K. Ahmad, H. Fulford, and M. Rogers, University of Surrey, UK

The elaboration of special language terms: the role of contextual examples, representative samples and normative requirements

ABSTRACT: The use of contextual examples appears to be gaining currency in general-purpose lexicography, but not yet in special-purpose lexicography ("terminography"). A range of lexicographical practice and opinion is described in this paper, incorporating new data collected by means of a questionnaire. We further report on our work concerning the use of contextual examples in terminography, focussing on a set of guidelines for the computer-based corpus-driven processing of contextual examples for a large-scale multilingual term bank for translators. Within this framework we discuss how our approach to terminography takes into account the equally important requirements of representativeness and normalisation in the processing of special-language terms. This work was carried out as part of the Translator's Workbench project (ESPRIT II No. 2315).

1 Introduction

By "contextual example" we understand a text fragment which contains the headword, or a morphologically-related form of the headword, which is presented together with the headword and other information in the dictionary entry. Elsewhere, contextual examples have been referred to as "citations", "illustrative quotations" or simply as "examples". The purposes of including contextual examples in dictionaries of different kinds, in so far as these are ever made explicit, are various. They include the illustration of semantic information (both denotational and connotational), grammatical information, and usage information, and are often considered as a supplement to definitions. Examples may be excerpts from authentic texts, or in some cases, sentences or phrases invented by the lexicographer. A third option is to modify text excerpts otherwise considered unsuitable. The majority of work on contextual examples relates to Language for General Purposes (LGP) for a variety of user groups, most notably foreign-language learners. Little work has been conducted until now on the use of contextual examples in Language for Special Purposes (LSP), i.e. in special purpose lexicography or "terminography" for specific user groups. We report here on work in this area currently being carried out at the University of Surrey.

The scope for treating contextual examples is indeed wide both in terms of possible purpose, and of procedure, thus making the task of the lexicographer or terminographer a daunting one. In order to remedy such a situation, Drysdale (1987, 223) pleads for greater "conscious editorial control" over examples, such as that "normally applied to definitions". In this paper we report on a set of guidelines for the selection of contextual examples in the specific context of a multilingual term bank for translators in the light of current lexicographical practice. Such guidelines are, however, only meaningful in the context of recent advances in computer-based working procedures in lexicographical work (Calzolari, Picchi & Zampolli 1987; Sinclair 1991) and terminographical work (Ahmad, Holmes-Higgin & Langden 1990). These procedures concern mainly the organisation of large text databases and the use of information retrieval techniques. They have led to a more descriptively-oriented corpus-driven approach to lexicography and now to terminography.

Given the means to systematically and automatically search large machine-readable corpora, the collection of data (including contextual examples) is no longer problematic at a logistic level; the problem is rather the selection of data from the wide range available. In terminography, computers have been used since the 1960s for the storage and retrieval of terminological data (term banks), but the application of computer-based procedures to other stages in the terminographical process, such as data capture, has until recently been neglected in favour of a conceptually-based approach (Felber 1984; Picht & Draskau 1985; Wüster 1985). Work conducted at the University of Surrey under the auspices of the Translator's Workbench (TWB) project (ESPRIT II 2315) has sought to fill this gap with the development of a fully-integrated set of software tools to cover all stages of a term's life-cycle and of terminographical work, from data capture from text, through the preparation and storage of terminological records, to editing and printing. The toolkit is known as the "MATE" (Machine-Assisted Terminology Elicitation) system (Holmes-Higgin & Griffin 1991). MATE is a generic set of tools: it is not domain-specific. The corpus at present contains c. 800,000 words in the domain of automotive engineering, and is structured according to language/language variety, text type and subdomain. We also refer later in this paper to other large-scale corpus-based work.

In the next section (section 2) we look briefly at dictionary users and contextual examples, before moving on to an outline of our survey of current lexicographical practice (section 3). The selection of contextual examples for special languages in a corpus-based computerised environment is then discussed (section 4). The penultimate section describes how such a corpus-based approach to terminology can satisfy the requirements of both representativeness and normalisation (section 5). Finally, some conclusions are offered (section 6).

2 Dictionary user types and contextual examples

The literature of lexicography seems to focus on the needs which foreign-language learners as a user group have for contextual examples. The potential usefulness of examples for mother-tongue users has also been discussed (Cowie 1989), the difference apparently being one of degree rather than kind. Contextual examples or illustrative sentences have long been a feature of dictionaries for foreign-language learners but the provenance of such examples has been a matter of some discussion in the literature: should they be citations taken directly from authentic texts, modified citations, or inventions?

