced the purpose and structure of the MACAS project, its history, context and the three modules. The present workshop will focus exclusively on module 2: the construction indicators based on mass media coverage of science in the comparative context of European countries and across the India. Chinese colleagues are paying close attention to the development of our procedures with a view of contributing a similar module. Bauer present the progress of the module arising from the July meeting in London, and defines the objectives of this workshop:

1. To discuss the attention indicators arising from the current state of the corpora
2. To advance the definition of linguistic indicators and the comparison with natural language corpora
3. To discuss and stabilize coding frame for comparative analysis

Petra Pansegrau presented briefly the MACAS website, which is now live for several month and documents in a simple and clear manner the progress of the project. Colleagues are invited to contribute their short CVs and materials arising from this workshop.

Ahmet Suerdem and colleagues (Istanbul) introduced the newly funded 'Tubitak project', which aims to build a platform for monitoring science news in Turkey, based on a crawler and thesaurus logic that is able to collect and classify 'science news' items on a continuous basis.

Unfortunately the group could not further discuss the interesting input with regard to efficiency of different text classification methods depending on their 'similarity index'. Important exams called the speaker away too early. But the issue was successfully raised: how to partition large matrices of n-grams x documents to classify the documents by clustering in both dimensions?

Migues Angel Quintanilla, Carlos Figuerola and Tamar Groves (Salamanca) present their own project on selecting science news in Spain with a thesaurus, and to code the content for intrinsic and extrinsic features of science, again thesaurus based classifications for Spain 2000-2010. They estimate the intensity of news coverage at about 6% of the total newshole, with annual fluctuations indicating particular events. They develop an index of 'intrinsicality', of relation of intrinsic versus extrinsic focus in the corpus of science news. This index can be compared across time, newspaper and news topic, science or technology. Intrinsciality is much higher in science news than in technology news, and intrinsicality in in science news seems to be increasing.

Discussion: The Spanish colleagues respond to the challenge of 'validity' of their results and present the problems they have with stabilising the corpus, duplications of articles due to multiple version, unusual high growth of article in one particular newspaper.

Petra Pansegrau and Joerg Roesing, the German MACAS-team, presented their progress on constructing the corpus. Different from the other groups they are building up a corpus based on **whole coverage** about science and scientists in 'Der Spiegel' (the leading German news magazine) supplemented by some years of 'Die Sueddeutsche Zeitung' (daily newspaper). The corpus includes 25 000 articles so far and will be ready to analyze soon.

Frederico Neresini and Andrea Lorenzet from Padua (Italy) presented their corpus (1990-2013, Il Corriere della Sella & La Republica) for media analysis and showed first results. The findings focused the level of attention for science in press as well as the detection of issues and themes.

Bankole Falade (LSE) presented progress in the construction of the UK science corpus for 1990 to 2012. The corpus is based TIMES and MIRROR on a bi-annual sample of 14 days each year (artificial week) and a set of keywords. Stabilizing and comparison of different sets of keywords is the problem. The search facilities of NEXIS was explored, efficient ways and problems of retrieval examined (duplicates and multiple editions), and the test corpus was examined on a 'relevance rating' of 0 to 5. The keyword set still produces too much irrelevant material.

Discussion: The presentations were followed by a discussion on the questions how the different corpora can be made comparable and how the potential comparability can utilized in an appropriate way.

Furthermore the decision was made to define for all corpora the 'level of relevance' like presented from the LSE-team. They defined six values (grades) that show how 'relevant' each article is in relation to science. MACAS will operationalise this relevance a) on the number of keywords one article includes (for example: 1-4 keyword ->little relevance, 5-8 keywords science used as a metaphor …. > 15 keywords -> full science article), or/and b) on a simple rating scale.

Guergana Popova (Goldsmith College) reacted directly to the UK presentation, having inspected examples of the UK corpus in preparation of the workshop. She presented a panorama of linguistic observations on the current state of the corpus including relevance of the material, syntax features, and the representations of science.

The following discussion revealed that a micro structural linguistic analysis, such as Guergana Popova pointed at a few examples, is very insightful and should therefore be carried out for the case studies of module 3. The idea is to show that language of science reporting forms specific features and expressions that provide further evidence on the extent to which culture is shaped by scientific knowledge or will be.

