[*Machine Translation Today: Translating and the Computer 15*. Papers at a conference…18-19 November 1993 (London: Aslib)]

NEW TRENDS IN TERMINOLOGY PROCESSING AND IMPLICATIONS FOR PRACTICAL TRANSLATION

Blaise Nkwenti-Azeh
Centre for Computational Linguistics, UMIST, Manchester UK

This paper examines how the changes currently taking place in terminology processing and documentation are related to the multilingual needs of translation, and also how progress in natural language processing in general, and terminology processing in particular, can contribute to the development of reliable, up-to-date terminology support tools for translators. The paper also describes some recent experiences in the automatic identification of terminological units from corpora. The paper concludes by identifying some specific areas in terminology software development which can benefit from the expertise of translators and other language professionals.

INTRODUCTION

Terminology is now firmly established and widely recognised as a distinct area of study concerned with the vocabulary of special subject languages, valiantly referred to as "Languages for Special Purposes" (LSP). Some scholars would even argue that terminology has attained 'discipline-status'. This identity manifests itself in at least three ways:

(i) the study of terminology is now backed by an established set of clearly defined theoretical assumptions (especially on the relationship between concepts, terms and extra-linguistic objects), methodological approaches and practical goals.

(ii) terminology now constitutes a separate component in an increasing number of translator training programmes; students in these courses are now given greater exposure to the methodological and practical aspects of terminology processing (i.e. terminography); many translation schools now offer regular seminars and short courses on terminology to professional translators.

(iii) several attempts have been made in the past and others are currently being pursued at national and international level to standardise the description of terminological items. We may mention, in this respect,

  (a) the pioneering efforts of the Nordic countries (1),
  (b) the ISO-led magnetic tape exchange format MATER (2), which was recently resurrected as MicroMATER (3) and taken on board by the Text-Encoding Initiative (TEI), albeit in much changed form (4); and,
  (c) two CEC-funded projects — EUROTRA-7 on the feasibility of standardised and reusable lexical resources (5), and MULTILEX on the definition of a multilingual standardised lexicon for the EC languages (6).

Terminological research is a time-consuming activity and occupies a considerable amount of time of specialised translation: estimates of up to 60% of total translator time have been cited in the literature. In a recent study reported in *Language International*, (7), translators have been found to spend between 20 and 42 minutes resolving a single terminological problem.

In the past, translators used dictionaries and other printed reference works for term equivalents; these were supplemented by personal collections of bilingual terminology. In some cases, industrial organisations compiled collections of terminology of product documentation to be used by in-house teams of translators and technical writers.

With the advent of computers and rapid advances in science and technology, the volume of technical literature has grown significantly; so too has the multilingual need for such information which is increasingly more complex and now requires greater specialised know-how or terminological research than even ten years ago. Consequently, a new range of computer-based lexical support tools has emerged (e.g. text databases, terminological data banks, CD-ROM dictionaries) in order to satisfy the LSP requirements of different groups of users.

In what follows I will focus on terminological data banks (or term banks, as they are more popularly known) which have evolved directly from the printed technical dictionary, and as such are most relevant to translators.

EVOLUTION OF TERM BANKS

Motivation for Term Bank Creation

Term banks have been intimately linked with translation since their inception in the mid 1960s and early 1970s. The earliest of these term banks were developed by translation departments in large organisations,

(a) to supplement printed dictionaries by providing up-to-date multilingual terminology;
(b) to preserve centrally the considerable effort of in-house language specialists, and to make this work more widely available;
(c) to permit greater terminological unity among translations split up among different translators by providing agreed, reliable and unified terminology;
(d) to speed up the translation process by giving the translator a single efficient reference tool.

In the past 10 years or so, we have witnessed a proliferation of term banks for research and commercial applications. More recently, term bank development tools have also been introduced for use in text-processing environments; these terminology-support tools are again aimed predominantly at translators.

The first of the 2 indicative lists below enumerates the well-known term banks and also some lesser-known ones. The approximate date of creation is entered alongside the early term banks.

