**Preliminary communication**

Alessandra Vicentini, Kim Grego*, Daniele Russo

# The Pro.Bio.Dic. *(Prototype of a Bioethics Dictionary)* project: Building a corpus of popular and specialized bioethics texts

ABSTRACT

This paper reports on an ongoing, long-term research project in the field of medical ethics and bioethics conducted by a multidisciplinary team combining medical, linguistic, IT and philosophical research interests: the *Prototype of a Bioethics Dictionary* (*Pro.bio.dic*).
Having already outlined (Vicentini *et al.* 2011) the reasons and needs to both redefine and update the lexicographic material available so as to provide a corpus-based collection of the English terms of contemporary bioethics to be published on a web platform, the *Pro.bio.dic* has now entered the key stage of corpus-building.
This stage requires establishing the criteria involved in creating a large, statistically-valid reference corpus of both specialized and popular bioethics texts, to be processed by means of text-mining and machine-learning techniques, and to serve as the basis from which the entries of the electronic online tool described as the *Pro.bio.dic* will be drawn by means of concordancing software.

**Keywords**: bioethics, English lexicography, corpora, corpus linguistics, online dictionary

[1] Research for this paper has been carried out jointly by the three authors. Alessandra Vicentini is responsible for § 1 and 3; Kim Grego for § 2 and 2.1; Daniele Russo for § 2.2.

* Correspondence address: Kim Grego, University of Milan Department of Language Studies and Compared Literature, Milan, Italy, e-mail: kim.grego@unimi.it

## 1. Background, aims and methods

This paper stems from an ongoing, long-term research project in the field of medical ethics and bioethics conducted by a multidisciplinary team[2] combining medical, linguistic, computer science and philosophical research interests: the *Prototype of a Bioethics Dictionary* (Pro.Bio.Dic). Its origins lie in the realisation that novel issues regarding the ethics of medical communication in the globalized, internet-connected world (Brügger / Bødker 2002, Slevin 2002, Mooney / Sarangi 2003, Vicentini / Grego *et al*. 2010) – together with the modern advances in technology and informatics applied to linguistics (Sinclair 1991, Thomas / Short 1996, Joachims 2002, Avancini *et al*. 2006, Bishop 2006) – call for both a redefinition and an updating of the lexicographic material available. Indeed, a detailed scrutiny of the latter showed it to be inadequate to meet today's societal and knowledge requirements in the domain of bioethics (Jonas 1997, Branningan 2001, 2004, Bellini 2008, 2012), since it was addressed only to a specialist public, written by a single (either a physician or a philosopher) expert, and created without referring to a scientific compiling methodology, but based on previous lexicographic works. Not only, the latter was usually merely re-elaborated, adding and adjusting some new information to the same core lexemes and contents, a process that inevitably left room for the compiler's own introspection and individual linguistic experience[3]. Contrary to this, the Pro.Bio.Dic will distinguish itself from the past lexicographic production in terms of a) the multidisciplinary approach adopted, necessarily deriving from the underlying linguistic objectives and from the inherently interdisciplinary nature of bioethics, which many scholars would, on that account, describe as a field rather than a specific discipline; b) the employment of an up-to-date and innovative scientific method based on the principles of statistics, corpus and computational linguistics and automatic classification of texts rooted in machine learning and text mining techniques (Sinclair 1989/1995, Sebastiani 2002, 2006, Liu *et al*. 2007, Manning / Raghavan / Schuetze 2008); c) a new modality of content sharing given by its publication on an online wiki platform, which will allow to involve other experts possibly contributing new insights into the research itself and making it an ever-changing and -improving instrument open to non-experts as well.

[2] The multidisciplinary team working on the project includes researchers, professors, research fellows and Ph.D. students based at the Universities of Varese, Milan, Turin, Pavia (Italy) – comprising a linguistics, a philosophy and a computer science section – and at the Institute for the Science and Technologies of Information of the National Council of Research (ISTI-CNR) (Pisa, Italy). It also avails itself of the supervision and consultancy of The College of Saint Rose (Albany, New York, US) in the person of Prof. M. C. Brannigan.

[3] The problem of bias and ideology connected with lexicographic practice is one of the subjects investigated by the 2009 PRIN project (a government-funded research programme of national interest) *Within and across borders: Usage and norm in Western European languages* coordinated by Prof. G. Iamartino (University of Milan, Italy).

