# TERMINOLOGY MANAGEMENT AND KNOWLEDGE PROCESSING IN LESSER-USED LANGUAGES

*Khurshid AHMAD*

AI Group. Department of Computing
School of Electronic Engineering, Information Technology and Mathematics
University of Surrey

## 1. Introduction

Terminology management, that is, the collection, analysis, validation and distribution of terms, is essential for specialist writing, translation, information retrieval, and knowledge dissemination. Specialists of all persuasions are involved in coining terms, modifying existing terminology, rendering terms archaic or re-introducing discarded terms with new meanings. It appears that the specialists perform many terminology-related tasks: they coin neologisms, introduce retronyms, translate terms, validate terms and, in a rather indirect manner, compile or help to compile terminology collections of their specialisms. Modern science, contemporary leisure and entertainment, innovative enterprises, all distinguish themselves from their older incarnations not merely through new theories and artefacts, through enhanced and accessible pleasures, through advanced goods and services, but also through the terminology they use to describe their subjects.

Terminology management is by and large a manual task that relies on the existence of well-motivated documentalists, translators and terminologists, the latter sometimes performing the function of the former two. Currently available terminology management (computer-based) systems have alleviated some of the storage and retrieval tasks associated with terminology management. The tasks of collection, analysis and validation are still, however, by and large, executed manually. The argument we would like to present here is that some aspects of terminology collection, of analysis and of validation, especially the routine and time-consuming aspects, can be automated through the use of computer systems. This automation will, we believe, reduce the costs of these hitherto rather expensive tasks.

In languages used by numerical majorities, the expensive task of terminology management is underwritten by the expectation that there is a potentially large number of people who require terminology data bases and are willing to invest in them. For languages used by a numerically smaller number of people this is not the case; terminology management here is often linked with the politically-motivated, and often emotionally charged, work of language planning. The dependence on fewer human beings for terminology management is greater in the lesser-used language communities than may be the case for other languages.

Computer-based terminology management is not merely a task of writing computer programs, although such an undertaking is onerous in itself. Such an undertaking requires an understanding of how specialist text is written, how human beings deal with semantics of specialist domains, how discourse patterns change according to the needs of the authors and readers of the texts. Specialist writing is sometimes used synonymously with the term 'technical writing'. Technical writing is occasionally used pejoratively, to denote a discourse pattern which is associated with (highly) skilled work like routing aircraft and repairing cars, for instance. Technical writing is one manifestation of specialist writing: for us, all writing related to any specialist enterprise is specialist. This includes, for example, science, technology, entertainment and leisure, culture and religion, administration and politics.

Over the last ten years we have been building terminology collections in languages used by numerically larger groups of people, like English, German and Spanish, whilst at the same time attempting to adapt such methods for lesser used languages like Welsh, Norwegian, Flemish and Catalan. This paper will discuss some of the challenges encountered, opportunities identified and solutions suggested for managing the terminology of specialist languages in multilingual environments where at least one language belongs to the lesser used category on numerical grounds. Our theoretical framework draws on recent work in corpus linguistics, the philosophy and history of science and computing sciences.

## 2. Teaching and learning in a second language

English has a dominant position as a scientific and technological language: the achievements of English-speaking scientists and engineers has persuaded many to learn this language in order to understand and apply the which they articulate. The proliferation of English as a Foreign Language, English as a Second Language enterprises in the UK, USA and Australia is sufficient proof of the achievements of these scientists. The spread of English can be viewed as the successful persuasion, on the part of the English-speaking communities, to convince others to speak their language. (Colonisation by English-speaking nations has also contributed to the spread of English.)

For many, whose first language is not English, the teaching and learning of science and technology is commonly in their second language, usually English. However, non-native speakers of English have difficulties in using the language as efficiently as say native speakers might. This is especially true of the school-children learning the methods and techniques of science during their pre-teens. The school-child has to deal not only with the complexities of abstract concepts and novel artefacts, but also has to understand a language which might be typologically different, for example, Arabic native speakers learning physics in English, or Welsh native speakers learning radiochemistry in English. The pedagogic overheads of teaching and learning science and technology in a second (or in some cases third language, for instance, the former non-Russian Warsaw Pact nations had Russian as a second language and English as the third language) are considerable.

The other overhead of learning science and technology in a second language is less objectively described. This overhead relates to the relative prestige of the first language, a language which might be used to buy and sell food and clothing, a language in which all legends are described, a language of kings and queens of centuries past. However, the potency of this symbol of *nationhood* is reduced when the lesser-used language cannot cope with ‘today’, that is the today of computers and satellite communication, the today of nuclear reactors and space walks. The lower prestige of such languages is then used by the prejudiced to persuade themselves that the solution to efficient specialist communication really lies entirely in the second language. Consequently, the first language need not be learnt in any reasonable depth since it does not provide the linguistic means for dealing with today.

The twin overheads of pedagogy-related and prestige-related problems that may impede the growth of a language, can be overcome by encouraging the scientists and the technologists of the language community to write about their disciplines in their first language. Of course, the precursor of such an exercise is the existence of terminology: one might use second-language terminology embedded in a first-language text, but then the overheads, though partially minimised, still exist.

The three minority languages studied in this paper are all Indo-European languages and spoken by numerical minorities of less than 15 million. These languages are used in the teaching of science and technology in schools and sometimes at universities. Two of the languages are from the Germanic family: the North Germanic *Norwegian*, spoken by over 4 million people; and the West Germanic *Dutch* spoken by well over 14 million people. The third language, *Welsh*, belongs to the Celtic language family and is spoken and used largely in North Wales; the total population of Wales is just over 2.7 million. There is a substantial literature in science and technology in lesser-used languages like Dutch or Norwegian, and, to a lesser extent, in Welsh; and terminology has been compiled in a range of subjects[1]. However, the compilation of terminology collections is by no means cheap: each scientific discipline requires between 2,500-10,000 terms (Alford 1971) and we have estimated that the cost of acquiring and elaborating each term is about 10 ECU.

Corpus-based methods can be of help here. Scientific texts have a distinct texture when compared with general-language text: there is a profusion of nominals in scientific texts, relatively more passives, more plurals and fewer pronouns when compared with newspaper text, works of fiction, leisure magazine texts and so on. Furthermore, scientific and technical texts contain neologisms. Therefore, the comparison of the frequency-ordered list of words in a scientific text with, say, a list of words from a representative corpus of general-language texts, can be used find neologisms and other terms. An alphabetically-ordered concordance of scientific texts will reveal the grammatical environment of terms. The use of specialist text corpora, (more specialist texts are becoming available on the Internet) together with computer programs for organising and analysing such texts, will reduce the labour outlay on collection terminology. The behaviour of terms can also be studied in a given text, a collection of texts, in texts differentiated on the basis of their genre, origins and so forth.

## 3. Text corpora and the study of written language

### *3.1. Corpus Linguistics: Some definitions*

A text corpus maybe defined as a collection of ‘linguistic data which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language’ (Crystal 1997:95). The resurgence in empirical and in language statistical methods of language study owes much to the tenacity of many scholars including Quirk, Greenbaum, Leech, Sinclair and Svartvik on this side of the Atlantic, and Francis, Kucera and Biber, amongst others, on the other side of the Atlantic. About 30 years ago Quirk suggested that 'the basis [for writing a grammar of English] must be copious materials, made up of continuous stretches of “texts” taken from the full range of co-existing varieties and strata of educated English...' (1968: 78-9). In 1985, Quirk et al published *A Comprehensive Grammar of the English Language*, comprising examples from the University College (London) based ‘Survey of [British] English Usage’, the Brown (University) Corpus of American English and the Lancaster Oslo/Bergen Corpus. The authors of the *Grammar* noted that 'assessment by native speakers of relative acceptability largely correlate with their assessments of relative frequency' (1985: 33). In the same year

[1] Consider, for example, the terminologies produced by the Welsh Joint Education Committee, Wales (UK) which include terminologies of computing, biology, chemistry, etc.

John Sinclair, using Birmingham University's 'Bank of English', a computer-based corpus, announced the *very first* dictionary which 'has been compiled by the thorough examination of a representative group of English texts' (1985: xv). It appears to us that what Quirk *et al* and Sinclair have said about general language, specifically the ways in which frequency of occurrence of a word is correlated with the acceptability of the word, sometimes also called the 'evidence' and the 'reference' for a word, applies more strongly to terms and their behaviour in specialist texts.

*3.2 .Closed and open class words*

Frequency studies of texts, dating back to Biblical concordances, show clearly that a given text comprises more of the words that are either prepositions, pronouns, determiners, conjunctions, modal verbs or primary verbs. The less frequently encountered words can be categorised as nouns, adjectives, full verbs and some adverbs. Thus there are two major classes of words: the closed classes, that is the prepositions and so on, and the open classes, that is nouns and so forth (See Quirk *et al* 1985:67-78 for a detailed discussion).

The closed class words are 'closed' in the sense that very seldom does one see new words added to the class and similarly it is rare to see words belonging to these classes being declared archaic. The stock of the open class words, on the other hand, is constantly being renewed, principally through neologisms, nominalisation of verbs, and the denominalisation of nouns and through a whole range of other devices available to a linguistic community.

The predominance of the closed class words is quite remarkable both in general language texts and in special language texts. In English language texts the first ten most encountered words, i.e. *the, at, on, of, in, be, ....* make up over 20% of most texts. Indeed, a look at a modern-day corpus of English will show that if a frequency-ordered list of words in, say, a 10 million word corpus like the Longman/Lancaster English Language Corpus[2] (Summers 1991), is compiled, the list has over 100,000 different lemmas (not differentiating the morphological variants). However, it is the first hundred words, the first *percentile*, of the frequency-ordered list that comprises just under half the total text (see Table 1a). Note that there are virtually no open class words in this percentile - with the exception of some commonly-used adjectives like *little*, or the nouns such as *man*, and possibly verbs, such as *time* - the percentile comprises determiners, pronouns, conjunctions and modal and primary verbs!

