A TERMINOLOGY-BASED RE-DEFINITION OF GREY LITERATURE

Claudia Marzi, Gabriella Pardelli, Manuela Sassi
Institute for Computational Linguistics "*Antonio Zampolli*"
National Research Council of Italy (*CNR*)
Via G. Moruzzi 1, 56124 Pisa, Italia
[claudia.marzi, gabriella.pardelli, manuela.sassi]@ilc.cnr.it

Abstract

The conventionally accepted definition of Grey Literature, as Information produced and distributed by non-commercial publishing, does not take into consideration either the increasing availability of forms of grey knowledge, or the growing importance of computer-based encoding and management as the standard mode of creating and developing grey literature.

Semi-automated terminological analysis of almost twenty years of terminological creativity in the proceedings of eleven GL International Conferences offers the opportunity to pave the way to a bottom-up redefinition of Grey Literature stemming from attested terminological creativity and lexical innovation.

In this paper, we focus on a set of automatically-acquired terms obtained by subjecting our reference Corpus to a number of pre-processing steps of automated text analysis, such as concordances, frequency lists and lexical association scores. Acquired terms allow us to throw in sharp relief developing trends and important shifts of emphasis in the current understanding of the notion of Grey Literature.

Theme: ***Redefining Grey Literature*** *– Proof of Concept – Research Process*

Keywords: Grey Literature definition, GL Conference corpus, Terminology extraction

1. Introduction

1.1 Grey Literature definition

The Luxembourg Convention on Grey Literature held in 1997 offered the following definition of Grey Literature (expanded in New York, 2004): "Information produced *and distributed* on all levels

of government, academics, business and industry in electronic and print formats not controlled by commercial publishing*, i.e. where publishing is not the primary activity of the producing body".*

The questions that immediately arise are the following: is this definition still valuable? Is it so far completely satisfactory? Or does it rather need important modifications?

And what about other conventionally accepted definitions and descriptions?

In considering the evolution of the role and definition of Grey Literature, Augur (1989) started from the beginning of the 20th century, where the notion of GL had been, for many years, coextensive with that of report literature: documents evolving out form research and development activities, particularly in the aircraft and aeronautics industries, were a very important means of communicating the results of research testing. In particular, World War Two had the greatest impact on report literature, transforming it into a major vehicle of communication. By the 1970s GL became the recognized medium for dissemination and promotion for many organizations and was considered an important reading throughout the world, though not easy to find. By the 1980s other scientific domains such as Social Sciences, Economics and the Humanities were included in the wide range of research reports, discussion and policy documents, working and conferences papers, etc. A huge increase in quantity as well as the advantageous effect of the flexibility and speed, however, didn't completely obscure problems of identification and acquisition; given the nature of this kind of literature, many categories contained security restrictions. In the 1990s GL attained its importance as an independent medium of communication because of an initial need for security of confidentiality classifications which prevented documents from being published in a conventional manner.

Hirtle (1991) gave a definition of GL as "the quasi-printed reports, unpublished but circulated papers, unpublished proceedings of conferences, printed programs from conferences, and the other non-unique material which seems to constitute the bulk of our modern manuscript collections".

IGLWG (*Interagency Gray Literature Working Group*) defined in 1995 GL as "open source material that usually is available through specialized channels and may not enter normal channels or

systems of publication, distribution, bibliographical control, or acquisition by booksellers or subscription agents".

Debachere (1995) described GL as "a range of materials that cannot be found easily through conventionally channels [..] but which is frequently original and usually recent".

Actually, quoting Wikipedia "Grey Literature is a term used [..] to refer to a body of materials that cannot be found easily through conventional channels such as publishers [..]".

All these descriptions of Grey Literature are phrased negatively; often GL is defined by contrast to other things. In other words, we notice that particular emphasis is laid on what GL is not, rather than on what it is.