The latter, that is directly opening the dictionary to the general public, constitutes one of the most innovative aspects in comparison with the past approach. One must consider that health and well-being are commonly associated with treatments, cures and procedures. However, especially in the case of well-being, it is possible to contribute to these notions by leading a health-conscious life, which includes receiving (and providing) the correct amount of correct information. The question of how the man-in-the-street can disentangle the huge amount of web-based documents on health(care) available in the internet becomes dramatically relevant when bioethical factors are involved. Indeed, any one citizen may find it necessary to ask at one point or another in his / her life questions concerning life-health(care)-death issues, such as euthanasia, abortion, stem cells, cloning, genetic manipulation etc., by which he / she is constantly and massively bombarded by the media. The Pro.Bio.Dic can provide an answer to all of the above and represent an authoritative and serviceable tool for society as a whole. Indeed, to sum up all the characteristics of the prototype, it is planned to be quality-based (professionally designed and compiled) but quantitatively available (web-based and publicly accessible); it will draw its entries from realia (specialised and non-specialised web-based texts), and thus be as close as possible to real societal needs; it will be regularly maintained up-to-date, edited and integrated; availing itself of the constant collaboration of a multidisciplinary supervising scientific committee, it will deal professionally, informedly, yet correctly from both a political and a religious view, with highly debatable and debated subjects; it will be in line with the legal framework supporting and the ethics informing the EU and use state-of-the-art web- and corpus-based, machine learning IT methodologies. In brief, it is believed it will fill an empty space at the national, European and international level, especially considering that the pilot model is designed to be in English, but a following step can involve turning it into a multi-lingual tool, starting from the EU official working languages (i.e. English and French).

Having already explained in detail the Pro.Bio.Dic's multidisciplinary research frameworks, target and methodologies in previous research (Vicentini / Grego *et al.* 2012), this paper will now report on the key stage of corpus-building. It will describe the criteria and processes involved in creating two large, statistically-valid reference corpora of bioethics texts – a specialised and a popular corpus, both necessary for the double-target and use envisaged – to be processed by means of the computer science techniques hinted at earlier, and to serve as the basis from which the entries of the electronic online tool will be drawn by means of concordancing software.

These two different corpora are needed to build the IT learning models necessary to extract the lemmas considered for inclusion in the dictionary. In particular, it will be necessary to assemble corpora in which the same terms used in bioethics (e.g. abor-

tion, stem cells, etc.) are found in non-bioethical contexts, so that the automatic classification model may learn and recognise semantically, but also lexically and grammatically, what is bioethical and what is not. As regards the specialised corpus, medicine texts dealing with concepts in a technical, professional way, without any primary bioethical interests will be used (see § 2.2). The same will be done with the popular corpus (see § 2.1), using newspapers, or sections thereof, dealing with topics distant from bioethics, but containing the same terms isolated in dissemination bioethical contexts. Once the lemmas have been extracted from the combined specialised and non-specialised corpora, both tested against corpora containing non-bioethical data, the computer science team will proceed with the phase of lemma extraction.

## 2. Corpus selection criteria: an overview

The label 'popular corpus' (hereinafter PC) will indicate here a collection of texts (articles, to be specific) taken from popular sources such as newspapers. 'Popular' will have to be understood in terms of English for Specific Purposes (ESP) vertical variation, as in Cloître and Shinn (1985), i.e. as aimed at a wide general, non-specialised audience. Opposed to that, 'specialised corpus' (hereinafter SC) will indicate the domain-specific collection of inter- and intra-specialistic texts on bioethics from specialised publications in the field. The general selection criteria identified for both corpora are as follows.

As regards the diachronic variation, both the PC and the SC will have to share the same time span. This will make them comparable and the extraction of information chronologically aligned. For instance, a ten-year period might be covered by both corpora, reflecting in their content what the situation was during those years both at the specialised (inter- and intra-specialistic) level and at the popular level, i.e. in the press. One sub-factor to consider will have to be the frequency of updating of the corpora. Once the Pro.Bio.Dic has gone online and started to be contributed to, how often will the definitions have to be updated and, therefore, how often will the corpora from which they are extracted have to be updated or, better, integrated with new content? This is going to depend basically on further practical factors such as the number of scholars and staff involved in the project and the funding available to carry on work on it. An annual or biannual updating frequency is recommended to keep the product viable and serving its underlying popularising purpose.