Amanda Potts (Lancaster)

Natural language corpora and how to access them at CASS: a demonstration

Discussion: Because of the long meeting day and the pre-ordered supper the discussion of the Unit was postponed to the following day and the working group II. The immediate questions focused in particular the comparability of Lancaster Corpus with other national corpora (like the Institut fuer

Deutsche Sprache, Mannheim), and the question how MACAS can get access to the different corpora of Lancaster. Amanda Potts assured that she will be helpful both in the question of access as well as in the preparation of the analysis. In addition, she presented materials for the workshop II available to all participants.

Day 2: Saturday 11 Jan 2014

Roberto Franzosi (Emery & Oxford) presents the opening lecture. In this paper he argues that innovation, when based on ignorance of historical antecedents, often amounts to the reinvention of the wheel less effective and less efficient. He makes his point on the analytical category 'framing', which is now trade in stock in the analysis of public controversies of science and technologies. Comparing frame analysis with the traditions of rhetorical analysis of three base arguments (logos, ethos, pathos), genre, topos and tropes exceeds by far the scope and power the 'innovation' called frame analysis. He reports how this argument arises from his long preoccupation with 'quantitative narrative analysis', which made him realise these short-comings in relation to the tradition of analysing narrative structures. He finishes with a short demonstration, unrelated to his main point, of the power of data visualisation, arising from narrative analysis when using dynamic network animations and GIS for the presentation of relational data over time and location.

Discussion: Andrea Lorenzet observes that tradition and history can be overwhelming, as often in the case of Italy. New terminology is often needed to liberate and to overcome unnecessary blockage of innovation. Amanda Potts observes that confusion and variety of analytical terms often arises from 'intellectual territory' that defines academic identities. Also methodology is a market that encourages spurious product differentiation as 'selling propositions'. It is important to distinguish will-full ignorance from de-facto ignorance of the tradition. Several colleagues wondered why rhetoric got its bad reputation and disappeared from the radar of general education, only to find itself reinvented?

For the rest of the day, we formed two groups focussing on particular tasks.

Group A: Issue arising from using QDA miner for text classification led by Ahmet Suerdem

The group worked with parts of German corpus of 'Der Spiegel'. Ahmed Suerdem and Ceyhun Aytekin showed a few automatic operations that make the analyses comparable and manageable. There is a agreement that all corpora should be prepared due to the item quantity by the QDA Miner to automatically process and analyze: Therefore some variables are needed, which can be created manually, for example, the date of publication of one article. However, the manual way would be very time-consuming for each corpus. That is why the group needs an algorithm according to which the QDA Miner the information in the article (1) detects, (2) extracted and (3) transfoms into a variable of appropriate definition. It already showed the possibility of using certain limiter to encode the appearance of data in the article automatically. The problem was, however, to convert the encoded locations in appropriate variables.

Ahmet Suerdem and Ceyhun Aytekin suggested the following solution: it is assumed that the individual journal articles has ever the same text structure, e.g. is the release date in every article on the same line. To determine the variables of QDA Miner suitable, it is advisable to register from the texts in a table information by a (program) script. By using this table QDA Miner can create the variables automatically. Ceyhun Aytekin offered to work out this script.

Group B on corpus linguistic tools offered on-line at CASS (Lancaster) led by Amanda Potts.

The group moved into a Bilgi computer lab, and Amanda shows the way to the CASS resources on-line. We all opened our computer on the corpus facility, and went through an exercise of querying the Brown Corpus Family on issues of grammar change, e.g. 'verb contractions', semantic preference using the inbuilt semantic tagging USAS, and keyword comparisons with a reference corpus. This included an introduction into the CASS command language and syntax, and discussion of the mark-up of the text sources, and the different query and parameter windows of the interface.

Discussion focussed on how to provide and construct meta-data for our own corpora, which then can be up-loaded onto CASS and compared to the various reference corpora that are available. The Brown Family corpus allows comparison to British English for 1991 and 2006 across a number of text genres including press.