To sum up, all these definitions and descriptions of Grey Literature do not take into account those aspects that, in our view, are most strongly associated with the increasing availability and accessibility of GL materials, and the growing importance of computer-based encoding as the standard medium of creating and developing GL.

Our general idea is that a domain-specific document repository offers the possibility to pave the way to a bottom-up redefinition of Grey Literature stemming from attested terminological creativity and lexical innovation.

We intend to inquire and monitor terminological creativity over almost twenty years of technical and scientific work in the frame of the International Conference on Grey Literature, and to ground suggestions for a re-definition on those terms that appear to be consensually shared by the various disciplinary sub-fields.

1.2 Reference corpus

The empirical basis of our work is represented by the Corpus of *GreyText Inhouse Archive,* available on http://www.greynet.org/opensiglerepository.html consisting of titles, themes, keywords and full abstracts, for a total amount of around ninety thousand tokens (containing around seventy thousand word tokens).

Although comparatively small, the corpus suits the purposes of our present investigation quite nicely. First, it is fairly well structured, allowing selective search of relevant terms in a context-sensitive way. Moreover, it contains highly informative text excerpts, as titles and abstracts are, conveying document contents in a quintessential way. The traditional haystack problem in information extraction from unstructured materials is here considerably reduced, as all texts belonging to the corpus are characterized by a high density of mostly salient terms. Thirdly, the corpus presents a longitudinal selection of documents ranging over several years of intensive research in GL. This will allow a terminological trend analysis in a diachronic perspective.

2. Methodology

2.1 Research rationale

Our general idea is that an interesting re-definition of GL can be based upon careful examination of the longitudinal trend of almost twenty years of terminological creativity in the proceedings of the eleven GL international Conferences, by focussing on a set of automatically-acquired terms (both single-word and multi-word terms) obtained by subjecting our reference Corpus to a number of pre-processing steps of automated text analysis, such as concordances, frequency lists and lexical association scores (e.g. *Mutual Information* on word pairs).

Although knowledge-poor, bag-of-words approaches to text mining have proved to perform effectively in traditional tasks such as document classification and indexing, intelligent access to the contents of a document repository requires going beyond the over-simplistic notion of a text as an unordered collection of loose word tokens. Automated identification of the most relevant terms in a domain-specific document repository represents an important step in this direction. It is commonly assumed that salient domain-specific concepts and relations are conveyed in text through statistically significant terms, whether they are simple words like *computer* and *web*, or structurally more complex word sequences like *computer science* and *world wide web*. This requires that a raw text is preliminarily marked up at different levels of linguistic analysis, ranging from tokenization and part-of-speech tagging, to chunking and dependency analysis. Relevant

terminological units are then tracked down by projecting abstract morpho-syntactic patterns such as “NP PP” (*i.e.* “find a syntactic structure made up out of a Noun Phrase immediately followed by a Prepositional Phrase) onto linguistically annotated texts. All text strings that fit into the targeted morpho-syntactic pattern (e.g. *networks of institutional repositories*) are then filtered out through a further step of statistical post-processing, to assess their potential for termhood.

Filtering methods considerably vary in the literature, ranging from raw frequency lists and traditional Information Retrieval measures such as TF-IDF (Baeza-Yates & Ribeiro-Neto, 1999), to more sophisticated indices like the C/NC-value (Frantzi et al., 2000) or lexical association functions such as “log likelihood” and “point-wise mutual information” (Manning & Schütze, 1999). The result of this filtering step is a list of relevant term candidates, possibly to be validated by a domain expert but already usable for advanced content indexing.

In fact, more can be done on the way to understanding their content and the role they play in a document repository. With a view to meeting these further goals, we need to take into account the particular context where terms occur, the network of textual relations they entertain with other words and the semantic roles they play. Such a finer-grained analysis can be carried out in many ways: i) manually, through inspection/classification of a relevant list of concordances of the terms of interest, ii) semi-automatically, by automatically clustering words that occur in the same contexts (Lenci et al., 2006), and then having experts classify the resulting clusters; iii) fully-automatically, by clustering words and then discovering their semantic relations by using machine learning techniques (Mitchell, 1995).