Table 1a. The 1st percentile of a frequency-ordered list of words in the Longman/Lancaster English Language Corpus

| Rank | Words | Cum. Rel.Freq. |
|---|---|---|
| 1-10 | the, of, and, to, a, in, it, that, i, was | 23.24% |
| 11-20 | he, is, for, as, with, his, on, you, had, be | 7.65% |
| 21-30 | not, she, they, her, by, this, from, or, have, are | 4.83% |
| 31-40 | which, we, all, were, an, one, there, said, him, so | 3.57% |
| 41-50 | what, would, their, when, if, no, my, been, out, up | 2.62% |
| 51-60 | them, more, about, can, me, who, like, into, has, then, | 2.20% |
| 61-70 | could, do, will, ***time***, only, some, other, its, than, now | 1.79% |
| 71-80 | two, very, these, over, any, did, down, way, back, first | 1.39% |
| 81-90 | ***man***, know, just, see, may, our, how, even, well, your | 1.15% |
| 91-100 | such, where, because, after, much, made, before, little, most, through | 1.06% |
| | **TOTAL** | **49.51%** |

[2] The Longman/Lancaster English Language Corpus was compiled by the publishers of Longman English dictionaries. The publishers were inspired and guided by Lord Randolph Quirk and Professor Geoffrey Leech. The 28 million word corpus comprises texts one half of which were selected randomly and the other half selected by a panel of experts. The corpus comprises texts from novels, textbooks, newspapers and magazines, including women's and popular science magazines. We have access to only 10.29 million words comprising selections from both halves.

The evidence provided by the second percentile is similar, except that some of the open-class words have crept in the lower end of this second percentile (see Table 1b, we have italicised the nouns):

Table 1b. The 2nd percentile of a frequency-ordered list of words in the Longman/Lancaster English Language Corpus

| Rank | Words | Cum. Rel.Freq. |
|---|---|---|
| 101-110 | before, little, most, through, don't, must, go, us, get, good | 0.94% |
| 111-120 | too, ***people***, should, new, also, here, own, between, never, still | 0.87% |
| 121-130 | come, think, thought, again, those, say, long, off, re, right | 0.79% |
| 131-140 | make, old, ***wo****rk*, being, ***life***, many, same, day, got, came | 0.74% |
| 141-150 | ***years***, another, going, each, went, might, all, great, away, three | 0.67% |
| 151-160 | take, while, something, ***world***, both, *men*, himself, though, always, under | 0.61% |
| 161-170 | 've, things, ***house***, without, look, last, once, put, used, ***Mr*** | 0.56% |
| 171-180 | ***place***, looked, nothing, didn't, why, left, found, ***women***, ***part***, ***mothe****r* | 0.53% |
| 181-190 | every, does, ***hand***, want, ***thing***, *face*, end, few, *eyes*, against | 0.50% |
| 191-200 | ***head***, course, ***father, room, water***, asked, far, since, small, told | 0.48% |
| | **TOTAL** | **6.70%** |

The texts in the Longman/Lancaster Corpus include newspaper text, magazine text, excerpts from fiction and non-fiction work, popular science articles and so forth, and one can argue that these texts are texts of everyday usage: in other words Longman's is a general language corpus.

*3.3 .Closed and open class words in Dutch, Norwegian and Welsh*

A percentile-wise study of tokens in Dutch, Norwegian, and Welsh shows roughly the same pattern as the Longman Corpus. The Eindhoven Corpus of Dutch, comprising general language texts including texts from newspapers, magazines, magazines, works of fiction and popular science texts, contains a total of over 605,000 words (Table 1c):

**Table 1c. The distribution of the first percentile words in the Eindhoven corpus of the modern Dutch language**

| Rank | Words | Cumulative Rel. Freq. |
|---|---|---|
| 1-10 | de, van, het, en, een, in, dat, die, is, ik | 21.36% |
| 11-20 | te, op, niet, zijn, met, maar, voor, hij, je, n | 7.96% |
| 21-30 | dan, als, ze, aan, er, was, ook, t, zo, ja | 5.25% |
| 31-40 | nog, om, bij, door, wat, of, naar, wel, uit, heeft | 3.53% |
| 41-50 | over, hebben, we, zich, had, tot, worden, haar, deze, dit | 2.65% |
| 51-60 | al, nou, meer, geen, daar, cdb, wordt, toen, zou, dus | 2.18% |
| 61-70 | kan, nu, d, zij, u, toch, s, hem, hun, moet | 1.86% |
| 71-80 | r, heb, a, weer, werd, kunnen, men, zal, want, veel | 1.63% |
| 81-90 | me, mijn, waar, cobl, hier, jaar, wij, andere, tegen, eens | 1.37% |
| 91-100 | goed, waren, alleen, mensen, twee, gaan, na, m, ons, zei | 1.15% |
| | **TOTAL** | **48.93%** |

The Surrey/Bergen Corpus of Norwegian general language was created during an hour-long 'trawl' of the World-Wide Web; the 'trawl' resulted in the collation of 50 texts, including newspaper text, magazines and public information documents, comprising 107,304 words (Table 1d):

Table 1d. The 1st percentile of the frequency-ordered list of tokens in the Surrey/Bergen Norwegian Corpus

| Rank | Words | Cumulative Rel. Freq. |
|---|---|---|
| 1-10 | og, i, av, for, som, det, er, til, på, en | 22.12% |
| 11-20 | å, med, at, de, har, den, et, om, kan, skal | 8.82% |
| 21-30 | ikke, vil, eller, dette, fra, ved, it, denne, mellom, være | 4.00% |
| 31-40 | også, seg, andre, men, artikkel, var, nye, alle, både, vi | 2.18% |
| 41-50 | bruk, mer, etter, slik, man, *informasjon*, blir, må, ulike, tjenester | 1.68% |
| 51-60 | over, disse, gjennom, utviklingen, utvikling, ble, felles, kunne, innen, nett | 1.48% |
| 61-70 | mange, forhold, så, eøs, når, *europeiske*, norge, mot, der, tiltak | 1.29% |
| 71-80 | bør, del, viktig, løsninger, du, under, flere, norsk, ut, offentlig | 1.14% |
| 81-90 | området, innenfor, bl.a., ha, samme, enn, gir, få, avtale, opp | 1.08% |
| 91-100 | vært, *elektronisk*, offentlige, gjelder, selv, sin, nå, hvor, staatens, store | 1.03% |
| | **TOTAL** | 44.81% |

The design of the Surrey Welsh corpora of special- and general-language texts is similar to that of various corpora of English outlined in Aijmer and Altenberg (1991:315-318). Our coverage is not exhaustive, the selection of text is partly intuitive and we have only included extracts of newspapers, books and magazines (see Ahmad and Davies 1994) (Table 1e). Despite the presence of nouns like *ysgol* (school) in the first percentile, most of the words in the Welsh, like the English, percentile are closed class. Some of the most frequent ones are *particles* (like *yn*, *a*, *'n*) and a number of them have at least two different meanings: *yn* is not only used as a particle but also as *in* and *into*; *i* is used as personal pronoun *I*, *he* or *she* and also as an interrogative marker.

Table 1e. The Surrey Welsh general language corpus and the behaviour of the first percentile

| Rank | Words | Cumulative Rel. Freq. |
|---|---|---|
| 1-10 | yn, y, i, a, r, o, n, yr, ei, ar | 27.60% |
| 11-20 | ac, oedd, mae, am, ond, wedi, un, fod, fel, na | 6.71% |
| 21-30 | ni, mewn, eu, l, gan, bod, wrth, hyn, hi, yw | 3.57% |
| 31-40 | nid, chi, hynny, neu, fe, ddim, w, at, mi, os | 2.53% |
| 41-50 | hen, dim, iawn, er, yma, beth, lle, gyda, sy, roedd | 1.86% |
| 51-60 | fy, pan, hyd, sydd, cael, yng, bob, ti, arall, mai | 1.64% |
| 61-70 | nad, rhaid, eich, ef, hefyd, ym, mynd, ma, heb, ein | 1.49% |
| 71-80 | dy, cyn, oes, rhyw, nhw, u, iaith, hwn, mawr, gael | 1.34% |
| 81-90 | d, hun, dyna, iddo, byd, peth, di, mor, ffordd, eto | 1.22% |
| 91-100 | ysgol, wlad, rhai, llawer, drwy, bydd, erbyn, hwnnw, ag, mwy | 1.10% |
| | **TOTAL** | **49.06%** |

The tables 1a and 1c-1e show that closed class words tend to be the most frequent in the four languages that were observed. However, we show that when one looks at specialist text (corpora) in any of the four languages, the specialist terms usually occur in the 20 most frequent single words. The terms are invariably included in the first percentile.

A note of caution: the compilation of a general corpus is a complex, expensive and time-consuming task. One has to compile a corpus comprising no less than 10-100 million words. The questions of genre and variety also complicate the issue. In this context our very small corpora of Welsh and Norwegian is not an accurate representation of these languages. However, we are not using the general language corpora to extract lexical data, rather we would use the general language texts to contrast these texts with specialist texts. We are currently building a Welsh and Norwegian corpora with a target of 10+ million words each.

## 4. The 'texture' of scientific texts

There has been substantial discussion in the literature on scientific writing. This discussion has focused on a range of text types. Halliday and Martin (1993) have looked at scientific textbooks for schools; early researchers involved in corpus linguistics used to include popular science texts, under the rubric of informative texts, in their corpora of British English (the Lancaster/Oslo-Bergen Corpus, Johnsson and Hofland 1989) and of American English (the Brown Corpus, Kucera and Francis 1967); Jan Svartvik has studied questions related to voice in English with reference to a corpus of learned papers (1966).