List 1: A Sample of Term Bank Centres

- AMSI *(USA)*
- BATEM *(Quebec)*
- BD-TERM *(Switzerland)*
- BELGOTERM *(Belgium)*
- BTQ : 1973 *(Quebec)*
- BTB*(UK)*
- BTUC *(Chile)*
- BTUSB *(Venezuela)*
- CEZEAUTERM *(France)*
- CILF *(France)*
- DANTERM *(Denmark)*
- EURODICAUTOM : 1971 *(Luxembourg)*
- LEXIS : 1966 *(Germany)*
- NORMATERM : 1973 *(France)*
- NoTe *(Norway)*
- RUHRGAS *(Germany)*
- SURVIT *(UK)*
- TEAM : 1967 *(Germany)*
- TERMCAT *(Spain)*
- TERMDAT *(Switzerland)*
- TERMDOK: 1968 *(Sweden)*
- TERMIUM : 1975 *(Canada)*
- UZEI *(Basque Country)*

The second list is intended to give some idea of the range of products (or rather, product-names) available in the market (8).

List 2: Some Terminology Software Products

- Aquila
- Ascom
- Dicoterm
- Index
- INKTextTools
- Lingua-PC
- MicroCezeau
- Phenix
- Profilex
- Superlex
- Termex
- Term-PC
- TermTracer

Since translation is essentially concerned with interlingual equivalence/matching of units of meanings (as represented in a text), it is not surprising that the primary emphasis and sometimes overriding preoccupation in the majority of translator-oriented term banks, appears to be the documentation of foreign language equivalents. Also, as the relationship between terms and their corresponding concepts is generally assumed to be one-to-one, the problem of finding (or more precisely, selecting) linguistic equivalents in a target language is assumed not to be as difficult as for general language concepts.

The fact however is that, as far as specialised translation is concerned, the target-language equivalent must be supported by e.g., information on conceptual equivalence and contextual appropriacy. But, as far as terminology is concerned, multilingual equivalence is a secondary consideration when compared to, say, definition. The importance of definition is illustrated by the frequency of monolingual dictionary consultation during translation. For, where more than one foreign language equivalent exists, definitions are by far the most reliable disambiguation guide.

I will use the label *terminology-support tools* (TST) to refer collectively to term banks and term bank software.

The quantitative growth in terminology support tools has, unfortunately, not been matched by a significant change in quality. Qualitative changes have been in the form of making more information separately available, i.e. increasing access points; very little has changed by way of the information categories that are available in the database as a whole, e.g. the sort of information normally supplied by cross-references.

There are a number of key problem areas which developers of TSTs have to address if real progress is to be made in terminological knowledge representation. Some results from NLP and Lexical Data Processing are relevant in this respect.

In the Recent Developments which I shall review below, the general orientation is towards the establishment of a separate identity for terminological data bases as reference tools for specialised vocabularies, notwithstanding the specific requirements of any one user-group. The emphasis will mainly be on the incorporation of fundamental principles associated with special reference so that term banks (or the terminological lexicon) can provide the information required for the identification, fixation of reference, and correct use of the terms both in a monolingual and multilingual environment.

Progress in Term Bank Design

The evolution of Term Banks can be subdivided into 3 major phases or generations which broadly correspond to different levels of complexity of terminological description (i.e. incorporation of terminological principles and methods).

(i) The first generation started off as conventional data bank (i.e. *electronic dictionaries),* and incorporated little or no terminological theory. These 'term-oriented' data bases are the predominant type today and include EURODICAUTOM, TERMIUM, TEAM, and LEXIS.

(ii) The second generation of term banks incorporated some ideas of structure, notably, hierarchies. In spite of advances in computer data management, the few implementations of **'concept-oriented'** systems that exist include the Danish term bank (multi-disciplinary), the Norwegian Term Bank (oil terminology), CEZEAUTERM (initially soil mechanics), SURVIT (virology), and the British term bank prototype (multi-disciplinary).
Although this is a significant improvement over the first-generation term bank, the theory underlying the design of this generation database is inadequate to represent the diversity of terminological relationships for any one domain (e.g. *type_of, part_of, cause-effect, process-product, raw_material-product, succession, means_of_operation,* etc.

