[*Translating and the computer 11*. Proceedings of a conference … 16-17 November 1989, ed. Catriona Picken (London: Aslib, 1990)]

The Translator's Workbench Project[1]

Khurshid Ahmad, Heather Fulford*†, Paul Holmes-Higgin*, Margaret Rogers†* and *Patricia Thomas*†*

*Department of *Mathematics and Department of †Linguistic and International Studies, University of Surrey, Guildford, Surrey GU2 5XH, UK*

Presented by Patricia Thomas

SUMMARY

A methodology is being developed for creating the terminology of emerging domains which are multidisciplinary in nature. To demonstrate the methodology, terminology of three sub-domains of automotive engineering including catalytic converters is being developed. Methods, tools and techniques are being developed to show how their terminology can be identified, organised and disseminated using corpus-based computer-assisted methods of text analysis, artificial intelligence (AI) and database systems within the linguistic framework of language for special purposes (LSP).

AIMS OF THE TRANSLATOR'S WORKBENCH PROJECT

The principal goal is to develop a set of computer tools to provide translators with facilities for multilingual text processing, grammar, style and spelling checking, with remote access to a machine-assisted translation system

[1] This study forms part of the Translator's Workbench Project under the aegis of the European Commission's ESPRIT II Programme. The project started in April 1989 and is scheduled to run for three years, involving a consortium of academic and industrial partners in West Germany (Triumph-Adler, Nürnberg; Mercedes-Benz, Stuttgart; Siemens, Munich; Fraunhofer Gesellschaft, Stuttgart; University of Heidelberg), Spain (Siemens, Barcelona; Technical University of Catalunya), Greece (L-Cube, Athens) and the United Kingdom (University of Surrey).

(METAL) and to termbanks (e.g. Eurodicautom). The style checkers will serve as pre- and post-editing tools by preparing a source text for translation and by suggesting further refinements to the target text. A multilingual core editor will integrate the facilities being developed using standard interfaces to ensure portability and incorporating an ODIF (office document interchange format) and lexical interchange format (MATER), thus offering compatibility with office documentation architecture (ODA).

The role of the University of Surrey is threefold:

— to elicit and quantify translators' requirements
— to develop a machine-assisted methodology for creating LSP terminologies using lexicographical methods and advances in information processing, as well as tools and techniques from artificial intelligence and knowledge engineering
— to incorporate the translators' requirements into the machine-assisted methodology by building a domain-specific termbank and prototyping a toolset for creating and maintaining termbanks.

TRANSLATORS' REQUIREMENTS: SPECIFICATION

This study is being carried out in conjunction with Mercedes-Benz. Throughout this investigation, translation is viewed as a process which starts with the arrival of the text and which ends with the despatch of the completed translation (in whatever form or medium). The initial task is to assess the requirements of translators who will be the users of the translating system. In order to elicit these requirements, a number of systems analysis requirements methods have been used (Birrell and Ould, 1985). Initially the approach has been questionnaire-based, followed by in-depth observation of translators' working methods, and structured and focused interviews. The questionnaire approach was introduced by Patricia Thomas and Margaret Rogers (1986/7) at the University of Surrey where questionnaires were formulated to assess the needs of LSP translators. These questionnaires have been extended by Heather Fulford and Monika Hoge (1989) to look at translators' needs as a whole, covering the analysis of the source language (SL) text, terminology look-up, access to machine-assisted translation tools, verification of the target language text, ergonomic issues and so on.

The questionnaire has already been distributed to targeted groups of translators in West Germany, the United Kingdom and other European countries. The targeted groups are: in-house translators, freelance translators and student translators. Results are currently being analysed. Secondly,

[2] This task is being carried out jointly by Heather Fulford (University of Surrey) and Monika Höge (Mercedes-Benz).

translators have been observed at work and informal discussions held about their methods and their attitude towards computer assistance in translation work. Thirdly, selected translators will be interviewed on a 'focused' topic, particularly on the question of the human computer interface (HCI). Fourthly, the same interviewees will be questioned in so-called 'structured' interviews about more general questions concerning the translation process. Interviews will be conducted both in West Germany and in the United Kingdom.