Halliday and Martin have argued that in order to construct 'scientific reality' the author/scientist harnesses his or her vocabulary and grammatical structures to record scientific observations, to produce critiques of ideas and theories, and so on. The harnessing of the lexicogrammar produces a variant of the natural language of the author/scientist. For example, an English-speaking physicist will write English texts but in a rather terse and formal manner which would at times appear quite opaque: the so-called language of physics. The terminology of physics is dominated by single and compound nominals, and uses frequent nominalisation of verbs; the same will be true of a Chinese-speaking physicist, that is, the text produced will be in Chinese, but the author will use certain aspects of Chinese vocabulary and grammar in a preferential fashion and will suppress other aspects like pronominals. Svartvik (1966) has noticed that the scientist-author uses significantly larger numbers of non- agentive passives in specialist texts than used in speech or in general writing.

Specialist writing has its own distinct texture. The avoidance of ambiguity is essential in specialist discourses; monosemy is at a premium here and naming conventions, formal or informal, are devised by specialist communities for talking about the ideas and objects of their specialist domains. There are, as we shall show in a more quantitative way presently, a profusion of *nouns* in specialist texts; collocation patterns are also more readily identifiable if only due to their higher probability as compared to general language texts. Another notable linguistic characteristic of specialist texts is the restricted syntactic patterns used; there are two sentence types, *imperative* and *declarative*, that account for much of the sentence types used in the specialist texts.

Perhaps the most quantitative difference between specialist texts and general language texts is in the frequency distribution of the different vocabulary items. Consider the frequency distribution of the first 25 words in Longman/Lancaster Corpus and compare this distribution with the first 25 words in three specialisms - nuclear physics (a physical science), automotive engineering (an engineering discipline), and dance analysis (a performing arts subject). We have compiled a corpus of each of these disciplines at Surrey and we have computed the frequency distribution of all the words in each of these corpora. See Table 2a below:

In the above table, 'RF' stands for relative frequency and is computed by dividing the absolute frequency of a word by the total number of words in the corpus. The three specialist corpora chosen are of different sizes in terms of the total number of words they contain and also in terms of their *pragmatic balance*: the Automotive Engineering corpus has the six different genres usually associated with specialist texts, whilst the Dance Analysis corpus consists largely of news reportage (nine texts) and features, essentially press reviews (33 texts). The nuclear physics corpus contains journal texts and book excerpts only. Furthermore, the automotive engineering and dance analysis corpora comprise British and American English texts, whereas the nuclear physics corpus contains only British English texts (see Table 2b):

Table 2b. Composition of the three corpora compiled at the University of Surrey (cf. Table 3a.)

| *Genre* | *Automotive Engineering* 369,751 words | *Dance Analysis* 44,607 words | *Theoretical Nuclear Physics* 81,946 words |
|---|---|---|---|
| | No. of texts | No. of texts | No. of texts |
| Popular Science | 6 | | 0 |
| Journals | 50 | 1 | 15 |
| Manuals | 7 | 0 | 0 |
| Books | 5 | 3 | 13 |
| Advertisements | 38 | 0 | 0 |
| News Reports/ Features | 24 | 42 | 0 |
| **Total number of texts** | **130** | **46** | **28** |

The difference in genres and language variety does not appear to stop the profusion of nouns in all these three specialist corpora. The six most frequent words in all four corpora are the same closed-class words, comprising over 15% of the total words in each corpus. The first ten words in all the corpora are still closed-class words and comprise just under 25% of the total words of each corpus

*4.1. Weirdness of scientific texts*

The preponderance of nouns in a specialist text, and indeed in a specialist text corpus comprising such texts, is one of the characteristic textures of specialist texts. Such preponderance can be demonstrated by computing the distribution of the closed and open class words in a corpus of texts related to *Radiation Physics* (developed at Surrey) and comparing this distribution with that of the same class of words in general language. The radiation physics texts (55 texts written in English, in all comprising a total of over 85,000 words) were collected through the World-Wide Web using various search engines. The radiation physics texts include learned papers, advertisements for conferences and courses, popular science texts in radiation physics and radiotherapy.

The results of the comparison between the two, Radiation Physics (specialism), and general language (Longman Corpus), are shown in Table 4a. In order to save space, we show the aggregated frequencies of batches of ten words. Note that there is only **<u>one</u>** noun amongst the 100 most frequent words in everyday language (Table 3a, column 1) and even that is found in the lower frequency regions. However, the nouns make up around 40% of the 100 most frequent words in the specialist corpus: 100 most frequent words make up just over 40% of all the words found in the two corpora. (The third and fifth columns of Table 3a contain the values of relative frequency which is equal to the absolute frequency divided by the total number of words or tokens).

Table 3a. The 1st percentile of a frequency-ordered list of words in the Longman Corpus compared with Surrey's Radiation Physics Corpus

| Rank | Longman/Lancaster Corpus 10 million tokens | Rel. Freq. (%) | Radiation Physics Corpus 85,109 tokens | Rel. Freq. (%) |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| 1-10 | the, of, and, to, a, in, it, that, i, was | 22.54 | the, of, and, in, to, a, for, is, are, with | 22.41 |
| 11-20 | he, is, for, as, with, his, on, you, had, be | 7.91 | be, on, nuclear, radiation, from, by, at, as, data, that | 5.56 |
| 21-30 | not, she, they, her, by, this, from, or, have, are | 4.99 | this, energy, dose, mev, or, an, which, neutron, cross, it | 3.45 |
| 31-40 | which, we, all, were, an, one, there, said, him, so | 3.69 | protection, image, was, have, beam, used, these, measurements, we, can | 2.35 |
| 41-50 | what, would, their, when, if, no, my, been, out, up | 2.71 | will, been, research, sections, has, al., et, electron, also, beams | 1.85 |
| 51-60 | them, more, about, can, me, who, like, into, has, then, | 2.27 | physics, were, not, may, other, power, new, more, results, absorbed | 1.64 |
| 61-70 | could, do, will, time, only, some, other, its, than, now | 1.82 | system, there, measurement, dosimetry, use, reactor, university, high, imaging, treatment | 1.46 |
| 71-80 | two, very, these, over, any, did, down, way, back, first | 1.40 | such, than, our, technology, total, well, using, all, figure, accelerator | 1.30 |
| 81-90 | ***man***, know, just, see, may, our, how, even, well, your | 1.18 | about, clinical, section, medical, gamma, medicine, science, some, ray, fission | 1.19 |
| 91-100 | such, where, because, after, much, made, before, little, most, through | 1.08 | radiotherapy, up, but, conference, exposure, calculations, electrons, studies, low, one | 1.08 |
| | **TOTAL** | **49.59** | TOTAL | **42.28** |

Amongst the frequent nouns in the 100 most frequent words in the Radiation Physics corpus are terms like *energy*, *neutron*, *dosimetry*, *image*, *beams* and abbreviations like *MeV* (which stands for *M*illion *e*lectron *V*olts). A more detailed analysis has shown that although current general language does include terms that once were the preserve of atomic and nuclear scientists, e.g. *electron*, *radiation*, *fission*, *gamma rays*, these terms are used with much greater frequency in the specialist Radiation Physics writings: for example, *electron* is used 100 times more frequently in above sample of writing of radiation physicists; words like neutron and reactor occur over 1000 times more frequently. There are terms that are used in the above corpus that are yet to reach the general language.

Specialist text corpora, comprising learned scientific texts, have been created to study the grammatical, semantic and pragmatic behaviour of terms. This is what we will discuss next.

*4.2 .A weirdness-informed method for extracting terminology*

We have noticed quantitative differences between the vocabulary used in specialist texts and that used in general language texts. This observation appears to hold across a genre of texts within a specialism, for example, learned texts, technical manuals, legislative documents, advertisements and textbooks: each of these genres shows a disproportionate number of nominals, and the authors of these texts use fewer closed class words when compared with a representative general language corpus of English texts (the Longman Corpus of Contemporary English). This observation also holds across a range of languages specifically French, Catalan, Spanish and Welsh. The scientist-authors uses a language which is *weird*, to use a term coined Bronislaw Malinowski in the context of the language used by South Sea Island magicians. Indeed, this weirdness in vocabulary usage can be used as a basis for extracting terminology in the following sense: the ratio of the *relative* frequency of a word in a specialist corpus (or text) with the *relative* frequency of the same word in a general language corpus can provide a measure of 'weirdness'. For instance, if we compare the relative frequency of the first six open-class words from the 25 most frequent words in the Surrey Automotive Engineering corpus with their relative frequency in the Longman-Lancaster Corpus, we find that the ratio of relative frequencies is some guide to the quantitative differences between special-language texts and general-language texts. The very high values of this ratio for some words indicates the possibility that these words are used exclusively, say, for automotive engineering. Terms such as *autocatalyst*, and *hydrocarbon*, have zero frequency in the Longman-Lancaster Corpus, but a finite frequency in the automotive engineering corpus; hence, the *weirdness* for these terms,

when computed by comparison against the Longman-Lancaster corpus, is *infinity*. (It is interesting to note that the general language texts have the plural *hydrocarbons* but not the singular *hydrocarbon*). Table 3b shows the weirdness of a few single word terms in the Surrey Automotive Engineering Corpus.