(iii) In the third generation of termbanks, currently still under development but already at an advanced stage, terminology is viewed as problem-oriented, specialised knowledge representation, and the terminological database is seen as an expert system for terminology. A prototypical example of this new generation of **'knowledge-oriented'** term banks is the knowledge acquisition tool, CODE (Conceptually Oriented Design Environment), which is being jointly developed at the University of Ottawa—Canada, by the School of Translation and Interpreting and the AI Laboratory of the Department of Computer Science (9). The CODE environment allows for explicit representation and subsequent retrieval of multidimensional relationships *(see Figure 1);* it is therefore a more realistic approximation of the conceptual complexity of the knowledge domain.

RETRIEVAL FACILITIES IN TERM BANKS

The range of queries that can be addressed at existing terminology-support tools is, in computational terms, minimal and very superficial. Within these environments, one can get responses only to simple queries such as spelling, usage (language variety, context, restrictions, etc.), foreign-language equivalent, definition, context of use, restrictions on use, bibliographic source, (other) subject(s) in which used, and synonyms/abbreviations, all of which require extraction of explicitly-coded information from within individual records, and access via the *main term* or other *index term.*

Because most TDBs still rely on conventional (or enhanced) relational database management systems for storage and retrieval, the 2-dimensional tabular representation of the model imposes restrictions on the information categories over the whole database. The uniform structure required by these packages means, for instance, that one cannot **elegantly** (i.e. without duplication) represent multifacetted or domain-specific relationships within the same multidisciplinary database. Assume the following entry (10) in one such database:

Lexical Entry :	arthritis
Def:	Any abnormality of a joint in which objective findings of heat, redness, swelling, tenderness, loss of motion, or deformity are present.
isa:	inflammation
g_affects:	joint
symptom:	heat/ redness/ swelling/ tenderness/ deformity/ loss_of_motion
g_affects_spec:	rheumatoid arthritis/ cricoarytenoid arthritis ...
cause_spec :	bacterial arthritis/ fungal arthritis/...
symptom_spec:	hemorrhagic arthritis/ deforming arthritis/...

It is difficult to represent the above relationships specific to the terminology of medicine alongside, say, those specific to automotive engineering and others specific, say, to information processing:

Relationships specific to medical concepts

- for diseases:
  - * *isa,*
  - * *g_affects,*
  - * *caused_by,*
  - * *has_symptom,*
  - * *transmitted_by,* etc.,

Relationships specific to automotive engineering concepts, e.g.

- for vehicles:
  - * *function,*
  - * *powered_by,*
  - * *transporting,*
  - * *medium,*
  - * *has_part,*
  - * *typical_agent,*
  - * *typical_size,* etc.

Relationships specific to information technology concepts, e.g.

- for storage media:
  - -- *recording_technology,*
  - -- *degree_of_writability,*
  - -- *physical_form,*
  - -- *content,* etc.

It should also be said that even the much publicised commercially available terminological packages (software and/or terminologies) only offer stop-gap solutions to the terminological needs of translators, and would need to be carefully hand-crafted to realistically handle complex terminological information. I will therefore exclude these when considering long-term solutions to the LSP needs of translators. Furthermore, processing textual information, for example, making 'string searches' in a particular field, is not straightforward because this function is generally not part of the software design and requires a separate program written to perform the task.

In a translation environment, users often require information of an inferential/evaluative nature as opposed to factual information, and which cannot be obtained in the majority of current systems, e.g.

(i) specific facets of interrelation: Which terms are related to Y (by *part, type, cause, process,* etc.)

(ii) nearest FL equivalent: What is the nearest foreign language equivalent for X?

(iii) contextual synonymy: Can term X be used in the context of Y?

(iv) conceptual environment: List the immediate conceptual information for X.

(v) functional aspects: What do you call a machine that does Y? Or, Has X got any parts? List them.

(vi) relational description: List all terms which have parts associated with them.

(vii) nature of interrelation: What is the relation between terms X and Y?

It is however doubtful whether general-purpose terminological reference tools will ever meet the LSP requirements of translators. Translators tend to specialise in a limited number of text types, e.g. legal texts, chemical texts, social legislation, medical texts, etc. Ironically, the areas where demand for translation is greatest, and therefore the expertise of language specialists is much sought after, are those where either

(a) the vocabulary is not yet consolidated, especially in the emerging disciplines, or

(b) the concepts are new to the language.