Interim conclusions show that considerable time is spent in consulting domain experts and other sources. Furthermore, the use of external translation services as a result of the heavy workload results in extensive post-editing by in-house translators.

MACHINE-ASSISTED TERM ELICITATION (MATE)

Towards a new methodology

When dealing with LSP texts in newly emerging domains, translators are faced with terminological problems. The terms of such domains emerge rapidly and often continue to evolve. Such terminology is not standardised. The translators' problems are compounded by the bilingual nature of their work. Hence there is a need to establish a methodology for term elicitation in such emerging domains which is efficient, speedy and accurate.

In our work we have identified four consecutive processes in terminology building: investigation, discovery, organisation and presentation. In each of these processes, linguistic methods have been primarily corpus-based, informed by LSP studies; the tools to assist the terminologist in building a termbank are computer-based, deriving inspiration from computational linguistics (e.g. concordance and collocational analysis) and AI (e.g. data structures for type hierarchies, supertype to subtype, see Figure 1).

The standardisation of the terminology of a domain always lags behind the development of the domain itself, partly because the domain experts are still consolidating their own knowledge of the domain; since its terminology helps to encode the domain knowledge, the delay is understandable. The acquisition of terms themselves, their definition and the identification of appropriate illustrative examples of contextual use require the scanning of a wide variety of technical documents in different languages and text types, e.g. journals, advertising material, workshop manuals. Once the term has been identified and acquired, with its definition and context, it has to be verified by an expert to delimit the scope of the term.

Finally, in the dissemination process, feedback from translators provides further information about the term in its natural habitat – the 'text'. Terminologists helping translators are generally people trained in languages who would find it difficult to scan, understand and elicit useful information from

a large number of technical texts. We believe that computational and text linguistic tools will help terminologists to scan texts and to 'scoop' contextual usage from a pre-stored computer-based corpus. However, we also believe that a terminologist should acquire a rudimentary knowledge of the domain before using such tools.