For the present purposes, a context-sensitive analysis of relevant terms (domain specific word forms) was carried out through manual inspection of relevant list of concordances and frequency (step i above), since the other approaches require availability of a considerably larger amount of textual data. Nonetheless, we believe that our preliminary analysis illustrates the potential of the corpus-based approach to domain definition we propose here.

2.2 Data extraction

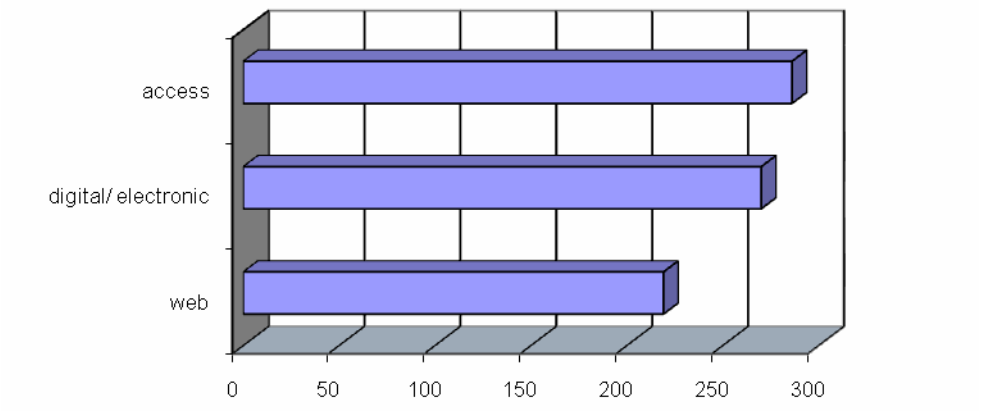
We started from a single-word frequency list, acquired automatically from around 90,000 tokens. Items in the list, that contains words that occur more than 9 times (an empirical threshold corresponding to 0.01% of the total size of our reference corpus), were ranked by decreasing frequency values to bring the most relevant terms to the top, as shown below.

985	grey
966	literature
737	**information**
477	research
220	**access**, conference
204	library
202	paper
191	**documents**
182	scientific
178	proceedings
162	development
159	project
142	**electronic**
137	system
130	use
128	science
128	**digital**
126	report
124	available
121	libraries, open, technical
120	data, national
119	collection
106	public
104	**knowledge**
103	publishing, university
102	study
100	results
98	international
97	resources, technology
95	metadata
92	repository
91	users
88	published
85	database, publications, **web**
84	academic, document
83	analysis
82	authors
80	communication
77	management
75	services
74	countries
73	researchers
72	government, **internet**, repositories
70	work
69	materials
68	health, projects, world
67	databases
65	community
64	bibliographic, theses
63	survey, systems
61	european, sigle, social
57	content
40	**accessible**
32	**online**
30	**www**
27	amsterdam, archive, author, economics, nature, reference, references
26	**accessibility**, agencies, citations, formats, intellectual, technological
25	networked, political, professionals, security, standards
24	communities, concept, industry, language, legal, virtual
23	italy, paradigm, physics, societies, uk
22	exchange, goal, japanese, preservation, purpose, scientists, sharing, useful
21	learning, retrieval, significant, topic
20	governmental, identification, networking, property, questions
19	historical, marine
18	india
17	school, searching, site, sites, tool, transfer
16	botswana, market, model, worldwide
15	agricultural, benefit, distance, financial, free,

	multicultural, multiethnic, poland, questionnaire		multimedia, visibility	10	america, benefits, bibliography, catalogue, collaboration, culture, engineering, engines, france, germany, literatures, media, participation, website
		12	catalogue, techniques, unpublished		
14	communications, companies, datasets, delivery, educational, networks, Russia	11	cognitive, czech, freedom, method, methodology, rural, semantic, words		
13	african, bank, catalogs, cooperation, cultural,				

Frequency distributions were then aggregated by putting in the same frequency class quasi-synonyms or semantically closely related terms (e.g. *internet, web* and *www*). This post-processing step allowed us to assess how often a concept, or ontological entity, was used in the corpus.