The diatopic variation will be represented, at this initial stage, by the inclusion of two main varieties of English: British English (BrE) and standard American English (AmE). From a quantity-based viewpoint – especially as regards the number of publications in science in general and in bioethics in particular – these varieties alone may well be considered representative of the English language as a whole. Of course, it

would only be sensible to evaluate the insertion of other major varieties such as Australian, Canadian, South African, Indian English, etc. However, two reflections must be made in regard to the diatopic dimension. The first is that, although a publication may be identified as belonging to one specific variety depending on its place of publication, the contents submitted to it cannot be guaranteed to belong to the same variety. This applies especially to the case of journals but also of newspapers and magazines, given the present globalized times. Only a deep, individual scrutiny of every single text in a publication could reveal the variety it employs, and it would still have to be confirmed by contacting its author(s). Indeed feasible, this would nonetheless prove an exquisitely sociolinguistic task that could itself constitute a research project of its own. The second reflection also stems from the effects of globalization and regards the increasing diffusion of the so-called Global English or English as a Global Language (Crystal 2003), English as a Lingua Franca (Seidlhofer 2004) and related phenomena. This ongoing trend of using English to communicate (in this case) science at the global level, on the one hand, conveniently collects all non-native varieties under one umbrella term. On the other, it tends to make efforts to identify non-native varieties little relevant, unless – again – for purely sociolinguistic aims, which are not (or not exclusively) those of this project. Similarly, if English as a lingua franca brings together those who write about science by facilitating communication across the world, the often international, inter-linguistic nature of research teams around the world makes it difficult to impossible to attribute authorship univocally to any of the individuals signing, for example, a research paper. For all the above reasons, the main publications in the bioethics field, irrespective of their place of publication, will be considered representative of English varieties at large, with a necessary prevalence of BrE and AmE reflecting the current statistic production and distribution of scientific research. The PC will try and include a choice of publications from English-speaking countries adopting a main standard variety of English. The criteria for building the PC will furthermore have to evaluate that the popular press, everywhere in the world, notoriously reports on major issues. If a specific bioethical debate happens to be going on in one specific English-speaking country at a given time during the period considered, the chance of that publication and of texts from that publication / country / variety being over-represented is very high. The possible calibration of such events by the statisticians in the project will have to be taken into consideration.

Concerning the diamesic level, the written media will be favoured, for the purposes of Corpus Linguistics analysis contemplated by the Pro.Bio.Dic. For the same reasons, but also because the project is itself a child of the digital revolution and of the new media that have widened the participatory frame, digital editions will obviously be preferred. As regards the choice between the online or offline editions of publications (with some

of them including exactly the same material, and others reserving special content for either edition), irrespective of the selection made, the consistency criterion must apply to all the publications of one corpus and if possible to those of both corpora.

The diastratic variation does not really apply to the SC, as research papers are expected to be written in academic English, sharing approximately similar standard features across countries. The PC, instead, is conceived as having to be representative of a wide sample of popular English. The classic 'broadsheets vs tabloids' distinction, which indeed can be represented in the selection, only strictly applies to UK newspapers (see e.g. Bell and Garrett 1998, Fairclough 1995). Newspapers published in the US and in other English-speaking countries are of course also placeable at various diastratic levels according to their readership. Well-known macro-distinctions such as that between the *Washington Post* (quality) and the *New York Post* (popular) in the US, for instance, may be integrated by a finer socio-linguistic evaluation of local readerships, especially as regards less well-known newspapers, i.e. those of smaller English-speaking countries.

The diaphasic variation in academic publications regards a small choice of genres. Regular research articles feature alongside with less frequent but no less important genres, e.g. the short article, the editorial, the review, etc. While diastratically they would all employ academic English, significant differences may apply to their quantity (length) and quality (e.g. a personal or impersonal stance and subsequent linguistic choices). Popular newspapers offer a much wider sample of popular genres, including the editorial, the feature article, the review, the agony aunt column, etc. The choice of from what genres to draw the texts for the corpora may vary, for instance by considering only one genre per corpus (e.g. the full research article for the SP and the feature article for the PC), but again it will have to be consistent for each publication used for each corpus.

Other factors to consider in the selection are of a more practical nature. Size and availability are the first concerns when very large archives are needed such as for the present project. Whether to employ existing archives or embark on putting together novel ones depends on both the existing or procurable financial and human resources, in turn depending on and influencing the time estimated or allotted to search and collect the material. The medium (support) of the archives are also fundamental, as digital texts are needed for Corpus Linguistics analysis, and digitizing printed matter of course requires time and also depends on financial and human resources. Any digital format, whether on- or off-line, e.g. the CD/DVD-ROM, the downloadable or just browsable internet archive, etc. The prices and possible limits (of time, quantity, users) of access to the archives also count. An important element too is the presence or absence of internal search tools: does the archive have its own

search engine, what kind of search engine is it, does it allow advanced (multiple criteria) searches, does it employ Boolean language, etc. – these are all possible issues.

