Khurshid Ahmad & Steven Collingham, University of Surrey

Renewable Terminology

Abstract

A term is coined, retrospectively defined, or made archaic usually through the medium of text. Once a term is coined or retronym proposed, then that becomes the basis of other texts and the text can be used as a basis for the evidence of the existence of a term and the potential ways in which the term could be elaborated. There is thus a synergy between the terminology of a domain and the text of the domain. Recently published texts can be searched on the basis of existing terms; these texts may contain neologisms or retronyms. Once validated these neologisms can be used as a basis for more focused search of still newer texts. A method for terminology extraction and validation using text corpora available on the global communications network is presented in this paper.

1. Introduction

Language growth and language change has been observed from a number of perspectives. These include socio-linguistic, psycholinguistic, historic, climatological, theological, economic and commercial and so on. More recently, attempts have been made to relate this growth and change to a corresponding growth and change in scientific thought. Such attempts indicate the ways in which lexicogrammatical resources of the scientists' general language are used in the coinage and elaboration of specialist terms. The following are good examples of the analysis of scientific writing: Geertz (1988) on anthropological writings, Heisenberg on the language of modern physics (1958, 1983), and Kuhn's analysis of the shift in the position of chemists in the late eighteenth century (from Priestley's *dephlogisticated air* and *elementary earth* to Lavoisier's and Scheels' *pure air* or *oxygen*, 1970:118). In these analyses one can see the link between terminologies, old, new and retrofitted, and scientific thought.

In this paper we argue that an understanding of scientific and technological writing is an important precursor for building terminology data bases, especially in the emergent disciplines: the features that distinguish between specialist writing with other linguistic outputs will play an important role in the identification of specialist terms. Furthermore, developments in corpus-based lexicography together with the availability of large quantities of computer-readable texts, particu-

larly through the burgeoning global computer-mediated communications network, makes it possible for us to design and build computer systems that will help in building term bases that contain neologisms and retronyms, and term bases that can systematically archive terms that are no longer used or are deemed to be archaic. We propose and describe a prototype system that can renew the terminology of a specialist domain by putting into operation the observations of those involved in studying scientific writing, those involved in building large corpora and those involved in searching for and capturing texts on the Internet and World Wide Web. The fact that scientific language overlaps with general language leads to two way traffic between general language and scientific language: terms of yesteryears are ensconced in general use and scientists continue to put a scientific *spin* on words used in general language. This notion of *renewable* terminology has some relevance for lexicographers involved in creating general-purpose dictionaries, college dictionaries and learners' dictionary.

2. Scientific Writing and Language Change

Halliday has taken a keen interest in scientific writing and he has looked at the writings of Chaucer (c. 11 century), Issac Newton (c. 18 century) and some recent writings in biology and geography (Halliday and Martin 1991). Halliday has noted the novel ways in which scientists deploy, innovate, and create lexicogrammatical resources of their language. These resources include the lexical resources, highly visible new technical terms, and 'grammatical' resources, resources that were used to construct nominal groups and clauses, that formed the basis of reasoned arguments in Chaucer's description of his Astrolabe and Newton's account of the refraction and reflection of light. The visible technical terms, their coinage, currency and obsolescence is discussed in major terminology books (Picht and Draskau 1985, Cabre 1993). Sager, Dungworth and MacDonald's comprehensive account of English as a language of science, deals with lexica and grammar-related issues within a neo-Hallidayian framework, together with some considerable thought on typology of scientific texts (1980:124–311). Thomas Kuhn's paradigmatic views of science were influenced by the works of his Harvard colleagues, including Holton, and has, in turn, influenced the writings of linguists from Halliday to Winograd. Kuhn's neologisms, for example 'normal' science, 'revolutionary' science, and 'crisis in science', appear to be particularly popular in this respect. One might even argue that such a metamorphosis, the transition from private to public science, is

signalled by a burst of neologisms, by a deluge of retronyms, and by a sad loss of terms that were more carefully crafted by human progenitors but at the time of 'revolution' were consigned to the dustbin of archaisms[1]. Indeed, this signal is used by scholars involved in techno-logical trend analysis and examine the distribution and co-occurrence of keywords in scientific publications (see, for instance Courtial 1994).