In the absence of up-to-date multilingual terminology records, translators will undoubtedly continue to be involved in

(i) term-creation, and more so in

(ii) systematic compilation of terminology from "grey literature", more of which is rapidly being assembled/made available in MR-form.

The terminology component of TR training should provide the necessary background for accomplishing task (i), via term-formation patterns. It should also provide the skills for identification and extraction of terminological units from texts in task (ii).

NEW DIRECTIONS IN TERMINOLOGY COMPILATION

Representational Aspects

A terminologically-oriented knowledge management system should facilitate the storage and retrieval of coherent collections of terms. A significant development with regard to representation is the terminology-specific software developed under the CODE system (9) which, as already mentioned, makes it possible to represent multi-hierarchical and multi-relational structures with minimal duplication of information (Figure 1).

Retrieval Aspects

The most significant development in human-assisted or machine-assisted terminography is the research into the use of an integrated package of terminographic and editing tools in the so-called "translator workstation" or "translator's workbench" (TWB).

A typical TWB should, among other things, provide translators with an integrated package of computerised terminology tools in a MAHT environment, with facilities for multilingual text processing, (remote) access to non-resident term banks and other terminological support tools (including other machine-assisted translation systems), and dynamic terminology management (i.e. machine-assisted creation/acquisition, extension and maintenance of collections of terminology).

Unfortunately, the term-acquisition modules currently being developed within the integrated translator workbench environments embody little terminological knowledge. They are inherently unsatisfactory because they have not resulted from a study of the term-formation and other sublanguage characteristics of the domains in which they are intended to be used.

Term Identification

Ideally, the extraction of terms from a machine-readable corpus should be performed automatically, if we are to benefit from the speed and consistency (NB: not **Accuracy**) which computational tools provide. Researchers are currently investigating various "semi-automatic" and "automatic" ways of identifying potential terminological units. Some collocational-type methods have already been incorporated in TWBs (11).

We at CCL have recently been examining the use of positional information of lexical items in term identification, using corpus texts and terms, e.g. from the field of satellite communications (12). This terminology-oriented method exploits the regularities in term formation which are characteristic of each special subject.

The work so far has focused on identifying positional values from existing term lists and using this information to extract new units from a corpus.
For example, if the input dictionary contains the terms:

- *frequency assignment*
- *carrier frequency*
- *constant frequency assignment*
- *available bandwidth*

the term-identification program should, and does in fact, recover the terms

⇒ *carrier frequency assignment*

⇒ *available frequency bandwidth.*

Using a list of approximately 600 terms manually extracted from a 50-page telecommunications text corpus, we have, for example, been able to automatically extract over 400 new potential terminological units from the same corpus. The list below shows examples of extracted terminological units having *satellite* as element:

List 3: Automatically extracted terms

- ionosphere sounding satellite
- justified satellite link
- land mobile satellite service
- long intersatellite link
- low altitude observation satellite
- low orbiting satellite
- major path satellite
- maritime mobile satellite service
- maritime radionavigation satellite service
- maritime satellite
- narrow beam satellite antenna
- near antipodal reverse frequency assignment satellites
- radionavigation satellite service
- artificial satellite
- operational global satellite communications system
- complete satellite communications networks
- recurrent earth track satellite
- reflecting satellite
- satellite antenna gain
- satellite antenna polarization
- satellite antenna radiation pattern
- satellite antenna reference pattern
- satellite antenna reference radiation pattern
- satellite redundancy
- satellite repeater

One of the significant aspects of the methodology (apart from the fact that identification is so far fully automatic,) is the fact that it can ensure comprehensive coverage of all the term combinations in a given paradigm. Term identification is not as straightforward as it may seem (anyone who has done thematic terminology research will attest to this). The following examples highlight some of the problems with the positional approach, namely, inclusion/extraction of non-term compounds:

List 4: Errors in term identification

- civil time
- magnetic disturbance
  ⇒ civil disturbance
- complete reflective surface
- digital information
  ⇒ complete information
- correct check
- picture information
  ⇒ correct information
- key pulsing signal
- picture element
  ⇒ key element
- aerodynamic force
- natural noise
  ⇒ natural forces
- outgoing country
- single sideband
  ⇒ single country

Eventually we hope, of course, to be able to do away with an input term corpus altogether, and to minimise the incidence of non-terminological units, by incorporating statistical, lexical-semantic and other parameters in the identification program.