**Table 3b. The preponderance of open-class words in special-language literature (Figures in columns *b* and *d* have been rounded up)**

| *Word* | *Surrey Automotive Engineering Corpus (369,751)* | | *Longman-Lancaster Corpus* (10,299,924) | | *Co-efficient of Weirdness* |
|---|---|---|---|---|---|
| | *Absolute Freq.* | *Relative Freq. (%)* | *Absolute Freq.* | *Relative Freq. (%)* | |
| | *(a)* | *(b)* | *(c)* | *(d)* | *(b/d)* |
| system | 1,795 | 0.4855% | 4197 | 0.00040748% | 12 |
| control | 1,517 | 0.4103% | 2152 | 0.00020893% | 20 |
| car | 1,790 | 0.4841% | 1903 | 0.00018476% | 26 |
| engine | 2,083 | 0.5634% | 437 | 0.00004243% | 133 |
| vehicle | 1,884 | 0.5095% | 131 | 0.00001272% | 401 |
| emission | 2,194 | 0.5934% | 28 | 0.00000272% | 2183 |
| hydrocarbons | 290 | 0.0784% | 3 | 0.00000029% | 2692 |
| catalyst | 1,700 | 0.4598% | 13 | 0.00000126% | 3643 |
| autocatalyst | 27 | 0.0073% | 0 | 0.00000000% | Infinity |
| hydrocarbon | 140 | 0.0379% | 0 | 0.00000000% | Infinity |

Most of the open-class words with very high *co-efficient of weirdness* are specialist terms. The ratio of infinity indicates that whilst a word occurs in the special language corpus but cannot be found in a general language corpus; *infinity* here is the result of dividing a non-zero number by zero. The above prescription of comparing the relative frequencies in two corpora, one general language and the other specialist language, has been implemented in System Quirk, a text and terminology management system. The results reported in this paper have all been derived from corpora organised using System Quirk and all the text analysis and comparisons were also performed using the System (Ahmad and Holmes-Higgin 1995). But we digress.

The point about comparing *relative frequencies* is that one can automatically produce a list of **candidate single-word terms** from a text corpus by designing a program that can compute the co-efficient of weirdness. These candidates have to be approved by competent terminologists and subject specialists before being entered into a terminology data base. Furthermore, one can design yet another program that can identify, and compute the statistical significance of the collocates of the candidate terms. The collocation patterns can then be used to identify **candidate compound word terms**.

## 5. Case Studies from Lesser Used Languages

The development of ever larger English language text corpora was initiated by the long-standing corpus linguistic tradition which can be traced back to Firth and is currently spurred on by the intense competition in the English language teaching market. Thus, most dictionary publishers have their own corpora, including CoBUILD, Longman, American Heritage Corpus of Houghton-Mifflin, and so on.

There are national initiatives across Europe, North America and the Asian-Pacific region for compiling ‘representative’ and ‘contemporary’ corpora of English and of the respective national/regional languages. And, some of these corpora are made available to university-based researchers on accessible terms. Unfortunately, such a resource-intensive initiative cannot be considered by dictionary publishers in lesser-used languages. One hopes that in due course, given ready access to texts –at least through word processors and international communication networks– such a useful compilation of texts will be undertaken and the results made available soon afterwards.

In this section we will look at the case of three languages: Dutch, Norwegian and Welsh. For Dutch, although a lesser-used language, there is a large general language

corpus of modern Dutch - the Eindhoven Corpus; we show how to use this national corpus for terminology extraction. The case for Norwegian is different: the author is not aware of a publicly available corpus of modern Norwegian. We created our own Surrey/Bergen Norwegian general language corpus by 'surfing' on the Internet. This rather unrepresentative corpus was used for terminology extraction. The same procedure was carried out to collate the Surrey Welsh corpus.

*5.1. Weirdness in Engineering and Law of Copyright texts written in Dutch*

We have compared the frequency distribution amongst the first 40 most frequent words in the Eindhoven Corpus with the first 40 words in the two specialist corpora compiled at Surrey. The first corpus comprises six texts in automotive engineering - a total of 16,842 words. The second 'corpus' is made up of three texts comprising 34,075 words. Table 4a below shows a weirdness in the specialist Dutch texts, when compared with the Eindhoven Corpus, as was the case for English specialist texts and English general language texts:

Table 4a. 40 most frequent words in the Eindhoven Corpus (Dutch general language corpus) and two specialist Dutch corpora[3] (RF stands for relative frequency)

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| | **Eindhoven General Language Corpus (605,773)** | | **Surrey Automotive Engineering Corpus (16,842)** | | **Surrey Copyright Law Corpus (34,075)** | |
| | **Tokens** | **RF** | **Tokens** | **RF** | **Tokens** | **RF** |
| 1 | de | 6.49% | de | 8.19% | de | 7.64% |
| 2 | van | 3.24% | van | 4.22% | van | 5.13% |
| 3 | het | 2.97% | een | 2.76% | een | 3.94% |
| 4 | en | 2.96% | en | 2.42% | het | 3.14% |
| 5 | een | 2.64% | het | 2.23% | in | 1.95% |
| 6 | in | 2.39% | in | 2.07% | is | 1.86% |
| 7 | dat | 2.16% | is | 1.47% | of | 1.80% |
| 8 | die | 1.55% | te | 1.03% | en | 1.39% |
| 9 | is | 1.52% | met | 0.98% | die | 1.36% |
| 10 | ik | 1.33% | voor | 0.98% | op | 1.31% |
| 11 | te | 1.30% | dat | 0.95% | zijn | 1.16% |
| 12 | op | 1.22% | op | 0.94% | te | 1.15% |
| 13 | niet | 1.17% | zijn | 0.83% | voor | 0.90% |
| 14 | zijn | 1.14% | bij | 0.78% | dat | 0.89% |
| 15 | met | 1.01% | **katalysator** | 0.78% | aan | 0.86% |
| 16 | maar | 0.96% | door | 0.74% | niet | 0.86% |
| 17 | voor | 0.89% | worden | 0.71% | tot | 0.85% |
| 18 | hij | 0.85% | aan | 0.68% | **art** | 0.85% |
| 19 | je | 0.81% | deze | 0.63% | door | 0.84% |
| 20 | n | 0.81% | wordt | 0.59% | worden | 0.74% |
| 21 | dan | 0.79% | die | 0.58% | kan | 0.68% |
| 22 | als | 0.75% | ook | 0.56% | wordt | 0.64% |
| 23 | ze | 0.74% | dit | 0.55% | met | 0.61% |
| 24 | aan | 0.73% | niet | 0.49% | **recht** | 0.58% |
| 25 | er | 0.73% | om | 0.49% | bij | 0.53% |
| 26 | was | 0.71% | tot | 0.43% | heeft | 0.52% |
| 27 | ook | 0.66% | **uitlaatgassen** | 0.43% | dan | 0.50% |
| 28 | t | 0.54% | **auto** | 0.42% | deze | 0.50% |
| 29 | zo | 0.54% | **benzine** | 0.40% | als | 0.49% |
| 30 | ja | 0.52% | dan | 0.40% | goed | 0.48% |
| 31 | nog | 0.51% | er | 0.39% | dit | 0.45% |
| 32 | om | 0.48% | als | 0.36% | hij | 0.45% |
| 33 | bij | 0.47% | **motor** | 0.36% | artikel | 0.38% |
| 34 | door | 0.46% | **lucht** | 0.35% | uit | 0.34% |
| 35 | wat | 0.46% | of | 0.34% | indien | 0.33% |
| 36 | of | 0.44% | zal | 0.32% | wanneer | 0.32% |
| 37 | naar | 0.42% | nog | 0.32% | **auteursrecht** | 0.31% |
| 38 | wel | 0.42% | kan | 0.29% | lid | 0.31% |

[3] This data was collected by Peter W. van Kersbergen (Maastricht) at Surrey during August-December 1992. His contribution to the corpus-based studies of (Dutch) terminology is gratefully acknowledged.

| | | | | | | |
|---|---|---|---|---|---|---|
| 39 | uit | 0.42% | co | 0.28% | **rechtshandeling** | 0.28% |
| 40 | heeft | 0.41% | **verbranding** | 0.26% | hem | 0.28% |
| | **TOTAL** | **48.61%** | | **42.00%** | | **47.60%** |

Note that there is no noun in the first forty most frequent words in the general language corpus and what we have here are determiners, prepositions, conjunctions, pronouns, and primary verbs. What is perhaps more important to note is that these closed class words make up just under 50% of the 600,000 word corpus.

For the Dutch automotive engineering and law corpora, we see that the closed class words still dominate in that the first ten most frequent (and closed class words) make up 26% and 29% of the texts in the automotive engineering and the law corpus respectively. However, the open class nouns (six amongst the 40 most frequent words) make up over 2% of the texts in the automotive engineering and 1.82% of the law corpus (3 words and an abbreviation).

We see that what gives texture to specialist texts, as compared to general language texts, is the frequency of open class words. Table 4b is an illustrative computation of the so-called weirdness co-efficient.

Table 4b. The weirdness co-efficient of some terms in the Surrey's Dutch automotive engineering corpus

| **Word** | **Eindhoven Texts Relative Frequency (a)** | **Auto. Eng. Texts Relative Frequency (b)** | **Weirdness Ratio (=a/b)** |
|---|---|---|---|
| verbranding | 0.2600% | 0.2700% | 1.04 |
| benzine | 0.0008% | 0.0400% | 50.00 |
| zuurstof | 0.0025% | 0.1700% | 68.00 |
| motor | 0.0045% | 0.3600% | 80.00 |
| schadelijke | 0.0010% | 0.1700% | 170.00 |
| mengsel | 0.0003% | 0.1700% | 566.67 |
| uitlaatgassen | 0.0003% | 0.4300% | 1433.33 |
| katalysator | 0.0003% | 0.7800% | 2600.00 |

## *5.2. Norwegian* Spider Terminology[4]

The Surrey/Bergen corpus of Norwegian general language may or may not be ‘representative’ of contemporary Norwegian, but shows systematic differences between Norwegian special-language texts and general-language texts. These differences are important for terminology management.