Science and technology has had a major, if sometimes deleterious effect on our lives in the 20th century. The *1926 Sunday Chronicle New English Dictionary*, clearly shows the effect of the First World War and the influence of chemistry, the dominant science of the time. The dictionary has a 20 page 'new and foreign words' supplement com-prising nominals like *alkane, eclair,* and *escalator* and denominalised verbs like *queueing* and *gas* (as 'in to kill by poison gas'). However, a number of the neologisms in the dictionary are not be found in current dictionaries: Words like *cryptoid* – applied to 'mysterious' phenomena as hypnotism, telepathy and animal magnetism; *cymometer* -a device for measuring the wavelength of electricity; or *shingling* – the process of hammering or pressing iron after puddling – appear not be in use now.

3. Exchanging Terms with Words and Words with Terms

Politicians and religious leaders talk about the role of science in ethics, in morals, and in eliminating poverty and in creating equality; literary critics take pride in their scientific analysis of literature that leads them to talk confidently about the ideological, historical and personal bent of an author. Hence, *scientific terms* are used to describe the human condition ranging from the bizarre to the rational and from the mundane to the sublime. The polemicist, William Safire bemoans the fact that 'you cannot write about world affairs without a *weltanschaung*; you cannot practice literary criticism without an understanding of *zeitgeist*; you cannot nibble your nails properly without *angst*, and you cannot report on turmoil anywhere without *Sturm und Drang*' and to describe a chess match 'you need *zugzwang*' (1990: 415).

The rise of electronics and computing has led to the introduction, not only of a number of neologisms, but also of retronyms, so much so that scientists differentiate how they have 'mailed a message' to their col-leagues, 'through email' or 'snailmail/steam mail' – see Table 1.

Token	Original implied meaning	Retronym (retrospective meaning)
detonation	chemical detonation	nuclear or electromagnetic detonation
guitar	acoustic guitar	electric guitar
intelligence	biological intelligence	mechanical intelligence, artificial intelligence
language	natural language	programming language
mail	paper mail	electronic mail
radio	valve radio	transistor radio
readable	human-readable	machine-readable
speed	speed of physical objects	speed of light; speed of signals
television	terresterial television (*rare*)	cable television; satellite television
watch	wristwatch	pocket watch
	analogue watch	digital watch

Table 1: A selection of retronyms found in newspapers and popular science magazines

The dominance of science and technology in our times does reflect itself in the composition of general language dictionaries, especially the so-called 'college dictionaries' (another retronym?) that include between 25–40%, by number, scientific terms. But this lexical and grammatical traffic, is not one-way only. Kelly-Bootle, who appears to follow in the footsteps of Ambrose Bierce, 'the under-appreciated inventor of cynical lexicography' has been worried about the 'effect' on computing-related writing of words from general language. In his introduction to the first edition of his *Computer Contradictionary*, Kelly-Bootle talks about surviving (!) merrily with "our anthropomorphic *memory*, our medical *bug*, our sexual *random access*, and homely *address*, our gastronomic *chip*, our sportive *jump*, our ornithological *nest*, our narcotic *hash*, our thespian *mask*, our law's *delay*, our daily *queue*, our slum's degradation", and goes on to add 500 new headwords.

3.1 Texts, Text Corpora and Digital Libraries on the Net

The Internet is expanding! The number of Internet users, if it continues to grow at its current exponential rate, will soon be greater than the worlds' populace. Or such is the hype. Nevertheless there is literally every kind of text available on the Internet. Some texts are essentially draft documents, some are written by illiterate people, but most of the texts are

good quality. Scientific texts, especially refereed journals and technical reports and pre-prints of scientific papers put under the authority of respectable organisation, are generally of a good quality and contain up-to-date terminology (see Table 2).