Part of the problem lies in the fact that a good knowledge of the domain is often necessary especially if general language words have specialised usage within the domain—as simple terms or in combination with other lexical items (general language words and special language term elements) to form compound terms. Any automatically-generated term list would therefore necessarily have to be post-edited by a human specialist.

NLP-ORIENTED TERMINOGRAPHY

From the earlier summary, it emerges that the main changes in terminographic orientation over the past few years have been from word-based to concept-based

systems, and from technology-influenced, database-dictated, inflexible structures to conceptually-motivated, dynamically-generated systems.

There are also significant methodological changes currently taking place in the field of NLP.

Firstly, MT system developers are now moving away from the pure rule-based approach—which has been characteristic of the domain over the last decades—, towards empirical (corpus-based or example-based) approaches which make direct use of information extracted from large corpus resources (typically parallel/translated texts), or hybrid approaches which consist of a rule-based core and add-on empirical modules (13).

Secondly, there is broad agreement on the need for separate GL and SL lexicon modules (or at least for different types of information for lexical and terminological entries) and for the need to incorporate sublanguage-specific information as an integral part of the grammar and lexicons of these systems (14).

The recent multinational efforts towards definition of standards for NLP lexicon description, in particular, Eurotra-7 (1990-91) and MULTILEX (1991-92) merit special consideration here because I consider them to be of particular significance to translators, if we take the view that the integrated translation environment will be the setting for the future.

The Eurotra-7 Study (5) identified two main categories of standards depending on their object: the contents of linguistic description and its representation. The study concluded in its Final Report (10), inter alia, that:

> "Within descriptive linguistics, different theories and descriptive models are basically interested in the same phenomena, but they classify the phenomena in different ways...; such classifications of individual objects of an observational domain allow for different, even *a priori* incompatible generalizations" (p.72).

The authors of the report recommended that research in general language and sublanguage be carried in parallel as it would then allow to answer the following questions:

> "- to what extent can we share descriptive devices between general language and sublanguage?
>
> - how can the peculiarities of sublanguage which are usefully described in terms of restrictions, deviations and preferences with respect to knowledge about general language items, be best accounted for in a formal linguistic specification?" (p. 112).

With respect to representational standards, the MULTILEX project (a follow-on from the Eurotra-7 study) description (10) is based on the assumption that

> "the same format/formalism can be used in SL and GL. It seems useful in order to accommodate descriptions from a whole range of sublanguages and from general language, to have one common representation or to have means of combining several representations." (p. 19)

The above re-orientation of NLP and re-definition of its components opens the way for translators and other language professionals—who have so far been marginalised in the development of MT grammars—, to play a greater role in helping computational linguists and computer scientists identify areas of potential

translational problems and formulate rules for resolving these issues. The pragmatic experience of translators can be deployed so that statistically-based preference mechanisms are more consistent with those of particular micro-environments. We can equally rely on these professionals to provide realistic descriptions of language and equivalence which they encounter in routine work in actual texts, rather than relying solely on the intuition of computer scientists or other 'non-language-professional' grammar writers.

CONCLUSIONS

(1) Translators must therefore learn to separate terms from words, identify compounds or other juxtapositions which may be single units or casual collocations, recognise variants and have criteria for finding the standard form, etc.

(2) Being able to learn and to recognise that they are dealing with a term rather than a word, will narrow down the search space in the reference works to be consulted.

(3) Translators/interpreters also need to know where there are conceptual or terminological incompatibilities between their working languages so that they know when a paraphrase or a neologism is necessary. Such incompatibility can only be identified at either by comparing or through a knowledge of the conceptual structures of the subject field in the different languages.

(4) Although term banks and other multilingual terminological reference tools are aimed primarily at translators, the contribution of the latter in the logical structure and content of these products has so far been marginal (apart, of course, from the use of translators' terminology cards to build up the collections).

(5) As the terminology requirements of NLP and MAHT/HAMT converge, translators will be called upon to play a greater role in lexicon design by providing NLP programs with various types of structural information (decoding, parsing, disambiguation, interlingual mapping, creation of new lexical items, etc.).