### *2.1. Sample archives for the PC*

Sample publications as identified by the general criteria set out in § 2 may be represented by the following examples of digital newspapers archives. These include a first group comprising two quality and two popular (in the sense of 'tabloid') British newspapers, and four corresponding US ones. Of course, more resources can and will probably be considered, especially according to the considerations made about the diatopic variation in § 2 above. Whereas for specialised publications full access is usually by subscription and sometimes only for research institutions (see the following § 2.2 on the SC), popular publications such as newspapers are obviously of common interest to the general public, which is also their intended readership and, as such, very often searchable archives are offered online for free. Reported below are therefore the direct links to the archives themselves.

*Table 1 – Digital archives of British quality and popular newspapers*

| Resource | Features |
|---|---|
| *The Guardian (1821-2003) and Observer (1791-2003) Digital Archive* http://pqasb.pqarchiver.com/guardian/advancedsearch.html | • Period: both stop in 2003; do not include contemporary debate.<br>• Archive is slowly being integrated, hopefully catching up with current time.<br>• Available: online via website.<br>• Price: £ 49.95 / month. |
| *The Financial Times Historical Archive (1888-2008)* http://gale.cengage.co.uk/financial-times-historical-archive.aspx | • Period: 1888- 2008.<br>• Available: online via website.<br>• Price: subscription only open to institutions, price not public. |
| *The Times Digital Archive (1785-2006)* http://gale.cengage.co.uk/times-digital-archive/times-digital-archive-17852006.aspx | • Period: 1785-2006.<br>• Available: online via website.<br>• Search tools: own search interface, engine, viewer.<br>• Price: £ 1,413.75 ca. (JISC Collections 2012) / year. |
| *The Daily Express Archive (1900-current)* http://www.ukpressonline.co.uk/ukpressonline/?sf=express | • Period: 1900 to present day.<br>• Available: online via website<br>• Price: £4,626.09/year for Universities. |

*Table 2 – Digital archives of US quality and popular newspapers*

| Resource | Features |
|---|---|
| *The New York Times Article Archive (1851-present)* http://www.nytimes.com/ref/membercenter/nytarchive.html | • Period: 1851-present day.<br>• Available: online via website.<br>• Price: post-1986: first 100 articles free, no limits for subscribers (subscription: ca. $ 180.00). |
| *The Washington Post (1877-current)* http://www.washingtonpost.com/wp-srv/newssearch/ | • Period: 1877-present day.<br>• Available: digital downloads (various formats: *.txt, *.PDF, etc.).<br>• Price: $29.95/25 articles. |
| *The Wall Street Journal (1979-present)* http://online.wsj.com/public/page/public_home_search.html | • Period: 1979-present.<br>• Available: digital downloads.<br>• Price: $2.95/article older than 90 days. |
| *The New York Post Archive (1998-present)* http://www.nypost.com/nypostarchives | • Period: 1998-present.<br>• Available: Online html versions.<br>• Price: Free. |

### *2.2. Specialised corpus*

A specialised corpus is a corpus which includes texts on a specific subject area. This specialisation has no definite boundaries, but some peculiar criteria need to be established to specify the type of the texts in question (Sinclair 2004). Typically such corpora may contain either some texts specialised in terms of a particular genre, topic (i.e. art, politics, medicine), or sub-domain (i.e. anatomy, ophthalmology, informed consent forms, informative material on HIV/AIDS). The specialised corpus on bioethics to be assembled for this project will be divided into two sub-corpora as the distinction in terms of target reader and editorial context has proved to be significant.

The former sub-corpus comprises specialised publications specifically dealing with bioethics, such as journals (in both print and online format), books, websites, etc. The number of publications and websites in the last decades testifies to a growing demand for bioethical competence, especially in specific domains, such as nursing and engineering (see Johnstone 2004 and Vallero 2007). The table below shows some examples (both websites and specialised publications) of resources selected for this sub-corpus through Google search queries (keywords: bioethics journal OR review) and meta-resources (i.e. Bioethics Resources on the Web http://bioethics.od.nih.gov/index.html)