In terminography, the provenance of contextual examples has not been an issue to date; LSP dictionaries only rarely contain contextual examples. In fact, in our terminology work, we have so far identified just one specialist dictionary in the exemplary domain of automotive engineering which contains some contextual examples: DICTIONARY OF AUTOMOTIVE EMISSION CONTROL (Schmitt 1986). A number of factors may have contributed to this apparent neglect of contextual examples in terminography. In general-purpose dictionaries, examples can be used to show "typical" features of a headword or to distinguish between a range of meanings and grammatical behaviour of semantically broad words such as "cast" which is shown in Collins, for instance, to have 37 meanings. By contrast, in LSP, terminologists have argued that a term has - or should have - a relatively fixed meaning within a defined domain. Furthermore, the conventional approach to terminology has rarely involved the use of real text as a primary source of data, preferring instead concept-based norm-oriented procedures. An important part of our corpus-based work has therefore been to initiate and explore the systematic and comprehensive use of contextual examples in LSP terminography, a step which has been greatly facilitated by computer-based resources and procedures.

One aspect of this work has involved an analysis of the criteria to be adopted when selecting contextual examples from those collected from our corpus of LSP texts. So far, we have found that references to selection criteria for authentic text excerpts in the lexicography literature mention only general language work. Fox (1987), for instance, notes a number of criteria for selecting contextual examples which were adopted in the *Cobuild* Project (COllins Birmingham University International Language Database): these criteria state that the examples should be typical, natural, and authentic. Where applicable, they should also contain collocations. In order to elicit explicit and current information about the criteria used in lexicography for selecting contextual examples, we decided to conduct a survey among lexicographers. This survey is the subject of the next section.

3 Current lexicographical practice

The intention of our questionnaire survey among lexicographers was to determine their views on, and treatment of examples. The questionnaire was distributed to a sample of 21 lexicographers. The first section contained questions about the usefulness of examples for different types of user, and the provenance for such examples (e.g. text corpus, inventions, and so on). We also posed questions about the criteria used for selecting examples from authentic sources, and the editing of such examples. The second section of the questionnaire related to background details about the lexicographer, including the type of dictionary projects they are working/have worked on and their length of lexicographical experience.

The response rate to the questionnaire was approximately 62% (13/21). Responses were received from a number of lexicographers working primarily on general language monolingual (both mother tongue and foreign-language learner) and/or bilingual dictionaries. All respondents, regardless of the type of dictionary they were working on, stated that contextual examples are useful. The majority of respondents had between five and ten years' experience in lexicographical work. The preliminary results of the survey are summarised below:

- the reasons given for the value of contextual examples included: to enable the user to distinguish between senses; to provide usage evidence; to show typical collocations; to amplify and clarify definitions; to show typical use of words.
- inventing examples is a less popular strategy than selecting examples from authentic sources.
- all respondents stated that they select examples from authentic sources; some readily edit these examples, others are more reluctant to do so, but acknowledge that editing is sometimes unavoidable.
- most respondents select examples from machine-readable text corpora using concordancing packages or "reading and marking" programs; others scan texts manually.
- a number of criteria were noted for selecting contextual examples; these included: typicality; naturalness; length; usefulness of syntactic information provided by example; semantic complexity of example.

In the next section we report on contextual examples within an LSP framework, where the intended user of our data is a translator who, somewhat like a foreign-language learner, has to become familiar with new vocabulary, its syntactic and semantic requirements, and its usage.

4 A text-based approach to terminography

The conventional approach to terminology of the Wüster or Vienna School stresses the need for what is called the "systematic" organisation of terms. By this is meant the organisation of terms according to relations between concepts of which terms are the linguistic designation. The "systematic" organisation of terms is usually seen as an alternative to alphabetic (or word-based) organisation, still the most common practice in general purpose dictionaries. It has been argued (Arntz & Picht 1989, 194-5) that for LSP domains the alphabetical ordering principle makes it difficult to establish whether a domain has been comprehensively covered and also requires the lexicographer to include polysemous terms under the same entry. In a complementary way, a conceptual organisation is said to have disadvantages for LGP vocabulary since it is difficult in such a disparate collection of words to find clear organising principles (Arntz & Picht 1989, 193). Terminologists have been reluctant to adopt a word-based approach to their work, following early unsuccessful attempts to deal with large amounts of data using this method (Wüster 1970, 207). Consequently, in their attempts to capture and record lexical data, terminologists have been more wary than lexicographers of using "real" texts, since the linguistic sign rather than the concept is considered primary. A concern with normative requirements, most notably the "elimination" of synonymy and homonomy, has also played a role here.