The analysis of the general language corpus showed results similar to those obtained by studying the English texts in that the most frequent words belonged to the so-called closed class words together with a smattering of adjectives (see Table 5a): Our small corpus contains some texts that deal with access to computers (hence *informasjon*, and *elektronisk*) and has texts that deal with enterprises in Europe (eøs, the acronym for European Economic Area): hence there is some infiltration of the nouns in the first percentile. Nevertheless, the majority of the 107,304 word text, i.e. around 44%, comprises closed class words. The second percentile, not shown here for reasons of space, comprises 7.71% of the total tokens and contains nouns. This figure compares well with that of 6.70% for the Longman-Lancaster Corpus (Table 1b above).

Table 5a. The 10 most frequent words in the Surrey/Bergen Norwegian Corpus comprising 107,304 tokens. The English equivalents were taken from Kirkeby’s *Norsk-Engelsk ordbook* (1993)

| Token | Absolute Frequency | Relative Frequency |
|---|---|---|
| og (*and, too*) | 4576 | 4.26% |

[4] This section is largely based on an earlier joint paper by the author in collaboration with Magnar Brekke and Johan Myking (see Brekke, Myking & Ahmad 1996).

| | | |
|---|---|---|
| i (*in, at, of*) | 3483 | 3.25% |
| av (*of, by, off*) | 2544 | 2.37% |
| for (*for, to, too*) | 2331 | 2.17% |
| som(*who, that, which, as, like*) | 2130 | 1.99% |
| det(*it*) | 2006 | 1.87% |
| er (*is*) | 1987 | 1.85% |
| til (*to, for*) | 1740 | 1.62% |
| på (*on, at*) | 1482 | 1.38% |
| en (*a, one*) | 1452 | 1.35% |
| **TOTAL** | **23731** | **22.11%** |

We have obtained a specialist text, from the WorldWideWeb, dealing with *spiders* in Norwegian which was written by a scientist to describe some aspects of the behaviour of spiders (the Norwegian equivalent of spider is *edderkopp*). The text contains 3658 word tokens and has a total vocabulary of 1281 words. Each of the morphological and syntactic variants of a word, if expressed as a compound, is treated separately. Table 5b shows the 20 most frequent words in the specialist text on spiders:

Table 5b: Freq. distribution of 20 most frequent words in the *Edderkopp* text

| Total No. of tokens = 3658 | | | Total Vocabulary = 1281 | | |
|---|---|---|---|---|---|
| Norwegian Token | Abs. Freq. | Rel. Freq. | Norwegian Token | Abs. Freq. | Rel. Freq. |
| i | 95 | **2.60%** | for | 50 | **1.37%** |
| er | 94 | **2.57%** | de | 50 | **1.37%** |
| som | 77 | **2.10%** | å | 42 | **1.15%** |
| og | 75 | **2.05%** | det | 41 | **1.12%** |
| en | 71 | **1.94%** | at | 38 | **1.04%** |
| av | 65 | **1.78%** | ***edderkopper*** | ***38*** | ***1.04%*** |
| på | 64 | **1.75%** | et | 32 | **0.87%** |
| med | 62 | **1.69%** | seg | 31 | **0.85%** |
| til | 59 | **1.61%** | kan | 29 | **0.79%** |
| har | 57 | **1.56%** | etter | 22 | **0.60%** |

Note that the term *edderkopper* is the 16th most frequent term in the first percentile. The first percentile contains many other terms: *insekter*, *tarentula*, *silke* (see Table 5c):

**Table 5c. The 1st percentile of the frequency-ordered list of words from the *Edderkopp* text**

| Rank | Words | Cumulative Rel. Freq. |
|---|---|---|
| 1-10 | i, er, som, og, en, av, på, med, til, har, | 19.66% |
| 11-20 | for , de, å, det, at, ***edderkopper***, et, seg, kan, etter, ***edderkoppene*** | 10.20% |
| 21-30 | dette, den, ved, han, man, også, men, ikke, ***edderkoppen***, mange | 5.25% |
| 31-40 | inn, bytte, blir, hun, nett, denne, store, når, ***edderkoppens***, hvor | 3.69% |
| 41-50 | ofte, hunnen, så, samme, disse, om, bitt, ***insekter***, henne, nettet | 2.98% |
| 51-60 | veldig, små, ble, *edderkopp*, være, dyr, alle, stor, må, kunne | 2.49% |
| 61-70 | slik, skal, vi, kommer, over, mellom, under, bein, ***silketråd***, brukt | 2.19% |
| 71-80 | sitter, opp, sitt, før, hvis, ut, hannene, bare, ***tarentula***, fangnett | 1.97% |
| 81-90 | ulike, pga, byttedyr, da, går, liten, ***silke***, mens, fra, finnes | 1.67% |
| 91-100 | mot, rundt, mest, gift, mennesker, ca, raskt, kg, samtidig, antall | 1.61% |

(Note there are six different variants of *edderkopp*. The singular definite, indefinite and genitive respectively are *edderkopp*, *edderkoppen* and *edderkoppens*. The equivalent plurals are *edderkoppene*, *edderkopper*, and *edderkoppenes* respectively.)

The word *edderkopper* actually is not to be found in that restricted corpus and this gives a clue to the existence of terms in specialist texts. Table 5d contrasts the difference in distribution of words in the *Edderkopp* text with that of the general language corpus:

Table 5d. Contrasting the special language (SL) *Edderkopp* text with the corpus of Norwegian general language (GL) texts

| **Word** | **SL RF (a)** | **GL. RF (b)** | **Weirdness a/b** | **Word** | **SL RF (c)** | **GL. RF (d)** | **Weirdness c/d** |
|---|---|---|---|---|---|---|---|
| **og** | 2.05% | 4.25% | *0.48* | **på** | 1.75% | 1.38% | *1.27* |
| **det** | 1.12% | 1.86% | *0.60* | **de** | 1.37% | 1.01% | *1.35* |
| **for** | 1.37% | 2.17% | *0.63* | **er** | 2.57% | 1.85% | *1.39* |
| **av** | 1.78% | 2.36% | *0.75* | **en** | 1.94% | 1.35% | *1.44* |
| **i** | 2.60% | 3.24% | *0.80* | **kan** | 0.79% | 0.55% | *1.44* |
| **å** | 1.15% | 1.30% | *0.88* | **med** | 1.69% | 1.10% | *1.54* |
| **at** | 1.04% | 1.06% | *0.98* | **har** | 1.56% | 0.90% | *1.72* |
| **til** | 1.61% | 1.62% | *1.00* | **seg** | 0.85% | 0.27% | *3.12* |
| **som** | 2.11% | 1.98% | *1.06* | **etter** | 0.60% | 0.17% | *3.52* |
| **et** | 0.87% | 0.75% | *1.17* | ***edderkopper*** | 1.04% | 0 | ***INF*** |

Once the candidate terms are examined and the terminologist wishes to elaborate further, then he or she can examine the 'company' a particular candidate keeps by looking at a concordance of keywords in context, again produced by System Quirk for the *Edderkopp* text: the numbers on the left refer to the line number of the text and the keyword in context is the word *edderkoppen*[5]:

| | | | |
|---|---|---|---|
| 1_170 | sees klart hos kors - | edderkoppen | ( araneus diadematа ) som |
| 1_91 | tråder til , sikringstråd som | edderkoppen | alltid legger etter seg under |
| 1_107 | som et horisontalt teppe hvor | edderkoppen | beveger seg på undersiden . |
| 1_62 | med disse chelicerene biter | edderkoppen | bytte og sprøyter inn den |
| 1_79 | dette gjør at | edderkoppen | kan produsere silketråder med ulike |
| 1_291 | hos ulve - | edderkoppen | padosa amentata som finnes i |
| 1_116 | et virvar av tråder som | edderkoppen | raskt og elegant beveger seg |

*5.3. The creation of a term base of radiochemistry in Welsh*

Ahmad and Davies (1994) have created a prototype dictionary of radiochemistry in Welsh, comprising 200 entries, from a text of 7,660 words in Welsh and their English equivalents.

The prototype Welsh/English dictionary of radiochemistry was created with a view to demonstrating the speed by which specialist dictionaries can be compiled using machine-readable corpora of specialist text. The dictionary was compiled in direct response to a questionnaire survey which showed that one of the main problems in teaching and learning through the medium of Welsh was the lack of Welsh language resources.

Figure 1 shows the results of the computation of the *co-efficient of weirdness* in the specialist text *Cemegion Ymbelydrol* text for words. The system, System Quirk, was asked to signal words that have a weirdness ratio in excess of 10,000.

[5] System Quirk also has facilities for providing information related to the morphological variants through a facility in the System that allows a terminologist to express morphological heuristics which are then used to a frequency lists of lemmas only.