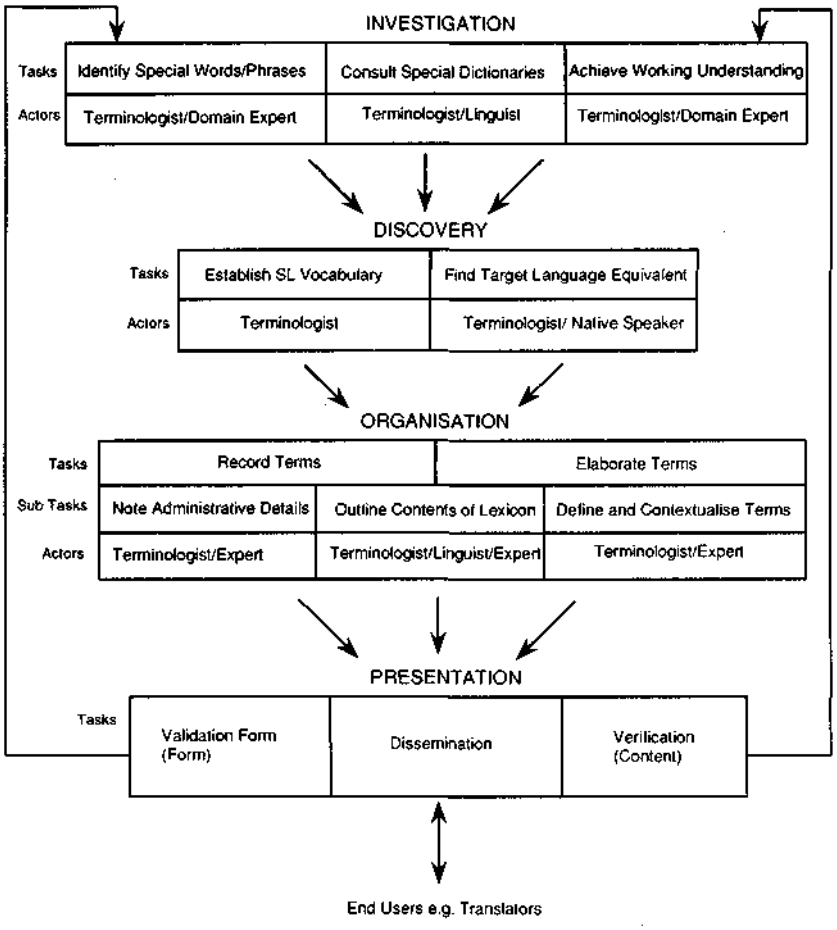


Figure 1. The process of developing a bilingual computerised termbank

The process of machine-assisted terminology building is being developed and tested, using a corpus-based approach in accordance with current lexicographical practice (COBUILD, 1987; *Longman's dictionary of contemporary English* (LDOCE), 1987). Some 160 texts on these sub-domains

have been gathered in German, British English and American English and put into machine-readable form using the optical character reader (OCR) at the Oxford University Computing Service and by copy-typing texts on to Surrey's mainframe computer. Spanish texts will be added later.

LSP corpora and text typology

Six text types have been identified by the end-users, Mercedes-Benz: books, learned journals, conference papers, workshop manuals, newspaper articles, and advertisements and marketing material. In addition, glossaries and dictionaries are being consulted.

In order to have a representative sample from the text which contains the majority of terms, instances of contextual use, grammatical and semantic variation, replication with synonyms, collocational patterns and so on, a machine-readable corpus of over 300,000 words distributed over 130 texts has been created. Our target is currently 500,000 words which we estimate will contain about 5,000 terms (these numbers are a 'guesstimate'), since initial estimates indicate that, from the total number of words in a text, approximately 10 per cent, will comprise different lexical items and of these, 1 per cent, will be LSP terms of the domain being studied.

These target figures can be related to other machine-readable corpora (of English), e.g. the Brown Corpus and the London-Oslo-Bergen (LOB) Corpus of written English, each containing one million word tokens over 500 texts of 15 text categories. Each text is exactly 2,000 words long. Of the 15 'text categories' (or 'genres') included in the two corpora, only the category 'learned' (including science and technology) contains as many as 80 texts for each corpus: a total of 160,000 words in each. All other categories contain fewer texts. The rest of the Brown and LOB corpora contain text categories such as: press reportage, religion, skills and hobbies, popular lore, etc. (Leech, 1987).

'Terminology in text'

Texts are being scanned using a custom-made concordance package called KonText for the elaboration of terms (see below):

— Various techniques are being explored for identifying terms in each sub-domain, each language (or language variety) and each text type, including frequency- and form-based methods.
— KonText is also being used to identify suitable examples of the contexts in which identified terms occur, as well as providing support in the construction of definitions.
— In addition, the text-based elicitation of terms provides information on the grammatical features of LSP terms, which cannot be generally assumed to be identical to those in language for general purposes (LGP), e.g. *formulae (LSP), formulas* (LGP).

— Lists of terms have been created from the corpus with corresponding (numbered) bibliographic source references.
— Each term is automatically allocated a number.
— Examples of terms in context are being recorded, again with numbered bibliographic references. Definitions are currently being acquired from LSP dictionaries, from domain experts and from the source texts; those devised by terminologists are being verified by experts. The presentation of the information to the user/translator on screen is being dealt with separately. The intention is to develop KonText to the extent that it can extract the 'term' or 'semi-finished' definition from the text.