While standardised terminology may indeed help to avoid potential misunderstandings in, for instance, expert-to-novice communication, it is nevertheless not only prescriptive (e.g. with regard to which terms are preferred), but also idealised (e.g. in contrast to the use of terms in specific textual contexts at specific times). Furthermore, translators, who are typically cited as the principal user group of terminologies, must deal with a variety of text types in which a wide range of terms and usage is represented. Consequently, their need is not simply for a normative collection of terms, but for a representative one. A structured corpus of texts representing those text types normally encountered by the translator therefore provides the most suitable source of terminological data for this user group. The Surrey automative engineering LSP corpus has in fact been constructed in close collaboration with Mercedes-Benz AG language services department, the intended end-user of the TWB terminology.

As language professionals concerned with the decoding and re-encoding of knowledge in text, translators need access to elaborative information which goes beyond a straight match between source language term and target language term. Of the information usually provided, the definition is usually considered the most important (see, for instance, Felber 1984, 160). What constitutes a definition is, however, a matter of some controversy. Within the framework of the Vienna School, a distinction is made between a definition and an "explanation", the difference being that definitions are said to take account of relative conceptual positions in the "system of concepts", whereas explanations simply describe a concept. Elsewhere, the international standard ISO/R 1087-1969 (E) (VOCABULARY OF TERMINOLOGY) refers to three types of definition, including the so-called "contextual definition", a definition which attempts to explain the meaning of a term by giving an example of how it is used. Such definitions are, according to Picht and Draskau, generally unsatisfactory as definitions for terminological purposes (see Picht & Draskau 1985, 51-5 for further discussion). However, we believe that in the framework of a descriptively-oriented approach to terminography, such examples are of particular value in the elaboration of terms.

The practice adopted in the Surrey term bank is therefore to include both a definition and a contextual example for each term. For a user group such as translators, the contextual example provides a valuable illustration of usage, collocational patterns, semantic and grammatical requirements (semantic selection restrictions and subcategorisation rules), as well as serving as a supplement to the meaning description given in the definition proper; e.g.:

entry term	combustion chamber
definition:	Space between the piston and cylinder head in which the fuel/air mixture is burnt.
example:	The peak temperature in the combustion chamber and the duration of its effect have a decisive influence on the concentration of nitrous oxide emissions.

Any particular contextual example is unlikely to provide an archetypical context for the term in question, and each example only shows what is possible, rather than what is not possible. But when viewed in conjunction with the descriptive information on usage, grammar, meaning, collocation, and so on, provided in discrete form elsewhere in the term bank, the contextual example performs the important function of integrating at least some of this information and illustrating its use in practice. Such an illustration is particularly valuable for translators when provided for the target language, i.e. for production purposes.

As indicated earlier, once the acquisition of potential contextual examples is no longer a problem, attention focusses on the more interesting problem of selection. In order to encourage a principled approach to this, a set of guidelines has been developed at Surrey for use by terminologists building term banks. The guidelines contain illustrations of both good and bad practice in the form of text excerpts, together with explicit descriptions of the criteria implicit in the illustrations. For example, the following potential example for the term "carbon monoxide" should be rejected on a number of grounds, including a non-interpretable reference outside the sentence, a list of terms, and the appearance of a number of terms in addition to "carbon monoxide": "The first way is to improve combustion, which reduces the quantities of hydrocarbons, carbon monoxide, and nitrogen oxides produced; this is a precombustion system." Instead, the following example is given as a guide to good practice: "Carbon monoxide is a major pollutant found in exhaust."