Publisher	Text Type	Total No. of Journals	Coverage
UK Institute of Physics	Research papers, Review papers, Popular Science articles	31 – all issues post 1995 – Back issues soon	Applied Mathematics, Mathematical Physics, Applied Physics, Atomic, Molecular and Optical Physics, Computer Science, Meteorology, Nuclear Physics, Plastics Physics
Association for Computing	Reviews, Surveys, Journals, Research Papers	24	Algorithms, Mobile Networks, Multimedia Systems, Computer-Human Interaction, Graphics, Information Systems, Mathematical Software

Table 2. Journals – issues in print and on the WWW

The professionals, the hobbyists, and the pressure groups also maintain readable and accurate documentation on the Internet. Furthermore, a number of daily newspapers, like the London *Times* and *Telegraph*, have started to produce their replica electronic versions. The electronic versions started their life as abridged editions. The US government is operationalising open-government through the Internet: legislation, draft and effective legislation, government edicts and other documents, including the President's daily schedule, are made available through the Internet.

Carefully crafted text corpora, like the Co-BUILD Direct, a 20 million word corpus together with a frequency ordered list of tokens and concordances that can be ordered electronically, are amongst the number of services offered by Collins. An analysis of nuclear and high-energy physics texts at CERN 'library', shows terms like 'vacuum degeneracy', 'perturbative quantum chromodynamic formalism'; texts at the 'Wine Net News', reveal the existence of 'aromatic lingering finish', 'earthy farmyard nose'; and many interesting new terms at the *Critical Inquiry* site, comprising last 20 years of quarterly issues, where the 'death' of text is discussed quite frequently! A neologism has been spawned to

describe these electronic repositories – the electronic-print (or E-Print) (file) servers:

Coverage	Title/ Organisation	Text Type	WWW Address	Update Frequency
General	The Daily Telegraph	News Reportage & Editorial	http:// www.telegraph. co.uk/	Daily
Physics, Astronomy, Cosomology	Council of European Nuclear Research	Learned Papers, Technical Manuals	http:// preprints.cern.c h/	Daily/ Weekly/ Monthly
Space Science, aeronautics	US National Atmospheric and Space Agency (NASA)	Technical reports, Learned Papers, Press Releases	http:// www.nasa.gov/	Daily and Weekly
Arts and Literary Criticism	Critical Inquiry	Learned Papers, Popular Articles	http:// www.uchicago. edu:80/	Monthly (Back issues available)
Food and Drink	Wine Net News	Popular Texts, News Reportage	http:// http.cs.berkeley. edu/~sethg/ Wine/wine.html	Monthly
Terresterial, Aquatic and Atmospheric Science and Technology	US Environmental Protection Agency's e-print server	Learned Papers, Legislative Texts, Press Release	http:// www.epa.gov/	Updated as per legislation

Table 3. Public domain E-Print Servers

There is also news that entire libraries are being converted for computer-assisted reading. For instance, Bibliothèque de France (Virbel 1993) is a FF7 billion project for building a *digital knowledge infrastructure* eventually to comprise 350,000 'works'; Random House's Modern *Library*, an imprint containing major works of 19th and 20th century literature, is to be rendered electronically on a CD-ROM and the Paper-less Library (Tomer 1993) is to disseminate literary masterpieces free of charge.

3.2 Terminology on the Net

The specialist communities are not only erecting full text repositories but some are also in the process of creating terminology data banks. Table 4 shows a list of some of these repositories which are currently being made available on the Internet:

Subject	Repository	World Wide Web Address
General Purpose	Term Bazaar	http://www.surrey.ac.uk/MCS/AI/pointer/bazaar.html
	Eurodicautom	http://www.uni-frankfurt.de/~felix/eurodictautom.html
	The Devil's Dictionary	http://www.cs.uit.no/~frankrl/Devil/dd_.htm
	Dictionary of Roadie Slang	http://searider.jpl.nasa.gov/~gms/text/slang.html
Techno-logy	NASA Terminology Collection	http://www.sti.nasa.gov/nasa-thesaurus.html
	Free On-Line Dictionary of Computing	http://wombat.doc.ic.ac.uk/
	Software Engineering Glossary	http://dxsting.cern.ch/sting/glossary-intro.html
Leisure	Dan's Poker Dictionary	http://www.universe.digex.net/~kimberg/pokerdict.html
Food and Drink	Whisky Glossary	http://www.dcs.ed.ac.uk/staff/jhb/whisky/glossary.html
Flora and Fauna	Vascular Plants glossary	http://155.187.10.12/glossary/glossary.html
	Aquarium glossary	http://www.actwin.com/fish/glossary.html
Com-merce	Real Estate and Mortgage Glossary	http://www.homebuyer.com/realestate/common.dir/glossary.html
	Credit, Financial and Legal Glossary	http://www.teleport.com/~richh/glossary.html