MATE: DEVELOPMENT OF A DEMONSTRATOR TERMBANK

An emerging domain

In order to demonstrate the methodology being developed at Surrey, particularly for testing the creation and maintenance of the terms in a termbank, we are currently developing the terminology for newly emerging sub-domains of automotive engineering, an area of interest to Mercedes-Benz. Standardised terminology for such domains is not yet available; it is planned to have the identified terms verified by translators and domain experts at Mercedes-Benz. Once the terms are identified, they are currently being stored in a termbank using a relational database management system, ORACLE. The termbank record format comprises three sections for each term: administrative, linguistic and terminological (conceptual), each section containing a number of fields (Figure 2). The underlying structure of a domain is clarified by terminological analysis (for example, by conceptual diagrams) which establishes the relationships between the concepts found in the domain (Figure 3). These diagrams are then verified by experts, thus helping the terminologist to organise the termbank logically. The knowledge encompassed in the definitions of the terms can be analysed and structured so that it may eventually form part of a knowledge base, enabling the context of the definition to be addressed by the machine.

Tools for eliciting terms from text: KonText

KonText, or Knowledge on Text, is a custom-built software package for analysing texts. It has been developed using PROLOG on a SUN workstation by Paul Holmes-Higgin (1989) at the University of Surrey and is WIMP (window, icon, mouse, pull-down menu)-based. Non-experts in a domain find it difficult to scan texts effectively, particularly in a multilingual context. The purpose of KonText is to provide a concordance facility to generate word frequency lists, frequency statistics, concordances, collocations and so on. One method for identifying LSP terms using KonText which has been implemented

is first to exclude frequently occurring ‘noise’ words from a text (e.g. common and LGP words such as the 6,500 words which comprise the Alvey Natural Language Tool's lexicon). The words in the text being scanned which do *not* appear in this list are automatically capitalised by KonText. In addition, terms which have *already* been identified are marked with asterisks. In this way it may be possible to highlight the LSP terms of the domain under investigation (Figure 4). A facility is also provided to identify and to mark compounds with question marks.

No. of characters	Field name	Explanation	
Administrative data			
5	KEY	Key no (00001 etc.)	
3	ORIGIN	Surrey (SUR)	
3	POOL	Sub-domain (cat.con. etc.)	
6	ORIGNTR	Originator (HU, DF etc.)	
8	ENTRYDAT	(03-JUL-89)	
12	DOMAIN	AU.EN (Automotive Engineering)	
5	LCCODE	Lang./country code (DE/DE, BR/UK, AM/US)	
80	ENTRY	Entry term	
3	TEXTTYPE	(BOO, JOU, GLO, NSP, ADV)	
5	BIBREF.	00001 etc. (Full references to appear automatically from stored file)	
Linguistic data			
80	SYN1	Synonym 1 (in-house jargon)	
80	SYN2	Synonym 2	
80	SYN3	Synonym 3	
8	ABFORM	Abbreviated form	
80	VARIANTl	Variant 1 (spelling etc.)	
80	VARIANT2	Variant 2	
80	VARIANT3	Variant 3	
80	ANT	Antonym	
50	COLLOC	Collocation	
80	SCOPE	Note on usage etc.	
10	GRAMMAR	Speech part/gender	
	CONTEXT	Piece of text; previous translation for reference	
Conceptual relationships			
10	CONCEPOS	Key no. of related term	
80	BT	Broader term)	
80	NT	Narrower term) for thesaurus building	
80	RT	Related term) if required	
20	CLASSIF	Classification (UDC, ISO, DIN, BS etc.)	
	DEFINITION	Intrinsic characteristics:	
80		IS_A	
80		HAS_A	
80		PART_OF	
80		MATERIAL	(What it's made of)
		Extrinsic characteristics	
80		FUNCTION	(What it does)
80		CAUSES	(Cause and effect)
80		LOCATION	(Where it is)

Figure 2. ORACLE termbank record format (draft version) at the University of Surrey

Another facility is that of specifying collocations, where collocation is understood to be the *occurrence of two or more terms in a specified length of text, not necessarily contiguously.*For example, a search for ‘mercedes’ and ‘abs’ (anti-lock