Currently, the guidelines contain 18 criteria; these are grouped into higher-level categories, including, comprehension and production, as illustrated in Figure 1 below:

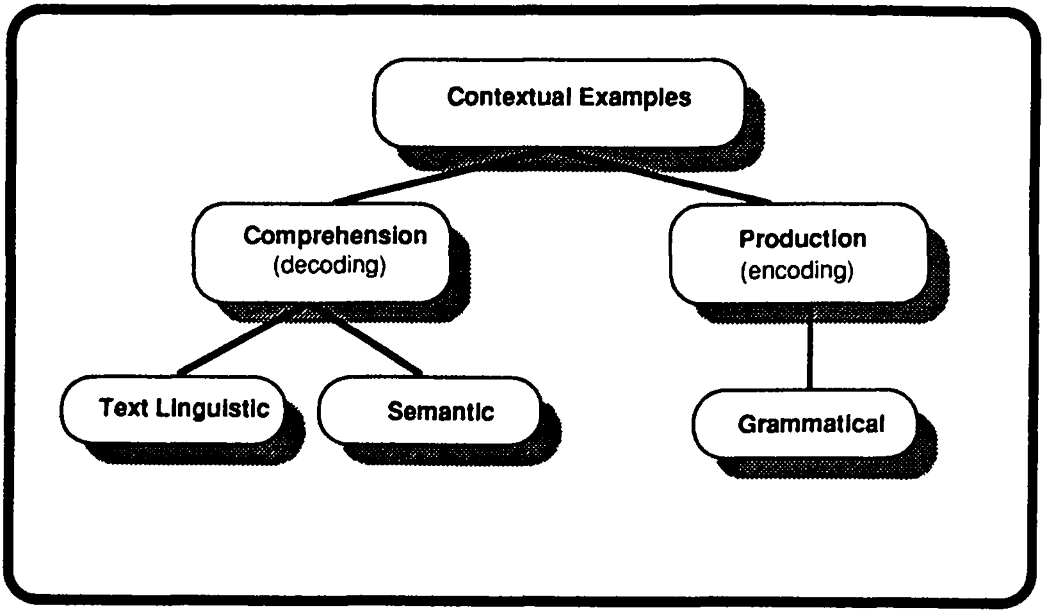
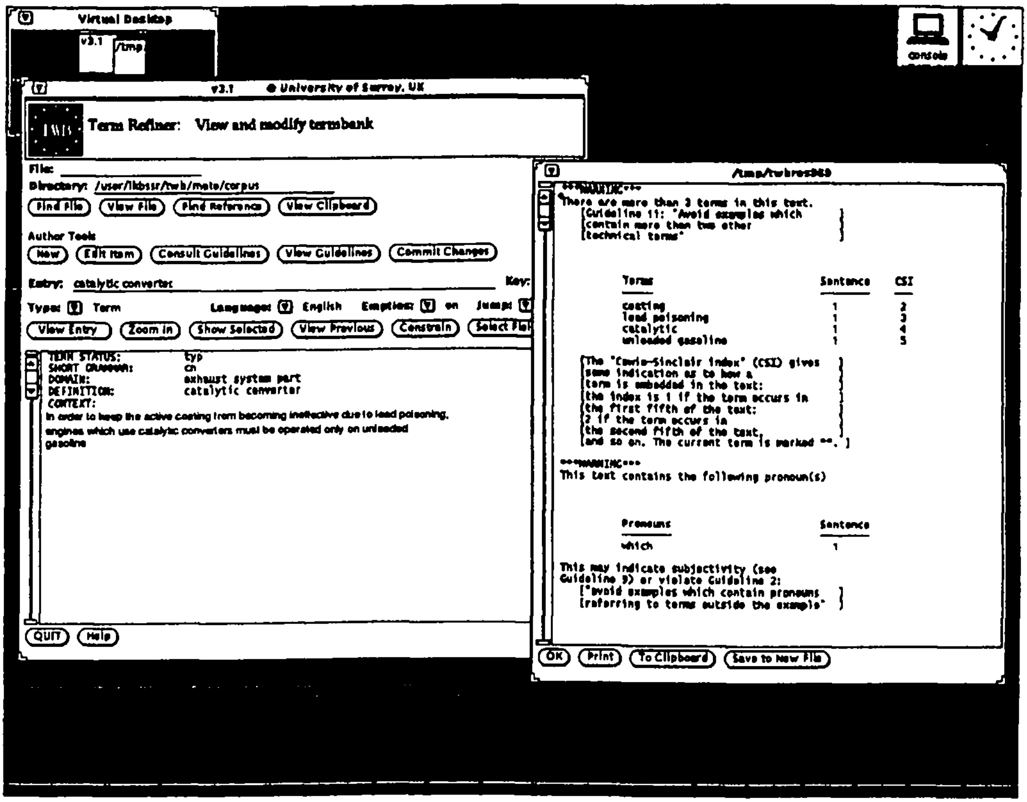