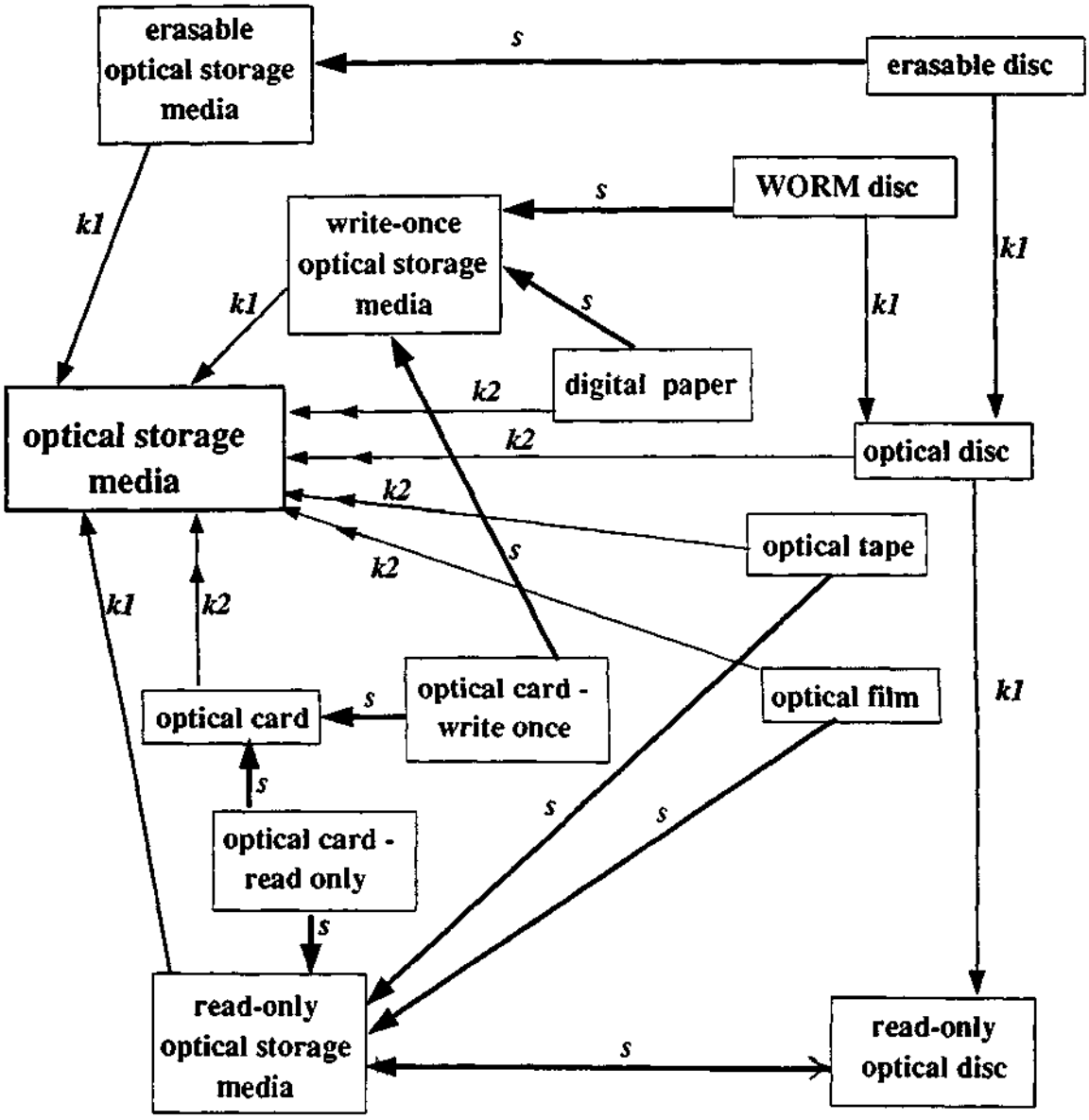
(6) Also, as up-to-date terminological reference tools become increasingly available mainly in MR form and as part of an integrated translation system, translators will not only need to be able to evaluate the utility of terminological products for particular operations. More importantly, translators must be able to construct, and/or update and maintain such reference tools in a way which enhances the quality and the sharing of information. They should be capable of choosing the most appropriate medium for this data (CD, disk, paper, etc.).