The analysis has been centred on those concepts which appear innovative with respect to the traditional definitions of Grey Literature reported above. Accordingly, core notions such as *information* and *documents*, which figure prominently in our list, although undoubtedly relevant to a proper characterization of GL, are taken to be too well established to deserve further analysis. Here, we rather intend to focus on highly salient concepts that appear to be shared by various disciplinary sub-fields, and mark, in our view, important steps in the evolution of current understanding of GL. In particular, we selected three such notions: *digital/electronic*[1]*, access* and *web*. Their aggregated frequency distributions are shown in the diagram below:



1 Contrary to our expectations and a general terminological trend, the attribute *electronic* continues to be used interchangeably with *digital* to characterize a document and/or its content; therefore we are considering both of them.

As a further step in our analysis, we considered lexical association scores between salient terms (e.g. m*utual information* on word pairs), focussing on terminological usages that are closely related to the ontological entities already mentioned above.

The typical collocates of a*ccess* are: *easy, electronic, facilitate, full, grey, information, internet, journals, literature, materials, movement, multicultural, open, public, repository, research, scientific.*

Digital combines with *document, grey, library, literature, network, object, project, repository, system, technology, theses.*

Electronic keeps company with *format, grey, information, journal, literature, network, paper, publication, report, resource, technical, theses.*

Reference to the notion of *web* is typically accompanied by *access, database, grey, information, network, literature, science.*

Finally, particular emphasis should be placed, in our view, on the use of *knowledge* coupled with *base, exchange, generation, infrastructure, management, scientific, service, share, society,* and, especially, *information*.

3. Results

Term aggregation by conceptual unity and manual inspection of the most recurrent contexts of use of selected terms shed considerable light on both established and innovative notions. The steadily increasing occurrence of the attribute d*igital/electronic* bears witness to the growing importance of computer-based encoding as the *standard medium* of GL. Here, availability in digital format appears to be the outcome of an integrated system of software tools for efficient, possibly metadata-oriented document production and management, and an essential prerequisite to ubiquitous dissemination and ready accessibility.

The noun *access* (defining the process of accessing text documents), is seen in the company of adjectives like *easy*, *full* and *open*. The usage underlines important conceptual innovations in the way GL material is distributed and eventually used; e.g. *open access* focuses on the free accessibility and reusability of digital contents. Coupled with *information, document* and *repository,*

access appears to point to a conception of world-wide available, structured digital contents, offering the combined advantage of ubiquitous accessibility and quality control under authoritative document management. As Farace (2006) puts it, “open access to information is the key to knowledge, both in its generation and transfer”. The management of valued resources in a global environment is in fact conducive to the extraction and combination of targeted information and, eventually, to the generation of innovative knowledge. This perspective lays emphasis on the increasing importance of information management systems for GL, and casts doubts on those definitions of Grey Literature as “a mere characterization of the distribution mode” (as already pointed out by Mackenzie Owen, 1997).

Finally, systematic reference to the notion of *web* throws in relief the huge importance of the World Wide Web as the standard means of disseminating GL, and the role of networking communities, acting at the same time as providers and users of GL material in a highly distributed, collaborative scenario.

4. Concluding remarks

Grey Literature defines an innovative approach and methodology for a wide information dissemination and exchange, by offering the web-based sharing facilities and distributed access to openly available scientific and technical document repositories, possibly under authoritative content management.