Figure 1: Classification of criteria for the selection of contextual examples from an LSP corpus

Those criteria which are statistically-based have been animated for the user, i.e. for the terminologist in prototype form. By animation, we mean that the computer interprets a set of criteria and uses them to analyse contextual examples already selected from the corpus by the terminologist using specially-designed MATE information retrieval tools. The results of the computer analysis are then presented on screen for the user as a basis for a decision on the suitability of the chosen examples for entry into the term bank.



File:
Directory: /user/lkbssr/twb/mate/corpus
Entry: catalytic converter Key:
Type: Term Language: English Emptions: on Jump:

```
TERM STATUS:        typ
SHORT GRAMMAR:      cn
DOMAIN:             exhaust system part
DEFINITION:         catalytic converter
CONTEXT:
In order to keep the active coating from becoming ineffective due to lead poisoning,
engines which use catalytic converters must be operated only on unleaded
gasoline
```

```
***WARNING***
There are more than 3 terms in this text.
    [Guideline 11: 'Avoid examples which    ]
    [contain more than two other            ]
    [technical terms'                       ]

          Terms                        Sentence   CSI
          coating                          1        2
          lead poisoning                   1        3
          catalytic                        1        4
          unleaded gasoline                1        5

    [The 'Cowie-Sinclair index' (CSI) gives ]
    [some indication as to how a            ]
    [term is embedded in the text:          ]
    [the index is 1 if the term occurs in   ]
    [the first fifth of the text:           ]
    [2 if the term occurs in                ]
    [the second fifth of the text,          ]
    [and so on. The current term is marked **. ]

***WARNING***
This text contains the following pronoun(s)

          Pronouns                     Sentence
          which                            1

This may indicate subjectivity (see
Guideline 9) or violate Guideline 2:
    ['avoid examples which contain pronouns  ]
    [referring to terms outside the example' ]
```

Figure 2: Sample output of animated guidelines

The information provided for the terminologist is still of a limited kind, reflecting the current limitations of computers and the paucity of programs which can perform language production, comprehension, or acquisition tasks at a comparable level with humans. Nevertheless, we still believe that animation is helpful in that the computer effortlessly and indefinitely analyses in a consistent manner all possible contextual examples presented to it, providing support to the user in the actual selection of the most appropriate example. The kind of information provided for the user includes: the number of terms contained in the example in addition to the term in question; the number of words in the example; an indication of any subjective words present in the example; and the position of the term in the example (see "Cowie-Sinclair" index). Currently, this facility operates for English.

In attempting to devise criteria for the selection of optimal contextual examples, one overarching concern has been the needs of the end-user (translator) in relation to the processing of the examples on screen. Consider, for instance, the following selected criteria (reformulated for brevity):

a. Avoid examples containing pronouns referring outside the example

b. Avoid examples containing more than two other technical terms

c. Avoid examples with a complex structure

d. Avoid examples which are long

e. Favour examples which are a complete sentence

f. Favour examples where the term appears early on rather than late

Texts are characterised by a number of features, the most well-known being cohesion and coherence. Excising a sentence from a text means isolating it from its many textual interconnections, both syntactic and semantic. Criterion (a), for example, is an attempt to ensure that the end-user is not faced with the impossible task of interpreting a pronominal reference for which there is no referent, as in the following example for the term "brake master cylinder": "Depending on the switching state, they connect the wheel brake cylinders either with the corresponding circuit of the brake master cylinder or with the return pump, or close off the wheel brake cylinder from both the circuit and the pump." Reducing the recommended minimal unit for an example below sentence level (criterion e) may further increase the problems of maintaining interpretability.

Pragmatic considerations also play a role in the interpretability of texts. While various models are possible for thematic progression in a text (Papegaaij & Schubert 1988, 94-9), it is still generally the case that within a sentence, known or given information (theme) tends to precede new information (rheme) in unmarked or neutral order (Firbas 1974). From the perspective of the end-user of a contextual example, the term can be assumed to be the given information, or, put a little differently, what is presupposed. Hence, examples where the term appears towards the end of the sentence are unlikely to meet the pragmatic expectations of the user (criterion f), as in the following contextual example for "sump": "One front wheel is driven directly by a shaft from the differential, the other by an intermediate shaft which passes through a tube in the sump."