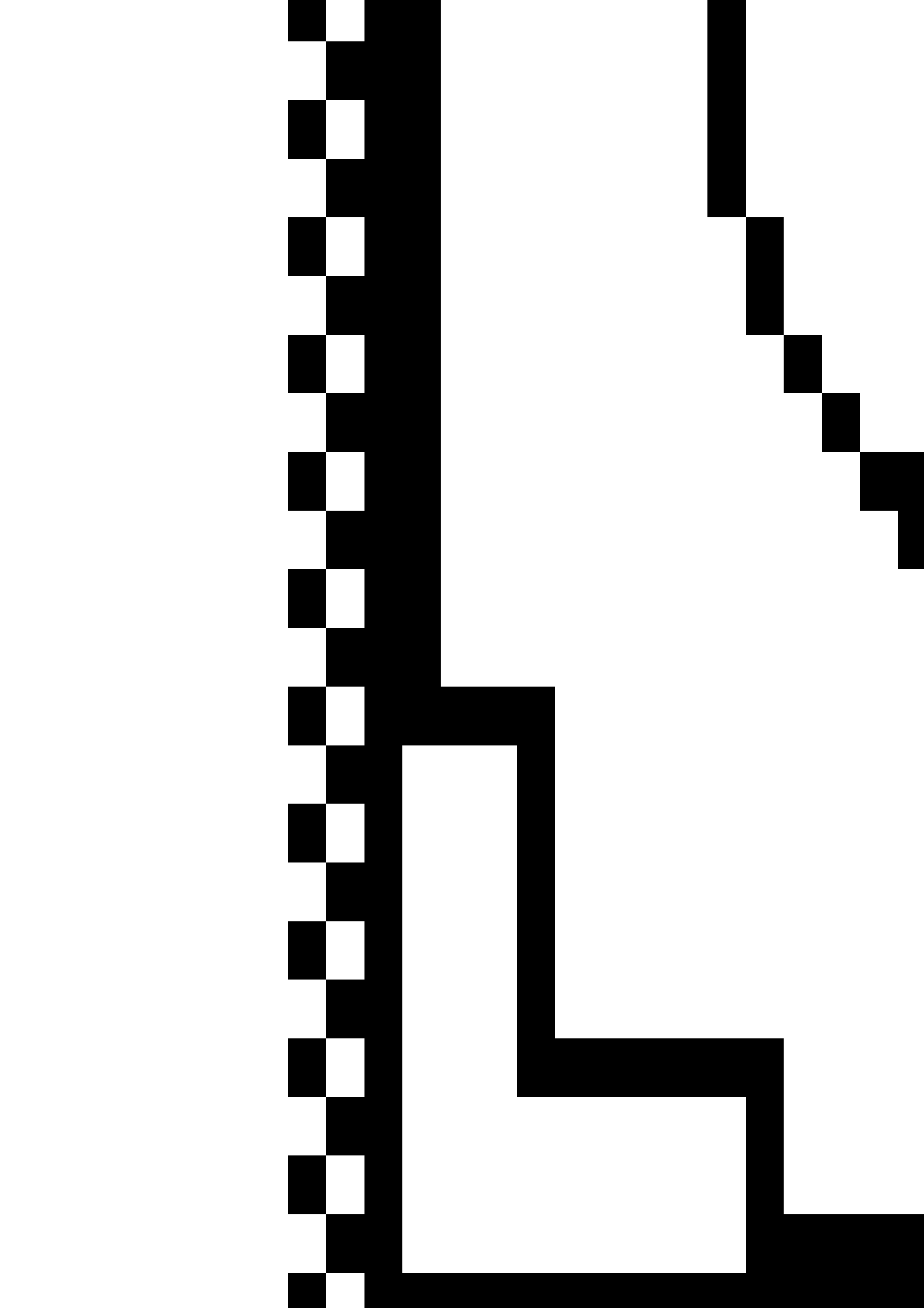

Figure 1: An analysis of the text *Cemegion Ymbelydrol* performed by System Quirk. The System signals a weirdness ratio of 10,000 or more. This was computed by comparing the relative frequencies of all words in *Cemegion Ymbelydrol* and then comparing these frequencies with Surrey Welsh Corpus of General Language.

The creation of our terminology data base of radiochemistry terminology in Welsh demonstrates how quickly and efficiently term bases may be built using computer-based corpora. Texts act as raw data for experts and lexicographers alike and, given text processing tools, the information which they contain may be extracted and manipulated very easily. The corpus of radiochemistry texts (11,900 words in total) proved to contain sufficient data to create a term base of 200 Welsh terms and their associated data at an average rate of 34 fully elaborated terms for a one-person day.

## 6. Resume

### *6.1. The promise of text-based terminology in a lesser used-language*

The intention has been to demonstrate what can be achieved with what would be regarded as a very small special language corpus, c. 260,000 words in a world of 20, 30, or now, 100 million word corpora.

There are two points that have emerged from the above discussion: first, corpora of specialist texts can be used to supplement a terminologist's intuition about the behaviour of certain terms. Second, a corpus-based approach, providing access to texts written by native speakers will help a non-native terminologist in understanding the lexicogrammatical behaviour of terms in second- or even third-language texts, thereby reducing the dependence on native speaker informants. A terminology collection related to a text corpus is, paraphrasing Sinclair (1987), based on measurable evidence!

The corpus-analytic methods and techniques suggested above should, in principle, help a terminologist to take into account as to how a term is used by the various authors, especially if a term has been borrowed from other languages, a common case when one deals with lesser-used languages. Once the evidence of the existence of a term has been collected, this can be given to a subject expert and a linguist. If the experts regard the evidence in a favourable light, then a terminologist can very easily extract the references relating to the use of the term.

### *6.2. Future direction?*

An interesting point, not discussed in this paper, relates to the *grammatical environment* of a term. Such an environment can be investigated by producing a concordance of the term within a text, or indeed over a whole corpus. Each instance of the use of a term can be extracted from the text (corpus) and the various usages of the term can be recorded. These records, used in conjunction with a frequency list and with the co-efficient of weirdness, can help a terminologist in deciding whether or not the candidate term is a term or not. And, if it is agreed that one indeed has found a term then the grammatical properties of the proposed term can be identified.

## BIBLIOGRAPHY

Ahmad, K., Davies, A. E. (1994). 'Weirdness' in Special-language Text: Welsh Radioactive Chemicals Text as an Exemplar. *Journal of the International Institute for Terminology Research*. Vol. 5 (No. 2), pp. 22-52.

Ahmad, K., Holmes-Higgin, P. R. (1995). System Quirk: A unified approach to text and terminology. In Picht, H., Budin, G. (eds.*): TAMA '94 Proceedings, Third TermNet Symposium, Terminology in Advanced Microcomputer Applications.* TermNet: Vienna, Austria. pp. 181-194.

Aijmer, K., Altenberg, B. (1991). (Eds.) *English Corpus Linguistics - Studies in Honour of Jan Svartvik.* Harlow (England): Longman.

Alford, Mark (1971). *Computer Assistance in Learning to read Foreign Languages: An Account of the Work of The Scientific Language Group*. Cambridge: Literary and Linguistic Computing Centre.

Brekke, M., Myking, J., Ahmad, K. (1996). Terminology Management and Lesser-used Living Languages: A Critique of the Corpus-based Approach. In (Ed.) Klaus Dirk-Schmitz and Christian Galinski. *Proc. of TKE '96: Terminology and Knowledge Engineering*. Frankfurt: Indeks Verlag.

Crystal, David (1997). *A Dictionary of Linguistics and Phonetics* (4th Edition). Oxford (UK) and Cambridge (Mass., USA): Basil Blackwell Ltd.

Halliday, M. A. K., Martin, J. R. (1993). *Writing Science: Literary and Discursive Power.* London and Washington D.C.: The Falmer Press.

Johnsson, S., Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar.* (2 volumes). Oxford: Clarendon Press.

Kirkeby, W. A. (1993). *Norsk-Engelsk ordbook (Annen Utgave).* Oslo: W. Nygaard Kunnskapsforlaget - Aschehoug A/S & Gyldendal Norsk Forlag A/S.

Kucera, H., Francis, W. N. (1967). *Computational Analysis of Present-Day American English.* Providence (R.I., USA): Brown University Press.

Quirk, R (1968). *The Survey of English Usage*. In (ed.) Essays on the English Language: Medieval and Modern. Harlow (Essex, England): Longman.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985*). A Comprehensive Grammar of the English Language*. London and New York: Longman.

Sinclair, J. McH. (1987). *Collins CoBUILD English Language Dictionary*. London and Glasgow: Collins Publishers.

Summers, Della (1991). *Longman/Lancaster English Corpus*. (Unpublished technical Report). Harlow: Longman Dictionary Publishers.

Svartvik, J. (1966). *On Voice in the English Verb*. The Hague: Mouton.

**********

## LABURPENA / RESUMEN / RÉSUMÉ / ABSTRACT

**Terminologia-kudeaketa eta ezagutza-prozesaketa hizkuntza minorizatuetan**

Terminologia-kudeaketa (terminoak biltzea, aztertzea, baliozkotzea eta antolatzea) oso garrantzitsua da informazioa ezagutza ulergarri eta aplikagarri bilakatzeko. Joera guztietako adituak dabiltza terminoak sortzen, dagoen terminologia aldatzen, termino batzuk arkaikotzat ematen edo baztertutako terminoak esanahi berriekin birsartzen. Badirudi adituen eginbeharreko bat dela neologismoak sortzea, erretronimoak (berrezarpen lexikalak) sartzea, terminoak itzultzea, terminoak baliozkotzea eta, nahiko zeharka, euren espezialitateetako terminologia-bildumak egitea edo egiten laguntzea. Zientzia modernoa, gaur egungo aisia eta dibertsioa, enpresa berritzaileak, euren enkarnazio zaharrengatik nabaritzen dira, ez bakarrik ondasun, zerbitzu edo tresnen bidez, bai eta zientziak, arteak eta kultura, eta negozioak eta enpresak deskribatzeko erabiltzen duten terminologiaren bidez ere. Terminologia-kudeaketa, oro har, eskulana da eta motibaturiko dokumentalistengan, itzultzaileengan eta terminologoengan bermatzen da; azken horrek beste bien funtzioa betetzen du. Termino espezializatuak artxibatzearekin eta aurkeztearekin lotuta dauden biltegiratze- eta berreskuratze-lan batzuk arindu dituzte gaur

egun terminologia kudeatzeko sistemek. Hala ere, gizaki adituek egiten dituzte bilketak, azterketak eta baliozkotzeak. Hiztun ugariko hizkuntzetan, datu-base terminologikoak behar dituzten eta horiek sortzeko dirua inbertitzeko prest dauden pertsona-kopuru handia egongo denaren esperantzak bermatzen du terminologia-kudeaketaren lan garestia. Hiztun gutxiago dauzkaten hizkuntzetan ez da hori gertatzen. Kasu horietan, terminologia-kudeaketa hizkuntz plangintzaren lanarekin lotzen da sarritan. Lan horrek, era berean, motibazio politikoak eta, askotan, emozio-karga bat ere badauzka. Hizkuntza minorizatuetan, terminologia-kudeaketak gizakiengan daukan menpekotasuna handiagoa da beste hizkuntza batzuetan baino.