*Table 3 – Specialised sub-corpus focused on bioethics*

| Resource | Features | Description |
|---|---|---|
| *Journal of Medical Ethics* http://jme.bmj.com/ | International, UK-based. Online & print. Monthly. Launch date 1975. Requires subscription. Some contents free. PDF/HTML. | *Journal of Medical Ethics* is a leading international journal that reflects the whole field of medical ethics. The journal seeks to promote ethical reflection and conduct in scientific research and medical practice. It features original articles on ethical aspects of health care, as well as case conferences, book reviews, editorials, correspondence, news and notes. JME has Editorial Board members from all around the world including the US, Europe, Australasia and Far East. |
| *Medical Humanities* http://mh.bmj.com/ | International, UK-based. Online & print. Bi-annually. Launch date 2000. Requires subscription. Some contents free. PDF/HTML. | *Medical Humanities* is a leading international journal that reflects the whole field of medical humanities. It features original articles relevant to the delivery of healthcare, the formulation of public health policy, the experience of being ill and of caring for those who are ill, as well as case conferences, educational case studies, book, film, and art reviews, editorials, correspondence, news and notes. Medical Humanities has Editorial Board members from all around the world. |
| *American Journal of Bioethics* http://www.bioethics.net/ | US, US English. Online & print. Monthly. Launch date 1999. Contents can be browsed by topics. Requires subscription. Some contents free. PDF/HTML. | *The American Journal of Bioethics* (AJOB) is a monthly peer-reviewed academic journal of bioethicspublished by Taylor and Francis. It publishes target articles, peer commentary, book reviews, qualitative research, literary criticism, photography and graphic arts, and comments on developments in law and medicine. |
| *The Journal of Clinical Ethics* http://www.clinicalethics.com/ | US, US English. Online & print. Quarterly. Requires subscription. HTML. | *The Journal of Clinical Ethics* is written for an by physicians, nurses, attorneys, clergy, ethicists, and others whose decisions directly affect patients. JCE is a double-blinded, peer-reviewed journal indexed in PubMed, Current Contents/Social & Behavioral Sciences, the Cumulative Index to Nursing & Allied Health Literature, and other indexes. The Journal of Clinical Ethics is an American Society of Bioethics and Humanities partner journal. |

| Resource | Features | Description |
|---|---|---|
| *Cambridge Quarterly of Healthcare Ethics* http://journals.cambridge.org/action/displayJournal?jid=CQH | International, UK-based. Online & print. Quarterly. Launch date 1992. Requires subscription. HTML | *The Cambridge Quarterly of Healthcare Ethics* is designed to address the challenges of biology, medicine and healthcare and to meet the needs of professionals serving on healthcare ethics committees in hospitals, nursing homes, hospices and rehabilitation centres. The aim of the journal is to serve as the international forum for the wide range of serious and urgent issues faced by members of healthcare ethics committees, physicians, nurses, social workers, clergy, lawyers and community representatives. |
| *The Hastings Center Report* http://www.thehastingscenter.org/Publications/HCR/ | International, US-based. Online and print. Bi-monthly. Requires subscription. Some contents free. HTML. | *The Hastings Center Report* is a bi-monthly journal that promotes incisive and balanced inquiry into the ethical issues in health, medicine, and the environment. It includes essays, commentary, original scholarly articles, and occasional Special Reports. The Hastings Center is an independent, non-partisan, and non-profit bioethics research institute founded in 1969. |
| *IRB: Ethics & Human Research* http://www.thehastingscenter.org/Publications/IRB/Default.aspx | International, US-based. Print. Bi-monthly. Launch date 2007. Requires subscription. | *IRB: Ethics & Human Research* explores issues in research with human subjects, including findings and analyses of empirical studies. Six issues are published each year, containing scholarly articles and columns. All submissions are peer-reviewed. IRB's readership includes administrators and members of institutional review boards, scholars, and researchers. The journal is issued by the Hastings Center (see above). |
| *The Journal of Medicine and Philosophy* http://jmp.oxfordjournals.org/ | International, UK-based. Online and print. Bi-annually. Launch date 1979. Requires subscription. Some contents free. PDF/HTML | *The Journal of Medicine and Philosophy* is one of the leading scholarly journals in bioethics and the philosophy of medicine. Its contributors and focus are international, addressing bioethical concerns across the world. Significant attention has been given to bioethics and foundational issues in health care policy in North and South America, Europe, and Asia. The journal's concerns range from clinical bioethics to studies in the philosophy of medicine, such as explorations of the nature of concepts of health and disease, as well as the character of medical explanation. |

