Christian Galinski
International Information Centre
for Terminology (Infoterm)
Vienna, Austria

Terminology and standardization under a machine translation perspective

This position paper considers MT on the basis of the experience gained in standardization at international level within the frameworks of ISO/TC 37 "Terminology (principles and co-ordination)", ISO/TC 46 "Documentation" and ISO/IEC JTC 1 "Information technology" as well as in terminology (science and work) in general and terminology unification and standardization in particular over the last decade. Because of its highly condensed presentation it is rather axiomatic in nature. It outlines the interfacing areas between MT on the one hand and terminology and standardization on the other hand, where immediate positive results can be expected.

1 The international environment of machine-translation

More or less well-founded criticism and even self-criticism of machine-translation (MT) abounds these days especially with regard to fully automatic machine-translation (FAMT). This contribution is not attempting another 'MT bashing', but rather tries to point out some aspects (focussing on terminology and standardization) which might help to render MT more effective.

There is a growing awareness of the fact that a number of methodology-oriented subject-fields, such as information science (and its practical application in documentation), terminology science (and its practical application in terminology work and terminography), language for specialized purposes (LSP) research (and its practical application in technical writing) translation studies (and its practical application in specialized translation) and to a certain extent management theory (focussing on information, communication and quality management) should be founded on a common interdisciplinary theoretical and methodological basis referring to

- *specialized* knowledge
- *specialized* information
- *specialized* texts
- *specialized* languages.

Dealing with such phenomena led to the diversification into a great variety of specializations such as

- specialized translation
- machine translation
- data modelling and database design
- knowledge engineering
- computer-assisted terminography
- technical documentation
- scientific writing
- information (resource) management
- communication management (in enterprises)
- quality management (especially in services).

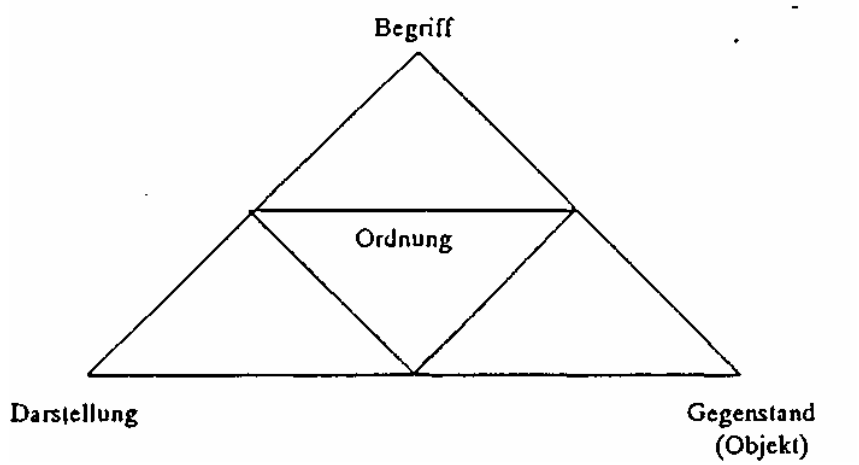
In addition today's world, which is driven by science and technology, is constantly moving towards a *world-wide multilingual information and communication society,* the so-called *global village. 'Innovation'* is increasingly dependent on the efficient use of information (from many countries in different languages). *'Global marketing'* of products and services can only be effective, if based on *multilingual approaches.*

This may have an impact on the approaches to MT, especially on the focus of its application (e.g. multilingual technical documentation) and on the design of (inevitably) sub-optimal systems (which might yield optimal results given a well-defined focus of application) as well as on new development areas supplementing hitherto's MT developments enhancing their effectiveness.

2 Fundamentals of terminology

'Terminologies' (i.e. 'concepts' within their respective 'concept system' and their 'representation') are the object of investigation of terminology science which also deals with terminological phraseology and the use of these terminologies in actual texts (discourse). Specialized concepts are represented not only by linguistic means (words etc.) but also by graphical symbols and other non-linguistic signs.

Figure 1: The 'terminological triangle'