Terminologia-kudeaketaren automatizazioa ez da programa informatikoak idaztea besterik gabe, lan hori, berez, zaila izan arren. Automatizazio hori egiteko jakin egin behar da zelan idazten diren adituen testuak, gizakiak zelan konpontzen diren arlo espezializatuen semantikarekin, diskurtsoaren ereduak zelan aldatzen diren testuaren egilearen eta irakurleen beharrizanen arabera. Hizkuntza minorizatuetan idatzitako testuak eta hizkuntza minorizatuei buruzko testuak ez dira erraz lortzen eta badirudi zaila dela arlo horietan azterketak eta datu-bilketak egiten hasteko inor konbentzitzea. Zelanbait, idazketa espezializatua “idazketa teknikoa” terminoarekin lotuta dago, makinekin zerikusia daukan diskurtso-eredu bat. Horren ondorioz, ez du lortzen hizkuntzaren eredu matematiko baten arabera esaldiak aztertzeari edo testuetan ikono kulturalak bilatzeari, lan abstraktuagoak baitira, ematen zaien estatus bera.

Ezagutza-prozesaketa terminoa hezkuntza, prestakuntza, irakaskuntza eta ikaskuntza, arazoak konpontzea, eta abarren moduko ekintza parekoak garatzeko erabil dezakegu. Terminologia espezializatuen bildumen eskuragarritasunaren menpe dago ezagutza-prozesaketa. Hori egia da, batez ere haurraren hezkuntza-urteetan, hasberriaren edo berriro prestatzen ari denaren hasierako ikasketetan: gizakiaren edozein ekintzarekin zerikusia daukaten gertakariak, printzipioak, teoriak eta esperientziak (ezagutza) asimilatu egiten dira eta gero irakasteko, ikasteko, arazoak konpontzeko eta abarretarako erabiltzen da. Liburuetan, egunkarietan, iraganeko esperientzia ez-artikulatuan eskuratzen den “material gordina” terminoen bidez komunikatu behar da, zeinen esanahiak ondo definituta baitaude eta adituek sarritan erabiliko baitituzte.

Beraz, ezagutza-prozesaketak eta terminologia-kudeaketak daukaten lotura askaezina da; terminologia-kudeaketa, era berean, hizkuntz plangintzarekin eta politikarekin dago lotuta. Azken hamar urteotan hiztun ugariko hizkuntzetako terminologia-bilketak egiten ibili gara, adibidez, ingelesean, alemanieran eta gazteleran. Aldi berean, metodo horiek hizkuntza minorizatuei egokitzen saiatu gara, adibidez, galeserari, norvegierari, flandeserari eta katalanari. Komunikazio honen gaiak izango dira aurkitutako erronkak, identifikatutako aukerak eta emandako irtenbideak hizkera espezializatuen terminologia inguru eleanizdun batean kudeatzeko, zeinetan hizkuntza bat gutxienez hizkuntza minorizatua baita. Gure marko teorikoa hizkuntza corpusari, filosofiari eta zientziaren historiari buruzko ikerketa berrietatik hartu dugu, alde batetik, eta informatikako ikerketatik, bestetik.

**Gestión terminológica y proceso cognitivo en lenguas minoritarias**

La gestión terminológica, esto es, la recopilación, el análisis, la convalidación y la distribución de términos, es una tarea crucial para convertir la información en conocimiento comprensible y aplicable. Hay especialistas de todo tipo de corrientes implicados en acuñar términos, modificar la terminología existente, definirlos como arcaicos o reintroducir términos desechados con nuevos significados. Una de las tareas de los especialistas es, al parecer, acuñar neologismos, introducir retrónimos, traducir términos, convalidar términos y, de forma bastante indirecta, recopilar o ayudar a completar recopilaciones de términos de sus respectivas especialidades. La ciencia moderna, el ocio y el entretenimiento de hoy en día, las iniciativas innovadoras, todo ello

se distingue de sus anteriores plasmaciones no únicamente a través de bienes, servicios y herramientas, sino también a través de la terminología que emplean para describir las ciencias, las artes y la cultura, los negocios y las iniciativas. La gestión terminológica es, sobre todo, una labor manual que se sustenta en la existencia de documentalistas, traductores y terminólogos bien motivados; los últimos desempeñando el papel de los dos primeros. Los sistemas de gestión terminológica asequibles hoy en día han aliviado algunas de las labores de almacenamiento y recuperación asociadas con el archivo y presentación de términos especializados. De todas formas, la recopilación, el análisis y la convalidación son funciones desempeñadas por personas cualificadas. En el caso de las lenguas mayoritarias, la costosa labor de la gestión terminológica se sustenta en la expectación que crea el hecho de que exista un gran potencial numérico de personas que reclaman y desean invertir en crear bases de datos terminológicos. Pero éste no es el caso de otras lenguas con menor número de hablantes; la gestión terminológica, en estos casos, a menudo está estrechamente vinculada con la políticamente motivada y, a menudo, emocionalmente cargada planificación lingüística. La dependencia en el factor humano de la gestión terminológica es mayor en el caso de las comunidades lingüísticas minorizadas que en el de otras lenguas.

La automatización de la gestión terminológica no es una mera cuestión de producir programas informáticos, aunque ésta sea una labor de por sí costosa. Esta automatización requiere un conocimiento de la forma en que se escriben los textos especializados, de cómo los seres humanos expresan la semántica de estos campos especializados, y de cómo el modelo discursivo varía de acuerdo con las necesidades del autor y de los lectores del texto. No es fácil encontrar textos en y sobre lenguas minoritarias, y parece que resulta difícil convencer a nadie para que asuma la labor de realizar estudios y recopilar datos en estas áreas. De alguna forma, la producción escrita especializada está asociada a la "producción técnica" de términos, un modelo discursivo que a su vez está asociado a máquinas, y por lo tanto no cuenta con el mismo estatus que la más abstracta tarea del análisis gramatical de frases mediante un modelo lingüístico matemático o de la búsqueda de iconos culturales en los textos, por ejemplo.

Pero el procesamiento cognitivo, término que puede emplearse para elaborar actividades interrelacionadas como la educación, la formación, la enseñanza, el aprendizaje, la resolución de problemas, etc. depende enteramente de la disponibilidad de recopilaciones de terminología especializada. Esto resulta especialmente cierto durante los años formativos de los niños, el inicio de carrera de un novato o para una persona que está siendo reentrenada; los hechos, principios, teorías y reglas que guardan relación con cualquier actividad humana y que conocemos colectivamente como conocimiento han de ser asimilados y aplicados a la enseñanza, aprendizaje, resolución de problemas, etc. La "materia prima" al alcance en libros de texto, prensa especializada, experiencia anterior todavía sin articular, ha de ser comunicada mediante términos con un significado bien definido y frecuentemente utilizado por una organización especializada.

El procesamiento cognitivo está, por lo tanto, inexorablemente unido a la gestión terminológica, la cual, por su parte, está unida a la planificación lingüística y a la política. Durante estos últimos diez años nos hemos dedicado a producir recopilaciones terminológicas de lenguas extensamente habladas, mientras, al mismo tiempo, intentábamos adaptar estos métodos a lenguas más minoritarias como el galés, noruego, flamenco o el catalán. Esta ponencia tratará de los desafíos con los que nos hemos encontrado, de las oportunidades identificadas y de las soluciones sugeridas para gestionar la terminología de lenguajes especializados en espacios plurilingües donde al menos una de las lenguas pertenece a la categoría de lenguas minoritarias. Nuestro marco teórico proviene de recientes estudios en el campo del corpus lingüístico, filosofía e historia de la ciencia por una parte, y de la informática, por otra.

**La gestion de la terminologie et traitement des connaissances dans les langues d'utilisation mineure**

La gestion de la terminologie, c'est à dire la compilation, l'analyse, la validation et la distribution de termes est un élément crucial pour convertir l'information en une connaissance compréhensible et applicable. Des experts de tout genre se voient impliqués dans la création de termes, la modification de la terminologie existante, la conversion de termes archaïques ou l'introduction de termes avec de nouvelles significations. Une des tâches des experts, semble être, celle de créer des néologismes, d'introduire des rétronymes, de traduire des termes, de valider des termes et, d'une façon assez indirecte, de compiler ou d'aider à compiler des collections de terminologie dans leur domaine de spécialisation. La science moderne, les divertissements et les loisirs contemporains, de même que les entreprises innovatrices se distinguent tous de leurs prédécesseurs non seulement à travers leurs biens, leurs services ou leurs engins mais aussi à travers la terminologie qu'ils utilisent pour décrire les sciences, les arts et la culture, les affaires et les entreprises elles-mêmes. La gestion de la terminologie est, en général, une tâche manuelle qui dépend de l'existence de documentalistes, de traducteurs et de terminologues très motivés; ces derniers mettent en pratique la fonction des premiers. Actuellement, les systèmes de gestion des terminologies disponibles ont soulagé quelques unes des tâches de stockage et de récupération associées avec les archives et la présentation de termes spécialisés. Néanmoins, les tâches de collecte, d'analyse et de validation sont réalisées par des êtres humains experts. Dans les langues utilisées par des majorités numériques, la tâche de la gestion de la terminologie est garantie par le fait qu'il existe un nombre immense de personnes potentielles qui en a besoin et qui est disposé à investir dans la création d'une base de données de termes. Pour les langues utilisées par une minorité de personnes, il se passe tout le contraire; dans ce cas la gestion de la terminologie se trouve souvent liée à aux personnes qui ont des motivations politiques ou à celles qui ont un travail de planification linguistique avec une certaine charge émotionnelle. La dépendance que la gestion terminologique a sur les êtres humains est plus grande dans les communautés de langues d'utilisation mineure que dans le cas des autres langues.