| Resource | Features | Description |
|---|---|---|
| *Kennedy Institute of Ethics Journal* http://muse.jhu.edu/journals/kennedy_institute_of_ethics_journal/ | US, US English. Online and print. Bi-annually. Launch date 1991. Requires subscription. Some contents free. PDF/HTML. | *The Kennedy Institute of Ethics Journal (KIEJ)* offers a scholarly forum for diverse views on major issues in bioethics, including analysis and critique of bioethics theories such as principlism and feminist perspectives in bioethics; the work of federal bodies such as the President's Council on Bioethics; and a wide range of topics such as enhancement technologies, health care reform, stem cell research, and organ transplantation. The Kennedy Institute of Ethics is one of the world's premier bioethics institutes. |
| *Nursing Ethics* http://nej.sagepub.com/ | US, US English. Online & print. Bi-monthly. Launch date 1994. Requires subscription. PDF. | *Nursing Ethics* is an international peer reviewed journal that takes a practical approach to this complex subject and relates each topic to the working environment. The international Editorial Board ensures the selection of a wide range of high quality articles of global significance. This journal is a member of the Committee on Publication Ethics (COPE) |
| *The Journal of Law, Medicine, and Ethics* http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN% 291748-720X | US, US English. Print. Quarterly. Launch 1973. Requires subscription. PDF . | *Material published in The Journal of Law, Medicine & Ethics* (JLME) contributes to the educational mission of The American Society of Law, Medicine & Ethics, covering public health, health disparities, patient safety and quality of care, and biomedical science and research. It provides articles on such timely topics as health care quality and access, managed care, pain relief, genetics, child/maternal health, reproductive health, informed consent, assisted dying, ethics committees, HIV/AIDS, and public health. |
| *Bioethics* http://www.blackwellpublishing.com/journal.asp?ref=0269-9702 | International, US-based. Online and print. Quaterly. Launch date 1987. Requires subscription. Some contents free. HTML. | *Bioethics* provides a forum for well-argued articles on the ethical questions raised by current issues such as: international collaborative clinical research in developing countries, organ transplants and xenotransplantation, ageing and the human lifespan, AIDS, genomics, and stem cell research. |

| Resource | Features | Description |
|---|---|---|
| *Eubios Journal of Asian and International Bioethics* (*EJAIB*) http://eubios.info/EJAIB.htm | International, US English. Online. Bi-monthly. Launch date 2000. Free. Supported by UNESCO. PDF. | *EJAIB* is the official journal of the Asian Bioethics Association (ABA). Its aim is to publish research papers, and relevant news, and letters, on topics within Asian Bioethics, promoting research in bioethics in the Asian region, and contributing to the interchange of ideas within and between Asia and global international bioethics. Asia is defined for the general purposes of this journal as the geographical area, including the Far East, China, South East Asia, Oceania, the Indian subcontinent, the Islamic world and Israel. |
| *Science and Engineering Ethics* http://www.springerlink.com/content/1471-5546 | Germany, UK English. Online. Quarterly. Launch date 1995. Requires subscription. Various contents open access. PDF/HTML. | A quarterly journal with articles on ethical issues of concern to scientists and engineers. Special topic issues. Includes many articles on responsible conduct in research. |
| *Yale Journal of Health Policy, Law, and Ethics* http://www.yale.edu/yjhple/ | US, US English. Online. Bi-annually. Launch date 2001. Requires subscription. Some contents free. PDF. | *The Yale Journal of Health Policy, Law, and Ethics* is a biannual publication of the Yale Schools of Law, Medicine, Epidemiology and Public Health, and Nursing. The Journal strives to provide a forum for interdisciplinary discussion on topics in health policy, health law, and biomedical ethics. It targets a broad and diverse readership of academicians, professionals, and students in medicine, law, and public health, as well as policy makers and legislators in health care. |

The latter sub-corpus consists in specialised medical publications that are not solely centred on bioethical issues (Table 4). It also includes journals (both in print and online format), books, websites, web portals, etc. Some of these resources, such as *PubMed.org*, consist in huge libraries of specialised articles.

All the resources of both sub-corpora are to be inserted, catalogued and indexed in a database according to these criteria: topic, text genre, availability, price, target reader, frequency, content, usability, and other meta-data aspects (see § 2). This will help establish what contents are most suitable for the project, as the two sub-corpora are expected to produce a large amount of data.

*Table 4 – Specialised sub-corpus not focused on bioethics*

| Resource | Features | Description |
|---|---|---|
| *US National Library of Medicine PubMed.org* www.pubmed.org | US-based. Online. Web portal. Requires subscription, some contents are free. HTML and XML. | US National Library of Medicine National Institutes of Health. *PubMed* comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites. |
| *British Medical Journal* www.bmj.com | UK English, international. Online and print. Weekly in three editions. Launched in 1840. Requires subscription, some contents are open access. PDF and HTML. | The *British Medical Journal* is an international peer reviewed medical journal and a fully "online first" publication. All articles appear on bmj.com before being included in an issue of the print journal (continuous publication). The website is updated daily with the *BMJ*'s latest original research, education, news, and comment articles, as well as podcasts, videos, and blogs. |
| *Medical Journal of Australia* www.mja.com.au | Australian English. Australasia-based. Online and print. Bi-monthly. Requires subscription, some contents are open access. HTML and PDF. | The *Medical Journal of Australia* (*MJA*) is Australia's leading peer-reviewed general medical journal. It covers all the important issues affecting Australian health care, publishing the latest Australian clinical research, evidence-based reviews, clinical practice updates, authoritative medical opinion and debate, and developments within the humanities with respect to medicine. The *MJA* encourages comment and debate from readers. |
| *Canadian Medical Association Journal* www.cmaj.ca | Canada-based. Online (weekly) print (18 issues per year). Launched in 1911. Open access. PDF and HTML. | *CMAJ* showcases innovative research and ideas aimed at improving health for people in Canada and globally. It publishes original clinical research, analyses and reviews, news, practice updates and thought-provoking editorials. *CMAJ*'s impact factor is 8.2 and the website receives over 2 million unique visitors a year. |