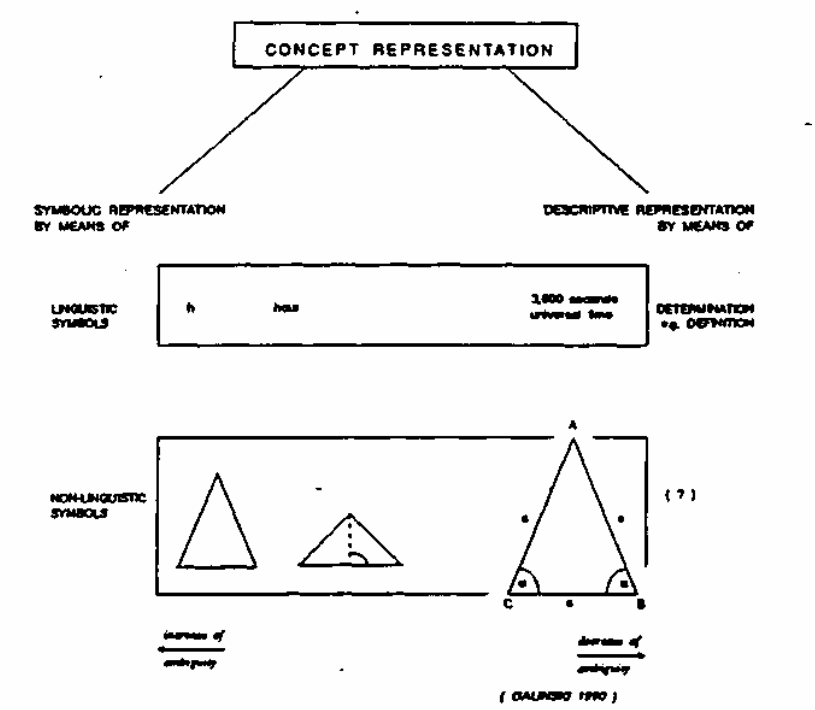
Terminology work properly carried out by subject-field specialists results in a *systematic representation* of the respective subject-field's knowledge at the level of concepts. *Terminography* comprises the activities to select, record and process terminological data. Definitions for these and other basic concepts of terminology are given in the International Standard ISO 1087 "Vocabulary of Terminology".

2.1 Concept representations

Given the hypothesis that concepts are the smallest units ('atoms') of specialized knowledge, the concept systems correspond to the respective theories and structure of the 'objects' of the subject-field in question. Basically concepts can be represented

- **by *symbolic representations,*** such as
  terms (and other linguistic symbols incl. abbreviations, alphanumeric codes etc.);
  graphical symbols (and other non-linguistic symbols);
  combinations of the above;
- **by *descriptive representations,*** such as
  definitions, explanations and other linguistic descriptions;
  (complex) graphemes, formula etc.;
  combinations of the above.

Figure 2: Different kinds of concept representation



Concepts are thus considered as
- ***units of thought*** (for recognizing 'objects' as part of reality);
- ***units of knowledge*** (for ordering these objects and related knowledge/information);
- ***units of communication*** (facilitating the representation and transfer of knowledge within a subject field or between subject fields).

Under this approach ***computer-assisted terminography*** has discovered a variety of applications, such as information, knowledge and language engineering, and more practical activities such as

- computer-assisted technical writing (concentrating on terminology, terminological phraseology, document management, text management);
- computer-assisted indexing and abstracting;
- computer-assisted conferencing (especially at conferences/meetings with a harmonizing objective incl. the function of a conference terminologist);
- project-concomitant terminology documentation etc.

Increasingly such applications show the main emphasis being laid on knowledge organization and information management aspects.

MT development projects would certainly profit from adopting and - if necessary - adapting some of the methods of terminography to enhance the quality and consistency of its components, in particular its knowledge base.

2.2 Terminology and language

The traditional distinction between general language and specialized language is still of some value for the analysis of texts to be translated by MT applications. Basically all texts to be translated automatically contain technical terms and words from general language. There is a continuous influx of general language words into specialized language ("terminologization") and vice versa ("popularization"). Nevertheless the steel terminology of a given country, for example, depends above all on the conceptualization by engineers of the composition of the raw material influencing the characteristics of the steel as well as the methods for producing and processing this steel. It is certainly not primarily a feature of language development. For decades now. the development of many languages has been influenced by the development of domain terminologies. Thus the "creativité de la langue" (creativity of language) is based on and driven by the conceptual creativity of scientists and other subject specialists, who in this way have also contributed to language development.

Terminologies are primary carriers of scientific-technical knowledge in texts. As a result of the proliferation of knowledge and due to the fact that science and technology are very much sectorized, the domain specific languages (in particular their terminologies) are developing in different ways and at different paces. Knowledge of the respective changes in the conceptual (semantic) base is essential in building MT knowledge bases for each domain. Social science and legal texts are particularly tricky in this respect because of the 'double identity' of many terms: they are at the same time technical terms representing a specialized concept and general words with rather fuzzy meanings. A thorough terminological analysis of such texts is an efficient method in determining the role of terminological elements in textual structures.

Due to the limited number of term elements that can be used to designate new concepts by means of terms, every language (including English) faces difficulties to provide a sufficient number of unambiguous terms for communication in specialized languages. Subject specialists, therefore, who are the main 'inventors' and users of 'their' respective specialized language, are at the same time also creating 'linguistic' problems (in the form of homonymy,

synonymy and quasi-synonymy). They are then trying to solve these problems via terminology work and particularly through terminology standardization.

2.3 Terminology and knowledge ordering

Macro-structures of knowledge are used as documentation languages, such as classification schemes or documentation thesauri for the purpose of ordering information. They are used in terminology databases and all types of databases for efficient (conceptual) information retrieval.

Thesauri can be efficiently used in MT systems in structuring knowledge bases and in enhancing information retrieval mechanisms both for analysis and text synthesis stages. Full-text search methods can be complemented by such concept-based retrieval techniques.

3 Specialized texts

3.1 Semiotic characteristics of specialized texts

Specialized texts are representations of specialized knowledge and information by means of terminological units and - depending on the subject field treated - a smaller or larger number of non-linguistic representations (e.g. formulae, flow charts, figures, pictures, diagrams etc.). The information which various kinds of users would like to draw from specialized texts consists of

- facts
- statements
- other kinds of knowledge units.