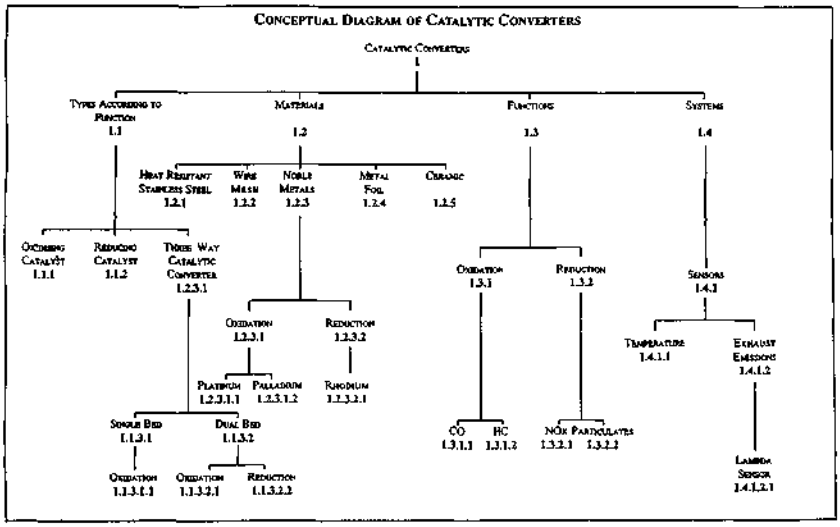


Figure 3. Conceptual diagram showing systems, functions and materials of catalytic converters (Meriel Hughes, 1989)

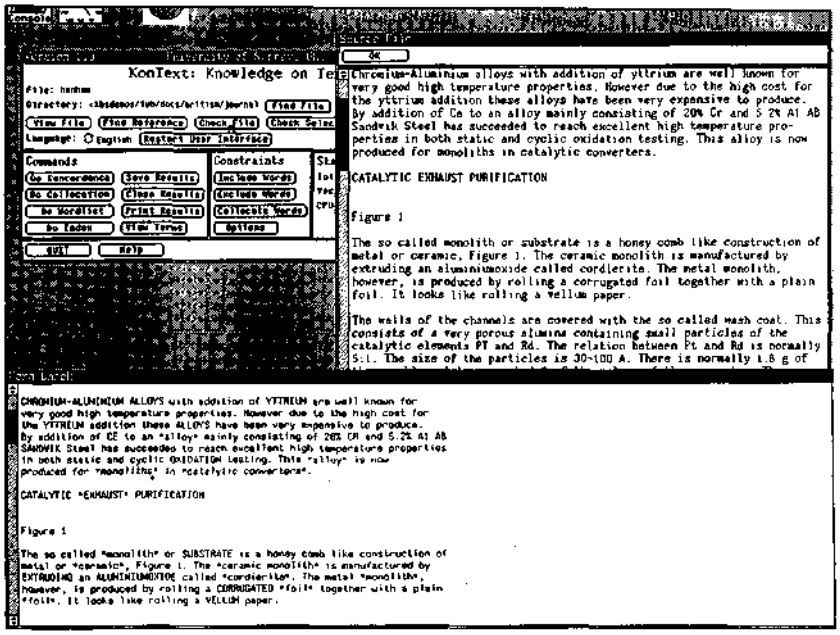
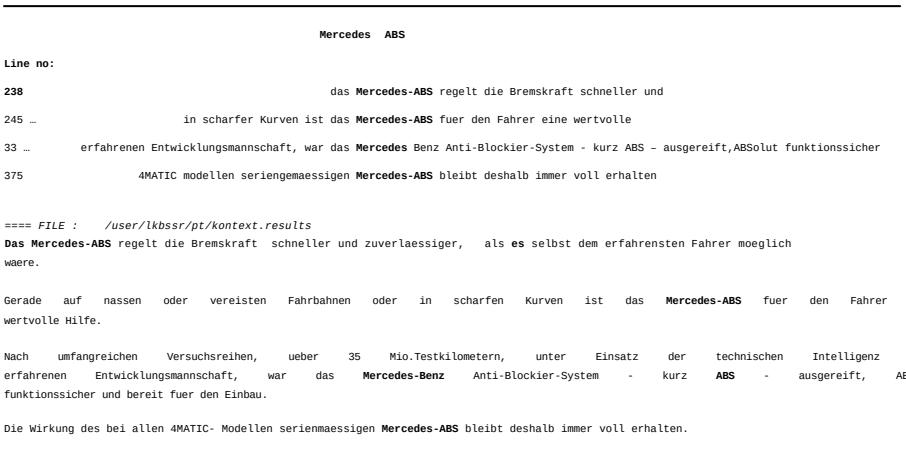