L'automatisation de la gestion et du travail terminologique n'est pas simplement la tâche de créer des programmes informatiques bien que cette entreprise soit onéreuse en elle-même. Cette automatisation a besoin d'une compréhension au sujet de la façon dont on rédige le texte spécialisé, de la façon dont les êtres humains traitent la sémantique des domaines spécialisés et de la façon dont les patrons du discours changent selon les besoins des auteurs et des lecteurs des textes. Les textes écrits dans, et au sujet, des langues d'utilisation mineure ne sont pas faciles à obtenir et il semble qu'il est difficile de persuader les gens à faire des recherches et à compiler des données dans ces domaines. Pour une raison quelconque, l'écriture spécialisée est associée avec le terme "écriture technique", un patron de discours qui à son tour est associé avec les machines et qui donc ne reçoit pas la même condition de la tâche abstraite d'analyser des phrases grammaticalement selon un modèle de langage mathématique ou, par exemple, la recherche d'icônes culturels dans les textes.

Mais dans le traitement des connaissances, un terme qui peut être utilisé pour élaborer les activités en relation comme l'éducation, la formation, l'enseignement et l'apprentissage, la résolution de problèmes, etc... dépend de manière décisive de la disponibilité des collections spécialisées de terminologie. Cela a une importance spéciale durant les années de formation de l'enfant, le commencement de la carrière d'un nouvel étudiant ou d'une personne qui est en train de se recycler: les faits, les principes, les théories et les normes empiriques en relation avec toute entreprise humaine qui sont connus de manière collective avec le mot connaissance, doivent être assimilés et ultérieurement appliqués dans l'enseignement, l'apprentissage et la résolution de problèmes, etc. La "matière première" disponible dans les livres de texte, les revues et l'expérience inarticulée antérieure doit être communiquée moyennant une agence de termes dont les significations sont clairement définies et fréquemment utilisées par une entreprise spécialisée.

Donc, le traitement de la connaissance est une gestion de terminologie particulièrement établie, qui, à son tour est en relation avec la planification linguistique et la politique. Au cours des dix dernières années nous avons construit des collections de terminologie dans des langues utilisées par des groupes de personnes numériquement grands, comme l'anglais, l'allemand et l'espagnol, tandis qu'en même temps nous avons essayé d'adapter ces méthodes aux langues d'utilisation mineure comme le gallois, le norvégien, le flamand et le catalan. L'exposé présent abordera les défis rencontrés, les occasions identifiées et les solutions suggérées par la gestion de la terminologie dans les langues spécialisées d'entourages multilingues où au moins une des langues appartient à la catégorie d'utilisation mineure du point de vue numérique. Notre cadre théorique surgit du travail récent réalisé dans le corps linguistique, la philosophie et l'histoire de la science d'une part et les sciences informatiques de l'autre.

**Terminology management and knowledge processing in lesser-used languages**

Terminology management, that is collection, analysis, validation and distribution of terms, is crucial for converting information into comprehensible and applicable knowledge. Specialists of all persuasions are involved in coining terms, modifying existing terminology, rendering terms archaic or re-introducing discarded terms with new meanings. One task of specialists, it appears, is to coin neologisms, introduce retronyms, translate terms, validate terms and, in a rather indirect manner, compile or help to compete terminology collection of their specialisms. Modern science, contemporary leisure and entertainment, innovative enterprises, all distinguish themselves from their older incarnations not merely through goods, services or artefacts, but also through the terminology they use to describe the sciences, arts and culture, and business and enterprises. Terminology management is by and large a manual task that relies on the existence of well-motivated documentalists, translators and terminologists; the later performing the function of the former two. Currently available terminology management systems have alleviated some of the storage and retrieval tasks associated with the archival and presentation of specialist terms. However, the tasks of collection, analysis and validation are undertaken with skilled human beings. In languages used by numerical majorities, the expensive task of terminology management is underwritten by the expectation that there is a potentially large numbers of people who require and are willing to invest in creating terminology databases. For languages used by numerically smaller number of people this indeed is not the case; terminology management here is often linked with the politically-motivated, and often emotionally charged, work of language planning. The dependence on human beings for terminology management is greater in the lesser-used language communities than say may be the case of other languages.

The automation of terminology involvement management is not merely a task of writing computer programs, although such and undertaking is onerous in itself. Such an automation requires an understanding of how specialist text is written, how human beings deal with semantics of specialist domains, how discourse pattern change according to the needs of the authors and the readers of the texts. Writings in and about lesser-used languages are not easy come by and it appears that it is difficult to persuade people to undertake research and data collection in these areas. Somehow, specialist writing is associated with term 'technical writing', a discourse pattern which in turn is associated with machines and thereby not given the same status as the more abstract task of parsing sentences according to a mathematical model of language or searching for cultural icons in texts for instance.

But knowledge processing, a term that can be used for elaborating related activities like education, training, teaching and learning, problem solving and so on, is crucially dependent on the availability of specialist terminology collections. This especially true during the formative years of a child, the early carrier of a novice or a person being re-trained; facts, principles, theories, and rules of thumb related to any human enterprise,

collectively known as knowledge, are to be assimilated and then applied for teaching, learning, problem-solving etc. The 'raw material' available in text books, journals, unarticulated past experience, has to be communicated through the agency of terms whose meanings are well defined and used frequently by a specialist enterprise.

Knowledge processing therefore is inextricably linked terminology management which, in turn, is linked with language planning and politics. Over the last ten years we have been building terminology collections in languages used by numerically larger groups of people, like English, German and Spanish, whilst at the same time attempting to adapt such methods for lesser used languages like Welsh, Norwegian, Flemish and Catalan. This paper will discuss challenges encountered, opportunities identified and solutions suggested for managing terminology of specialist languages in multilingual environments where at least one language belongs to the lesser used category on numerical groups. Our theoretical framework draws from recent work in corpus linguistics, philosophy and history of science on the one hand and computing sciences on the other.

*Table 2a. A contrastive view of frequency distribution in text corpora from three disciplines, automotive engineering, dance analysis and nuclear physics, compared with the distribution of the 25 most frequent words in the general language Longman-Lancaster Corpus*

| Rank | *Longman-Lancaster* 28 Million words | | | *Automotive Engineering* 369,751 words | | | *Dance Analysis* 44,607 words | | | *Theoretical Nuclear Phys* 81,946 words | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Word Form | RF (%) | Word Class | Word Form | RF (%) | Word Class | Word Form | RF (%) | Word Class | Word Form | RF (%) | Wo Cla |
| 1 | the | 6.09 | closed | the | 7.15 | closed | the | 6.23 | closed | the | 9.39 | clos |
| 2 | of | 3.06 | closed | of | 3.34 | closed | of | 3.45 | closed | of | 5.81 | clos |
| 3 | and | 2.80 | closed | and | 2.36 | closed | and | 3.16 | closed | in | 2.35 | clos |
| 4 | to | 2.51 | closed | to | 2.23 | closed | a | 2.69 | closed | and | 2.34 | clos |
| 5 | a | 2.19 | closed | in | 2.10 | closed | to | 2.15 | closed | to | 2.19 | clos |
| 6 | in | 1.88 | closed | a | 1.92 | closed | in | 2.08 | closed | a | 1.95 | clos |
| 7 | it | 1.13 | closed | is | 1.33 | closed | is | 1.42 | closed | is | 1.59 | clos |
| 8 | that | 1.10 | closed | for | 1.09 | closed | s | 1.09 | closed | that | 1.38 | clos |
| 9 | I | 1.08 | closed | with | 0.86 | closed | as | 0.96 | closed | for | 1.16 | clos |
| 10 | was | 1.05 | closed | on | 0.75 | closed | with | 0.93 | closed | be | 0.98 | clos |
| 11 | is | 0.91 | closed | as | 0.68 | closed | that | 0.87 | open | by | 0.88 | clos |
| 12 | he | 1.04 | closed | be | 0.66 | closed | it | 0.84 | closed | from | 0.83 | clos |
| 13 | for | 0.75 | closed | are | 0.64 | closed | by | 0.69 | closed | with | 0.82 | clos |
| 14 | as | 0.70 | closed | by | 0.62 | closed | on | 0.65 | closed | it | 0.78 | clos |
| 15 | with | 0.69 | closed | that | 0.62 | closed | **dance** | 0.61 | **open** | i | 0.72 | clos |
| 16 | his | 0.65 | closed | **emission** | 0.59 | **open** | this | 0.59 | closed | **particles** | 0.70 | **ope** |
| 17 | on | 0.65 | closed | this | 0.58 | closed | for | 0.56 | closed | this | 0.66 | clos |
| 18 | you | 0.60 | closed | at | 0.56 | closed | was | 0.56 | closed | on | 0.66 | clos |
| 19 | had | 0.59 | closed | **engine** | 0.56 | **open** | are | 0.56 | closed | **atoms** | 0.64 | **ope** |
| 20 | be | 0.59 | closed | **vehicle** | 0.51 | **open** | her | 0.51 | closed | are | 0.62 | clos |
| 21 | at | 0.52 | closed | **system** | 0.48 | **open** | from | 0.50 | open | as | 0.60 | clos |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | but | 0.52 | closed | **car** | 0.48 | **open** | he | 0.48 | closed | **nucleus** | 0.58 | **ope** |
| 23 | not | 0.51 | closed | **catalyst** | 0.46 | **open** | at | 0.45 | closed | was | 0.56 | **ope** |
| 24 | she | 0.49 | closed | it | 0.45 | closed | which | 0.45 | closed | which | 0.54 | clos |
| 25 | they | 0.46 | closed | which | 0.44 | closed | **work** | 0.45 | **open** | an | 0.53 | clos |
| **To-tal** | | 32.6 | | | 31.6 | | | 32.9 | | | 39.3 | |