Table 4: A selection of the specialist term repositories on the Internet

Inspired by this burgeoning text and terminology 'on the net', we have developed Term Bazaar (which can be found on the Internet at http://www.surrey.ac.uk/MCS/AI/pointer/bazaar.html), a program for man-aging terminology either at a given Internet site or across a number of Internet sites and for accessing texts from Internet sites based on a selection key-words either provided by a given user or keywords derived

from term bases. The system is currently being interfaced to a text and terminology system, System Quirk for text management (Ahmad and Holmes-Higgin 1995).

The Term Bazaar aims to facilitate the exchange of terminology across the World Wide Web. Currently the program provides access to over 12 term bases, and uses hypertext mark-up and relational data bases for storing and retrieving terminology. The system has been operational since March 1995 (see Ahmad and Collingham, 1995).

The availability of texts and terminology on the Internet is important for two reasons. First, the practical reason that such repositories will make access to and modification of the terms easier and faster; the availability of texts has its obvious utilitarian benefits, but such text repositories can be used for extracting terms, elaborating terms and for determining pragmatic and syntactic properties of a given term. Second, the terms in the repositories are in some sense peer-reviewed and approved and can be used as a basis for searching (peer-reviewed and authoritative) documents produced in a given specialist domain. Once a relevant document is found, and if it has neologisms or retronyms, then these terms can be extracted and incorporated in the relevant term base.

4. Renewing Terminology: Crawling the Internet for Texts and Terms

We have been involved in the development of a system that will select a set of terms from, say, a popular science text and then search the Internet for texts containing some of the key terms in the text. Once texts containing the key terms are found, then these would be analysed and some of the *new* key terms will be noted. These key terms will then be used for searching the Internet again, more text will be collected, new key terms noted and the search continues again. The 'newly' identified key terms may be checked against an existing terminology data base, if one is to be found. If the key terms are already in the data base then the system may be instructed to replace these terms with another set if possible.

This search-and-bring-back *(SEBRIBA)* strategy can be further refined by grounding the choice of key terms in a frequency-based criteria: for instance, top ten most frequent terms may be chosen, together with the five to ten least frequent terms, to start a search and to sustain it.

In an experiment to determine the efficacy the SEBRIBA strategy, we have conducted a four day 'vigil' of the Internet for terms in automotive engineering, specially in the area of low-emission cars. We began by analysing a popular science article on how to buy a catalytic converter-

fitted car. This 3,307 word article was analysed and a frequency ordered list of single words was produced and the 7 most frequent words in the text, namely *car, converter, engine, emission, limit,* and *catalyst,* together with 3 least frequent words, that is *conversion, pollute,* and *inhaling* were chosen. By using an Internet search engine, a program for searching texts available on the World Wide Web, we found 12 texts of which 5 appeared to be of much relevance. The five texts include two technical reports, popular science texts, like the so-called texts comprising *frequently asked questions* about autos/gasoline, and a composite text comprising of 127 abstracts from publications of the Dutch Centre for Energy Research (ECN – Energieonderzoek Centrum Nederlands). The five texts contain over 70,900 words of specialist language.

The 70,900 word corpus was analysed and a list of terms was extracted by analysing the texts. This list comprises: *fuel, methanol, engine, vehicle, vehicle, transportation,* and *emission.* Five least frequent terms in the corpus comprises: *logo, quantification, conservatism, castro,* and hydro. Note that some of the key terms above have also appeared in the previous list; all we can say is that terminology behaves quite erratically and requires an autonomous status. The 'new' terms were then 'sent' on a *SEBRIBA* strategy. Our system dutifully returned with 31 texts and we selected five new texts comprising 25,054 words in all. This trawl also included a *Transportation Glossary.* Amongst the new terms we found we had terms like *brake specific fuel consumption, cold start emission control* and so forth.