Figure 4. Highlighting of LSP terms using KonText (see Word Check window) developed by Paul Holmes-Higgin, 1989

braking system) will produce examples of these terms in context (Figure 5). A number of options is provided for the user to control aspects of the text processing, such as the number of words to be shown either side of the term being sought, alphabetic or frequency listings and so on. In addition to displaying these extracts, the multi-windowing system allows the whole sentence to be displayed showing the term operating in its 'natural habitat' when the user 'clicks' on the line number provided. A facility will be provided to paste suitable examples of the term in context directly into the appropriate record field in the termbank, when the system is fully operational on the SUN.



Mercedes ABS

Line no:

238 das **Mercedes-ABS** regelt die Bremskraft schneller und

245 … in scharfer Kurven ist das **Mercedes-ABS** fuer den Fahrer eine wertvolle

33 … erfahrenen Entwicklungsmannschaft, war das **Mercedes** Benz Anti-Blockier-System - kurz ABS – ausgereift,ABSolut funktionssicher

375 4MATIC modellen seriengemaessigen **Mercedes-ABS** bleibt deshalb immer voll erhalten

==== *FILE : /user/lkbssr/pt/kontext.results*

Das Mercedes-ABS regelt die Bremskraft schneller und zuverlaessiger, als **es** selbst dem erfahrensten Fahrer moeglich waere.

Gerade auf nassen oder vereisten Fahrbahnen oder in scharfen Kurven ist das **Mercedes-ABS** fuer den Fahrer eine wertvolle Hilfe.

Nach umfangreichen Versuchsreihen, ueber 35 Mio.Testkilometern, unter Einsatz der technischen Intelligenz einer erfahrenen Entwicklungsmannschaft, war das **Mercedes-Benz** Anti-Blockier-System - kurz **ABS** - ausgereift, ABSolut funktionssicher und bereit fuer den Einbau.

Die Wirkung des bei allen 4MATIC- Modellen serienmaessigen **Mercedes-ABS** bleibt deshalb immer voll erhalten.

Figure 5. Examples of collocation (Mercedes + ABS) located by KonText with line numbers indicated (top). By ‘clicking’ on the line numbers in the extract, samples of context may be elicited from the source text (bottom).

Knowledge engineering and terminology

In order to demonstrate the relationship between knowledge engineering and terminology, we have used a PROLOG-based system, MARVIN, to implement the definitions of a small number of automotive engineering terms using knowledge representation schemata, e.g. frames and objects. This exercise has encouraged us in our belief that terms ‘engineered’ in a knowledge base will reduce the duplication of repeated definitions in a termbank, e.g. the definition of each subtype necessarily contains the definition of the supertype; the definition of a part contains references to the whole; the use of knowledge representation schemata capable of ‘inheriting’ and ‘communicating’ will help to achieve an essential economy of definition. The definition has been analysed by seeking relationships with other terms: equivalence *(synonyms,* and *antonyms)*, hierarchy *(belongs_to),* part-whole *(is_part_of),* association (causes, *etc.),* material

(is_made_of), function *(carries_out)*, location *(is_placed_on)*. These relationships can be supported by MARVIN and are those often used in knowledge engineering. In addition to these facilities, MARVIN provides a 'concept browser' linked to appropriate images (Figure 6) (Holmes-Higgin, 1989).

A term which is implemented in a knowledge base will also help a knowledge engineer to define and implement a number of physical and conceptual objects of a given domain (Ahmad *et al.*, 1989).