In order to show the significance of the division between the two sub-corpora in terms of communication type, two examples from the resources described above can now be discussed: the *Journal of Medical Ethics* (first sub-corpus) and the *British Medical Journal* (second sub-corpus). The first is an international specialised journal (although it is UK-based). It requires subscription (but some sample contents are free), it is published in both print and online version, the online contents are available in both PDF and HTML format. The main text genre is the academic paper. The *Journal of Medical Ethics* is specialised in bioethics, the communication type is intraspecialistic, although this discipline is per se multidisciplinary. The main issues actually concern matters of medicine, law, philosophy and even psychology (especially as for the relationship between healthcare specialist and patient). The second example of specialised resource not exclusively dealing with bioethical issues is the *British Medical Journal*. It is also UK-based and requires subscription (but some sample contents are free), it is published in both print and online version, the online contents are available in both PDF and HTML format. The main text genre is academic paper. The communication type is intraspecialistic and interspecialistic, and lexis is mainly technical. For the purposes of this project the multidisciplinary nature of bioethics as a subject area and the inter- and intraspecific communicative dynamics play therefore a key role in the design of the text corpora.

## 3. Conclusions

The phase of selection of the criteria needed to correctly assemble the two corpora herein described is being accompanied by some preliminary tests on limited and small portions of texts carried out by the computer science and the linguistic teams involved in the project. These are necessary to a) verify whether the methodology to be employed is valid; b) assess which corpora / texts / documents could be more suitable for and representative of the scopes envisaged and c) develop methods and algorithms capable of extracting sets of terms to be included in the dictionary. After this preliminary phase, the computer science team will be able to pass on the lemmas extracted to the committee of experts for evaluation, with the top-ranked terms being the most authoritative candidates for inclusion. Several tests are needed to develop many such algorithms to associate to each extracted term a numerical estimate of the probability that the term is indeed bioethics-related, as well as to generate lists of terms characterised by the smallest possible quantity of spurious terms. One such recent sample tests consisted in the selection of 100 specialized texts on bioethics, 100 specialized texts belonging to a field / discipline other than bioethics, 200 (150 for machine training, 50 for testing) popular texts on very different topics than

bioethics, 20 popular texts on bioethics for testing, 20 popular texts on topics close to bioethics, according to the methodology summarised in § 1. This proved that the machines can learn to recognise texts dealing with bioethics and thus validated the methodology devised. Other sample tests on the possible corpora to include in the project will contribute to making the latter more robust and verifying the consistency and representativeness of the sets of documents / texts selected for analysis.

BIBLIOGRAPHY


1. Avancini, H., Lavelli, A, Sebastiani, F. & R Zanoli 2006, *Automatic Expansion of Domain-Specific Lexicons by Term Categorization. ACM Transactions on Speech and Language Technology*, 3(1), pp.1-30.
2. Bell, A. & P. Garrett 1998, *Approaches to Media Discourse*, Oxford, Malden, Blackwell.
3. Bellini, P. 2008, "La bioetica tra naturale e artificiale", in METABASIS.IT (www.metabasis.it), 3(6), pp. 1-9.
4. Bellini, P. 2012, "Pour une éthique de la technique", in "Repenser la nature. Dialogue philosophique Europe, Asie, Amériques", sous la direction de J.-P. Pierron et M. H. Parizeau, Quèbec, Presses de l'Université Laval, pp. 79-92.
5. Bioethics Resources on the Web 2012, http://bioethics.od.nih.gov/index.html.
6. Bishop, C. 2006, *Pattern recognition and machine learning, Heidelberg, Springer.*.
7. Branningan, M. C. 2001, Medical issues in human cloning: cross disciplinary perspectives, New York, Seven Bridges.
8. Brannigan M. C. 2004, Ethics across cultures, New York, McGraw-Hill.
9. Brügger, N. & H. Bødker. (eds) 2002, *The Internet and Society?,* Århus, The Centre for Internet Research.
10. Cloître, M. & T. Shinn 1985, "Expository practice: social, cognitive and epistimiological linkages", in T. Shinn & R. Whitley (eds), *Expository science. Forms and functions of popularization*, Reidel, Dordrecht, pp. 31-60.
11. Crystal, D. 2003, *English as a Global Language*, 2nd ed., Cambridge, Cambridge University Press.
12. Fairclough, N. 1995, *Media Discourse*, London, Edward Arnold.
13. Hundt, M., Nesselhauf, N. & Biewer, C. 2007, *Corpus linguistics and the web*, Rodopi.
14. Joachims, T. 2002, *Learning to Classify Text using Support Vector Machines*, Dordrecht, Kluwer.
15. Johnstone M. J. 2004, *Bioethics*, Churchill Linvingstone, Chatswood.
16. Jonas, H. 1997, *Tecnica, medicina ed etica. Prassi del principio di responsabilità*, a cura di P. Becchi, Torino, Einaudi, pp. 122-154.
17. Jonsen, A., R. 1998, *The Birth of Bioethics*, New York, Oxford University Press.
18. Jurafsky, D., Martin, J. H., & Kehler, A., 2008*, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition,* Boston, MIT Press.
19. Liu, B., Xu, Z. & Wang, X. 2007, *Automatic Domain-Specific Term Extraction and Its Application in Text Classification, Acta Electronica Sinica 2007-02.*
20. Manning, C. D., Raghavan, P. & H. Schuetze 2008,. *Introduction to Information Retrieval, Cambridge,* Cambridge University Press.