(7) Computerized terminography and the availability of databases of parallel texts offer opportunities for extensive coding of text-type-specific contextual information on terms, their textual variants and foreign language equivalents. Contextual information will in future constitute a more central component of the description of terminological items. In fact, the function of

manually-entered definitions may have to be re-assessed if elaborate systems of terminological relationships are represented in the terminological database, and the facility exists for automatically-generating terminological definitions from these and other information fields.

(8) Data-preparation requires enormous human resources and is therefore uneconomical for small-scale organisations. But, with the availability of term-identification tools, systematic collections can quickly be assembled from MR data. Significant economies can be made by being able to look up all/most potential terminological units before embarking on the text-conversion task itself. In these circumstances, the role of the language professional would typically involve development of firm-specific terminology collections, and evaluation and recommendation of commercial packages. In order to carry out any meaningful evaluation, they have to have knowledge of such benchmarks as user-friendliness, relevance of information, completeness, flexibility, etc.

(9) Finally, it is well-known that translators have a distrust of theory or theorising. In order for any of the above goals to be attained, we need first of all to convince translators that the solution of practical translation problems is assisted by an understanding of the underlying principles of terminology and that a sound methodology for developing terminology must also be based on the same theoretical foundation.



Partial Graphical Representation of relationships in the *media* subfield: concepts with OPTICAL STORAGE MEDIA as superordinate (9)

Note: *k* = multi-dimensionality s = normal subconcept

Figure 1: Multidimensional relationships in a terminological knowledge base

REFERENCES

1. Picht, H., and Draskau, J., (eds.), 1985. TermNet News 12/1985. Special Issue on the Nordic Countries.

2. ISO, 1984. Magnetic tape exchange format for terminological/lexicographical records (MATER), ISO/DIS 5156, Geneva.

3. Melby, A., 1991. "MicroMATER: A proposed standard format for exchanging lexical/terminological data files", META 36/1, March 1991,135-160.

4. Sperberg-McQueen, C.M., and Burnard, L. (eds.), 1990. "Guidelines for the Encoding and Interchange of Machine Readable Texts", TEI P1, Draft Version 1.1., Chicago & Oxford: 1 November 1990.

5. Heid, U., and McNaught, J., 1991. "EUROTRA-7 Study: Feasibility and project definition study of the reusability of lexical and terminological resources in computerized applications", Final Report, Submitted to the CEC. IMS-Stuttgart, August 1991.

6. McNaught, J., and Smith, S. (eds.), 1992. "MULTILEX: Definition of the Standard Monolingual Description of Lexical Items", ESPRIT Project 5304 (MULTILEX: A Multilingual Standardized Lexicon for the European Community Languages), Final Report, Submitted to CEC.

7. Language International 4/1 (1992), p.28

8. Pulitano, D., and de Besse, B., 1989. "COMPUTERM: Banques de données terminologiques et systèmes de traduction assistée par ordinateur, outils de la bureautique moderne", Aperçu du marché. Symposium CompuTerm, 27 septembre 1989, Bâle.

9. Meyer, I., Bowker, L., and Eck, K, 1992. "COGNITERM: An Experiment in Building a Terminological Knowledge Base", Proc. 5th Euralex Int. Congress, 4-9 August 1992, Tampere, Finland.

10. Martin, W., ten Pas, E., Demeersseman, H., and others, 1992. "The terminological description of lexical items in MULTILEX", 3rd review report, May 1992, Submitted to CEC.

11. Ahmad, K, Fulford, H., Holmes-Higgin, P., and others, 1990. "The Translator's Workbench Project", Translating and the Computer 11, C. Picken (ed.), 9-19.

12. Nkwenti-Azeh, B., 1993. "Exploiting term-formation regularities in automatic term recognition", CCL/UMIST Report 93/5.

13. Sadler, L., 1992. "Rule-Based Translation as Constraint Resolution", Proc. FGNLP Workshop, S. Ananiadou (ed.), 1-21.

14. Tsujii, J., Ananiadou, S., Carroll, J., and Phillips, J., 1990. "Methodologies for Development of Sublanguage MT Systems", CCL/UMIST Report 90/10.