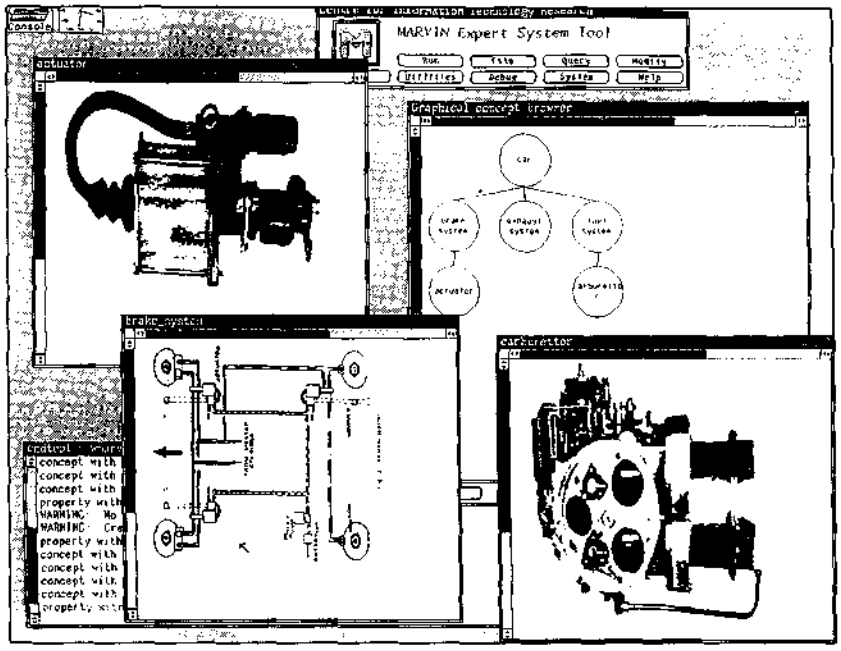


Figure 6. Concept browsing linked to diagrams

CONCLUSION

The Translator's Workbench Project aims to provide an integrated set of computerised aids for the translator, consistent with the current emphasis on machine-aided human translation. The University of Surrey's principal contribution is to combine corpus-based linguistic methods with computer-based tools inspired by aspects of computational linguistics in the development of a methodology for terminology building.

ACKNOWLEDGEMENTS

The support from the ESPRIT II Programme of the European Commission and from other partners working on the project is acknowledged. Particular thanks are due to Mr Khai Le-Hong and Mrs Edith Kroupa at Mercedes-Benz, Stuttgart, Professor Kurt Kohn and Mr Stefan Pooth at the University of Heidelberg, and Dr Gerhard Heyer and Dr Ralf Kese at Triumph-Adler, Nurnberg.

REFERENCES

Ahmad, K., Picht, H., Rogers, M. and Thomas, P. *Terminology and knowledge engineering: a symbiotic relationship.* University of Surrey, Guildford: Report CI-1, 1989.

Birrell, N.D. and Ould, M. A. *A practical handbook for software development.* Cambridge: Cambridge University Press, 1985.

Collins COBUILD English language dictionary. Ed.-in-Chief: J. McH. Sinclair. London and Glasgow. Collins, 1987.

Holmes-Higgin, P. *KonText user guide.* University of Surrey, Guildford: Report CI-2, 1989.

Holmes-Higgin, P. *MARVIN: the expert system tool user guide.* University of Surrey, Guildford: Internal Report, 1989.

Hughes, M. *The development of a bilingual terminology in automotive engineering (German to English): catalytic converters.* University of Surrey, Guildford: Report CI-1, 1989.

Leech, G. 'General Introduction'. In: *The computational analysis of English - A corpus-based approach,* eds. R. Garside, G. Leech and G. Sampson. London: Longman, 1987, pp. 1-15.

Longman's dictionary of contemporary English (LDOCE). London: Longman, 1987.

AUTHOR

Patricia Thomas, Department of Linguistic and International Studies, University of Surrey, Guildford, Surrey GU2 5XH, UK