21. Mooney, A. & S. Sarangi 2003, "Click here for health information/advice: interaction pathways via the NHS Direct Website", in Crivelli, B. & S. Rubinelli. (eds), *Televisione, stampa e internet tra medico e paziente*, Numero speciale di Tribuna Medica Ticinese, pp. 21-27.
22. Sebastiani, F. 2002, "Machine learning in automated text categorization", ACM Computing Surveys 34(1), pp. 1-47.
23. Sebastiani, F. *2006, Classification of text, automatic,* in Brown K. (ed.), *The Encyclopedia of Language and Linguistics,* Amsterdam, Elsevier Science Publishers, vol. 14, 2nd ed., pp. 457-462.
24. Seidlhofer, B. 2004, "Research perspectives on teaching English as a Lingua Franca", in *Annual Review of Applied Linguistics*, 24, pp. 209-239.
25. Sinclair, J. (ed.) 1989/1995, *Collins Cobuild. Essential English Dictionary*, London, Harper Collins.
26. Sinclair, J. 1991, *Corpus, concordance, collocation*, Oxford, Oxford University Press.
27. Sinclair, J. 2004, *Trust the Text: Language, corpus and discourse,* London, Routledge.
28. Slevin, J. 2002, "The Internet and Society: central themes and issues", in Brügger, N. & H. Bødker, (eds) 2002, *The Internet and Society?,* Århus, The Centre for Internet Research, pp. 7-12.
29. Thomas, J. & M. Short (eds) 1996, *Using Corpora for Language Research*, London, Longman.
30. Vallero D. 2007, *Biomedical Ethics for Engineers*, London, Elsevier.
31. Vicentini, A., Grego, K., Berti, B., Bellini, P. & G. Orizio 2010, "Medical cybercommunication: imaginaries, practices, languages - Toward A Dictionary of Bioethics Terms", presented at the 8th Conference on Communication, Medicine and Ethics (COMET) 2010, Boston University, Boston (USA), 28-30 June 2010.
32. Vicentini, A., Grego, K., Berti, B., Bellini, P. & G. Orizio 2012, *Intercultural and Ideological Issues in Lexicography: A Prototype of a Bioethics Dictionary*, in Roberta F. (ed*.), English Dictionaries as Cultural Mines*, Newcastle, Cambridge Scholars Publishing, pp. 247-264.

Alessandra Vicentini, Kim Grego, Daniele Russo

# Pro.Bio.Dic. projekt

SAŽETAK

Ovaj članak proizlazi iz dugoročnoga istraživačkoga projekta, koji je u tijeku, u području medicinske etike i bioetike. U projektu su uključeni različiti timovi stručnjaka iz medicine, jezikoslovlja, računalnih znanosti i filozofije. Cilj ovog rada je izrada prototipa digitalnoga bioetičkoga rječnika (Pro.Bio.Dic.) za poboljšanje u razumijevanju pojmova iz područja bioetike koja se spomenutim rječnikom žele postići. Potom se navode planovi kojima se želi osigurati kvalitativna i kvantitativna potpora izgradnji budućega korpusa rječnika. U ovome članku prikazana su načela u odabiru leksičkoga materijala odnosno skupine tekstova na kojima bi se leksik temeljio te nalaze se primjeri digitalnih izvora i osnovne karakteristike za nespecijalizirani i specijalizirani dio korpusa.