An updated re-definition of GL should take into consideration the key notions of digital medium, web-based distribution channels, information access policy and access and management tools for GL. By bringing these innovative elements into the picture, we are in a position to do justice to recent developments in the evolution of GL, where traditional core notions such as *information*, *distributed access* and *electronic/digital format* appear to acquire novel, cooperative and interactive undertones, coupling the advantages of flexibility, speed and quantity, with the further bonus of ubiquitous accessibility and content quality control in a global cooperative environment. In fact, by blurring the traditional divide between providers and users of document repositories, GL not only defines a policy for distribution and access of information, but does promote new, creative modes

of production and use of innovative knowledge.

At its core, Grey Literature is about producing and distributing the seeds of new knowledge.

References

AUGUR, CHARLES P. (1989). *Information Sources in Grey Literature*. Bowker-Saur, London.

BAEZA-YATES R., RIBEIRO-NETO B. (1999). *Modern Information Retrieval.* Addison Wesley, ACM Press New York.

BAAYEN R. H. (2001). *Word Frequency Distributions.* Kluwer Academic Publishers, Dordrecht.

CARROLL B. C. AND COTTER G. A. (1997). A new generation of grey literature: The impact of advanced information technologies. *Publishing Research Quarterly*, 13 (2), 5-14.

DEBACHERE, M. C. (1995). Problems in obtaining grey literature. *IFL4 Journal,* 21 (2), 94-98.

CHURCH K. W., HANKS P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* n. 16 (I), 22-29.

FARACE D.J.(2006) (Guest Editor). Introduction: Open access to grey resources*. Publishing Research Quarterly*, 22,(1), 3.

FRANTZI K. T., ANANIADOU S., MIMA H. (2000). Automatic Recognition of Multi-Word Terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), 115-130.

FRIEDMAN, T.L. (2005). *The world is flat. A Brief History of the Twenty-first Century.* Farrar, Straus, Giroux New York.

HIRTLE, PETER (1991). *Broadsides vs. Grey Literature.*

JEFFERY K.G., ASSERSON A. (2007). Hyperactive Grey Objects. *Publishing Research Quarterly*, 23(1), 71-77.

ERIC C. KANSA, SARAH WHITCHER KANSA, MARGIE M. BURTON AND CINDY STANKOWSKI (2010). Googling the Grey: Open Data, Web Services, and Semantics. *Archaeologies*, 6 (2), 301-326.

LENCI A., MONTEMAGNI S., PIRRELLI V. (2006). Acquiring and Representing Meaning: Computational Perspectives. In A.. Lenci , S. Montemagni ,V. Pirrelli . (eds.) Acquisition and Representation of Word Meaning. Theoretical and computational perspectives. *Linguistica Computazionale*, XXII-XXIII, IEPI, Pisa-Roma, 19-66.

MACKENZIE OWEN J. (1997). The expanding horizon of grey literature. *GL3 Conference Proceedings. GreyNet*, AMSTERDAM,11-13.

MARCUS A. BANKS (1996). Towards a continuum of scholarship: The eventual collapse of the distinction between grey and non-grey literature. *Publishing Research Quarterly,* 22 (1), 4-11.

MARZI C., PARDELLI G., SASSI M. (2010). Grey Literature and Computational Linguistics: form Paper to Net*. GL11 Conference Proceedings*, TextRelease, Amsterdam,118-121.

MANNING C., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

MITCHELL TOM M. (1995), *Machine Learning*. McGraw Hill, New York.

SCHÖPFEL J. (2006). Observations on the future of Grey Literature. *The Grey Journal,* 2, (2), 67-76.

SEYMOUR D.J. (2010). In the Trenches Around the Ivory Tower: Introduction to Black-and-White Issues About the Grey. *Archaeologies*, 6,(2), 226-232.

DBT Software (CNR patent) http://www.ilc.cnr.it/pisystem/prodotti/index.html

GREYNET WEB SITE http://www.greynet.org