Currently, we are developing an integrated system that will comprise System Quirk, for text analysis and terminology access, Term Bazaar, for accessing remote terminology databases, and a new search program, written in *Java,* a programming language for distributed applications. The search program will be able to browse the Internet and retrieve texts.

5. Conclusions

The demands for validated multilingual terminology are expected to grow quite dramatically within the 15 nation Europe and across the world. Ready access to terminology can be made possible by providing terminology data bases on, say, the World Wide Web, and by providing access to the burgeoning specialist text corpora on the World Wide Web. Such a symbiotic approach will help in identifying neologisms, retronyms and archaisms. The renewal of terminology, as described above, can be used as a basis for studying language change and growth on the

one hand and for studying the progress in science and technology on the other.

Notes

1. Even if one had doubts about the meta-theories of the Harvard school and one adopts a neo-positivist view of growth and change in science, that is like Popper (1981:119) argues about an iterative refinement of scientific thought: One can still posit the neologisms, retronyms and archaisms, throughout the various stages of the Popperian cycle:*Problem P₁ -> Tentative Theory (TT) -> Error Elimination (EE) -> Problem P₂*

References

Ahmad, K., S. Collingham 1995. *Aspects of Terminology Infrastructure in Europe: Volume 4 – Terminology Infrastructure and the Information Superhighway.* (Part of the EU project - POINTER).

Ahmad, K., P. Holmes-Higgin 1995. "System Quirk: a unified approach to text and terminology", in H. Picht and G. Budin (eds), *TAMA '94 Proceedings, Third TermNet Symposium, Terminology in Advanced Microcomputer Applications – Recent advances and User Reports.* Vienna, TermNet, pp 181–194.

Bierce, A. 1911.*The Devil's Dictionary.* New York, Neale Pub. Company.

Cabre, M. T. 1993. *La terminologia: Teoria, metodologia, aplicaciones.* Barcelona (Spain), Editorial Antartida/ Empuries.

Courtial, J. P. 1994. "A Coword Analysis of Scientometrics". *Scientometrics.* Volume 31 (No.3) pp 251–260.

Geertz, C. 1988. *Works and Lives: The Anthropologist as Author.* Stanford (Calif., USA), Stanford University Press (Also distributed by Blackwell Publishers, Oxford, UK).

Heisenberg, W 1963. *Physics and Philosophy: The Revolution in Modern Science.* London, George Allen & Unwin Ltd.

Heisenberg, W 1983. *Encounters with Einstein (& other Essays on People, Places and Particles).* Princeton (New Jersey, USA), Princeton University Press.

Heribert, P., J. Draskau 1985. *An Introduction to Terminology.* Guildford (Surrey, UK), University of Surrey.

Holton, G. 1973. *Thematic Origins of Scientific Thought: Kepler to Einstein.* Cambridge (Mass., USA), Harvard University Press.

Kelly-Bootle, S. 1995. *The Computer Contradictionary* (2nd Edition). Cambridge (Mass., USA) & London (UK), The MIT Press.

Kuhn, T. S. 1970. *The Structure of Scientific Revolutions* (2nd Edition). Chicago (Ill., USA) and London, Chicago University Press.

Landon, G. P., P. Delany 1993. (Eds) *The Digital World: Text-Based computing in the Humanities*. Cambridge (Mass, USA) and London, The MIT Press.

Popper, K. R. 1981. *Objective Knowledge: An Evolutionary Approach.* (Revised Edition). Oxford (England), Oxford University Press.

Safire, W. 1990. *Language Maven Strikes Again*. New York (USA), The Cobbett Corporation.

Sager, J. C., D. Dungworth, P. F. McDonald 1980. *English Special Languages: Principles and practice in science and technology.* Wiesbaden (Germany), Oscar Brandstetter Verlag KG.

Tomer, C. 1993. "Emerging Electronic Library Services and the Idea of Location Independence". In (Eds) George P Landon and Paul Delany. pp139–161.

Virbel, J. 1993. "Reading and Managing Texts on the Bibliothèque de France Station". In George P Landon and Paul Delany (eds), pp31–51.

Whitehead, A. N. 1948. *Essays in Science and Philosophy.* London, Rider and Company.