Psycholinguistic evidence suggests that a number of factors may improve the speed and accuracy of comprehension. Among these are: the absence of certain structures such as centre embeddings, and knowledge about "what normally happens" (Clark & Clark 1977, 60-1; 73-5). Hence, in the first case, it is advisable to avoid selecting examples with certain types of structure such as centre embeddings. This is one measure of "complexity" (criterion c). Syntactic complexity may be, of course, defined in a number of ways. In their discussion of the analysis of the Brown corpus of English texts, Francis & Kucera (1982: 550-3) suggest that the number of predications per sentence is a useful measure. They show that informative texts are characterised by a larger number of predications per sentence than imaginative texts (2.78 and 2.38 respectively). For the genre "learned text" the number is even larger: 2.84. In Francis & Kucera's definition, a predication is indicated by a verb or verbal group, including both finite (any tensed verb with a subject) and non-finite (infinitives, gerunds, participles). Such a criterion of complexity cannot at present be animated in our system, since our corpus is not grammatically tagged. The heuristic for selection of an optimal contextual example could, however, in the future be related to the mean number of predications per sentence related to text type.

A further measure of complexity is a lexical one: the density of special-language terms (criterion b). The following example, for instance, is not recommended since it contains a left-branching embedded structure as well as five terms other than the search term, "Hinterräder": "Die Hinterräder werden direkt über das am Motor angeflanschte Schalt- und Ausgleichgetriebe über Seitengelenkwellen angetrieben." The difficulty here is deciding where the borderline lies. This may be both user- and domain-specific and needs to be resolved by further empirical work.

With regard to the user's "knowledge" and expectations about the world referred to in the example, we have already mentioned pragmatic considerations of information order. Semantically, however, it is less clear how such a consideration can be made relevant to special language texts, where the user is a non-expert in the domain concerned.

Finally, there is anecdotal evidence that users of term banks, particularly busy translators, are not prepared to read through extensive material on screen (criterion d). But how long is "too long" for a contextual example? One objective measure would be to establish the mean sentence length for selected text types, and to use this as a guide for judging the suitability of contextual examples with regard to length. For example, in the Brown corpus, the mean number of words per sentence in informative texts is 21.06; for learned texts, the mean is higher (22.31), for press reportage, lower (20.72) (Francis & Kucera 1982, 552).

Our guidelines for selecting contextual examples also contain criteria concerning the avoidance of subjective material, proprietary names, and material from headings and legends, as well as preference for examples containing typical collocational patterns, for examples where the word class of the entry term is retained (i.e. avoidance of partial homonyms), and for examples where certain grammatical features of the entry term are retained, such as plurality if this is the default as in: "hydrocarbons", or countability as in the case of "gasoline" and "fuel": "While unleaded gasolines are becoming available in Europe, the maximum allowable Pb content in the Federal Republic of Germany is 13 mg Pb/11 (49 mg/gal), that is fifteenfold higher than the Pb level currently found in unleaded fuels in the United States."

5 Representativeness and normalisation

We fully acknowledge the need for normative requirements in terminology and for the principal artefact of those requirements, standardised terms. Standardisation of terminology increases the potential for safe and efficient use of the domain knowledge, given that the knowledge of a domain is partially encoded in its terminology. However, terms do not come into existence merely with the announcement of a list of standardised terms. Terms are the building blocks of all the language-based communication in any domain, and the knowledge of the domain has its own lifecycle: birth (the emergence of a new domain or revision of an established domain), maturation (growth and development), maturity (the establishment of codes of practice and knowledge sources such as books, journals, papers and articles, sales/marketing literature, and so on), mutation (within the domain or by transfer between domains), and finally, death (obsolescence). Terms, it appears, evolve accordingly in a variety of communicative environments, and it is only at the established or mature stage that standardisation becomes crucial and even possible. Even then, efficient and effective standardisation depends on the availability of a sufficient volume of communicated material (as manifested by the existence of large text corpora in conventional print). Hence the thrust of our approach is that representative samples of texts will (and do) eventually lead to standardised terminology. The analysis of text-type specific LSP syntactic characteristics will also be facilitated by the creation of representative LSP corpora.