Preparation for corpus comparisons will include the following steps

- Preparation of news items as text files
- Each file headed by meta-data in XML; guidance is available from Andrew Harvey;
- Each new item + metadata defines one file with filename
- An EXCELL file with all metadata + the filename of the items
- The EXELL file + files can then be uploaded onto a personal account at CASS, free of charge
- The uploaded file can then be automatically tagged for syntax features and semantic fields using the USAS system using Xmatrix3 for an annual change of £50
- Once the corpus is uploaded and tagged, it can then be queried and compared to reference corpora in the CASS system; and the metadata allows to partialize the comparisons according to the variables offered.

For the British corpus, these steps should be achievable during 2014, and Amanda Potts will be available to help and assist as needed, even to join us in London for this purpose. We will form a task force (Bauer, Falade, Popova, Potts) to work on this problem during 2014.

The objective is to make the UK news corpus CASS ready later by autumn 2014.

Summary session

Martin Bauer summarized the main insights of the workshop and draws the implications for future work on MACAS module 2.

<u>Attention and news itensity</u>

Stabilising the corpus has immediate priority until spring 2014 in all three contexts present: UK, Germany and Italy. This includes a) making transparent the keywords used to select materials, and b) coding each items on 'relevance' (0 = irrelevant, 4 = science focussed news item). Relevance should be a selection (inclusion) criterion as well as a internal filter. The internal filter, on already selected materials will allow us to conduct sensitivity analysis on any constructed indicator such as intensity or topic and text method variability.

Linguistic markers

Developing linguistic indicators for the comparison of science news with general language usage. Such indicators should reflect the relative distance of science news from everyday language, and be able to show convergence or divergence. Identifying such indicators will need further consultation of the literature and qualitative inspection of the corpus (Popova). Identifing these indicators will proceed in parallel with build-up of the UK science news corpus 1990-2012 to CASS compatability (UK CASS taskforce).

Comparative Coding frame

The Stabilising the comparative coding frame will continue step-wise and continuous through 2014, building in each team on the preliminary ideas and staying in close contact on any new developments. The coding frame includes the following sections to be developed.

Meta-data:	day, date, year, genre, version. It was considered important to record the 'version' of articles that have several versions in the same source, for example the TIMES in the London, Scotish or Northern Irish edition. There was some disagreement whether only one version should be counted, or several versions retained. It was decided, that for purposes of 'discourse analysis', all versions are relevant, but for purposes of 'intensity and attention' only one version must count in order to avoid inflationary figures.
Actor/action:	list of actors and action which allows to create actor/action links
Dynamics of controversy	
Valuation	n: two rating scales, positivity and negativity; Score E = (w)N – (w)P.
Scientific discipline	e.g. following Encyclopedia Britannica as used 1946-1995
Themes	the big science topics of the past 20 years: nuclear power, environment/climate change, biotechnology,

	nanotechnology, cloning/reproduction, brain science, AIDS, Space/asteroids etc.
Issues, policy issues	energy, economic growth, education, family, religion
Framing	representations of science: see list developed by Popova, also the Biotechnology frames

Other outcomes

Powerpoint presentations and related materials of the proceedings will be available on the website

The Spanish team from Salamanca has asked to be formally become an affiliated partner in the MACAS project. There should be nothing in the way to grant this request. The objectives of Salamanca and MACAS are entirely compatible in the objective of thinking through and developing 'cultural indicators of science'. Their thesaurus based expertise is welcome to enhance similar efforts in MACAS (Tubitak/Italy). The idea of defining 'intrinsic' and 'extrinsic' features (Quintanilla) of science and technology reportage brings welcome conceptual breath to MACAS.

MACAS has as of January 2014 a new fully affiliated partner team in Salamanca, Spain with focus on module 2.

Outlook for PCST Bahia 5-8 May 2014; followed by MACAS workshop on Friday 9th May at Bahia University with partners from Sao Paulo.

Bielefeld workshop in June 2014 with Indian colleagues in India (Subhasis Sahoo)

Publication ideas:

Special issue of PUS on the construction of science news corpora for 2015

A book publication arising from the MACAS project for 2016, finalised in a conference at ZIF in Bielefeld in September 2015; with contributions from outside the project

Martin W Bauer & Petra Pansegrau

m.bauer@lse.ac.uk

January 2014