6 Conclusion

We believe that our work provides an important bridgehead between what have erroneously been regarded as two rival approaches to terminology. These two approaches are reflected on the one hand in the ever-pressing need to standardise terms and their use, and on the other hand, in the need to achieve representativeness of the range of terms used in the full range of LSP communication.

Our approach to terminography is a novel one in so far as it is corpus-driven and user-informed; both translators and terminologists have been involved in the selection of text types and texts in building an LSP corpus, as well as in the continual evaluation of our guidelines for the selection of authentic and appropriate examples and the software which partially animates these guidelines. These two factors, i.e. the use of machine-readable corpora and the focus on the user, characterise a trend which emerged in LGP-based lexicographical work during the 1980s, but which had until now not been applied to LSP-based terminographical work.

Bibliography

AHMAD, K., HOLMES-HIGGIN, P. & LANGDEN, A. (1990): A computer-based environment for eliciting, representing and disseminating terminology. Periodic Report for the Translator's Workbench Project (ESPRIT 2315) for the period 1.4.89 - 31.3.90.

ARNTZ, R. & PICHT, H. (1989): Einführung in die Terminologiearbeit. Georg Olms Verlag, Hildesheim, Zürich, New York. 2te Auflage.

CALZOLARI, N., PICCHI, E. & ZAMPOLLI, A. (1987): "The use of computers in lexicography and lexicology". In: The Dictionary and the Language Learner. Ed. by A. Cowie. Max Niemeyer Verlag, Tübingen.

COWIE, A. P. (1989): "The language of examples in English learners' dictionaries". In: Lexicographers and their works. Ed. by G. James. University of Exeter, Exeter.

CLARK, H. H. & CLARK, E. (1977): Psychology and Language. Harcourt, Brace Jovanovich, San Diego.

DRYSDALE, P. (1987): "The role of examples in a learner's dictionary". In: The Dictionary and the Language Learner. Ed. by A. Cowie. Max Niemeyer Verlag, Tübingen.

FELBER, H. (1984): Terminology Manual prepared for the General Information Programme and UNISIST and for the International Information Centre for Terminology. Unesco & Infoterm, Paris.

FIRBAS, J. (1974): "The Czechoslovak Approach to FSP". In: Papers on Functional Sentence Perspective. Ed. by F. Danes. Mouton, Paris/The Hague.

FOX, G. "The Case for Examples". In: Looking Up. Ed by J. Sinclair. Collins, London and Glasgow.

FRANCIS, W. N. & KUCERA, H. (1982): Frequency Analysis of English Usage: Lexicon and Grammar. Boston: Houghton Miffin Co.

FULFORD, H. & ROGERS, M. (1990): Draft Guidelines for Entering Data in the TWB Term Bank. Translator's Workbench Project, ESPRIT II No. 2315, Workpackage 1.1 Report, Guildford: University of Surrey

HOLMES-HIGGIN, P. & GRIFFIN, S. (1991): MATE User Guide. Translator's Workbench Project, ESPRIT II No. 2315, Workpackage 1.3 Report, Guildford: University of Surrey.

ISO/R 1087-1969 (E) Vocabulary of Terminology.

PAPEGAAIJ, B. & SCHUBERT, K. (1988): Text Coherence in Translation. Foris, Dordrecht and Providence RI.

PICHT, H. & DRASKAU, J. (1985): Terminology: An Introduction. University of Surrey, Guildford.

SCHMITT, P. A. (1986): Dictionary of Automotive Emission Control. Brandstetter Verlag, Wiesbaden.

SINCLAIR, J. (ed.) (1987): Looking Up. Collins, London and Glasgow.

SINCLAIR, J. (ed.) (1988): Collins Cobuild English Language Dictionary. Collins, London and Glasgow.

SINCLAIR, J. (1991): Corpus, Concordance, Collocation. OUP, Oxford.

SUMMERS, D. (ed.) (1987): Longman Dictionary of Contemporary English. Longman Group UK Ltd., Harlow.

WÜSTER, E. (1970; 1931): Internationale Sprachnormung in der Technik. Besonders in der Elektrotechnik. 3. abermals ergänzte Auflage.

WÜSTER, E. (1985): Einführung in die allgemeine Terminologielehre und terminologische Lexikographie. Copenhagen School of Economics, Copenhagen. 2nd edition.

KEYWORDS: computerised terminography, term banks, translation, language for special purposes (LSP)