Terminological units are the main building blocks of the representation of specialized information and knowledge in specialized texts. There is no way to 'retrieve' these facts, statements and other knowledge units in a systematic way, if not by means of terminological units.

3.2 Sorts of texts

A specialized text is - or should be - written with specific user groups and/or with a special purpose and audience in mind. This decides on the 'sort of text', to which the specialized text in question belongs. Text linguistics is occupied with the study of various sorts of texts, which in many cases represent groups of types of text belonging to the same sort of text. This has an impact on text management facilitating a highly effective and efficient preparation and production of specialized texts. Here is also much room left for standardization with regard to

- the reduction of linguistic phenomena (viz. terminological and stylistic variation)
- the quality of the text (incl. aspects of consistency and homogeneity)
- the way and layout of representing specialized information and knowledge etc.

No need to stress that such standardization efforts would have a beneficial effect on the translatability of specialized texts, whether by human translation (HT) or by MT. There is a difference to 'controlled language' insofar as a restriction to one (or few) subject-field(s) and one (or few) sort(s) of text in terminology aims at completeness, whereas controlled language as a rule aims at the reduction of linguistic phenomena across a certain number of subject-fields and sorts of text. The two approaches can converge, if controlled language is applied to a very limited number of subject-fields and sorts of text

3.3 SGML view of text

On the one hand every kind of symbolic representation of information and knowledge according to modern text linguistics has to be considered as a text, including non-linguistic elements. Since parts of a text according to SGML philosophy can again represent identifiable texts, such a text can, in the extreme, consist of

- one symbol only
- non-linguistic representations only.

On the other hand every text also constitutes a document. This connection between text linguistics and documentation has not yet sufficiently been explored.

The above-mentioned text linguistic and documentational aspects of specialized texts very much influence our understanding of 'text', of what we can do with it and *how* to go about it. They may have an impact on the the way texts could/should be prepared with MT in mind as well as on the future re-use of machine-translated texts for a variety of other purposes (e.g. in databases containing facts, statements, terminological and phraseological units etc.).

4 Terminology and documentation (T&D)

'Terminology & Documentation' (T&D) is a new concept, which emerged in the mid-eighties from terminological activities (such as terminology work and terminography). It responds to the need to integrate advanced documentation methods into advanced terminological methods in order to structure terminology data banks designed to process terminological data as well as related bibliographic and factual data for knowledge data processing at the level of concept knowledge.

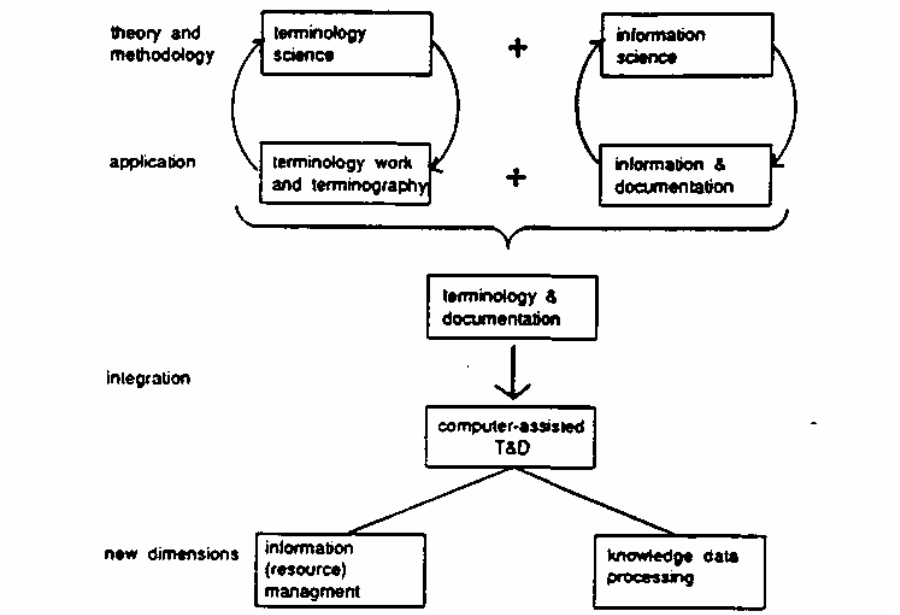
4.1 From 'terminological documentation' to 'terminology & documentation'

In the early sixties of this century UNESCO recognized the close relationship between terminology and documentation. At the end of the sixties 'terminological documentation' became one of the aspects of UNESCO's UNISIST Programme and was a frequent item of discussion at various conferences and meetings. From the "UNISIST decade" (1969 onwards) until today the meaning of 'terminology documentation', as it was later called, has differentiated into several concepts due to a variety of applications:

a) documentation of terminological information → (later:) terminological lexicography → (today:) terminography (covering all aspects of selecting, recording and processing terminological data);

b) terminological documentation → (today:) documentation in the field of terminology (covering the recording of reference data such as bibliographic data on documents as well as factual data on terminological activities, institutions etc. in the field of terminology);
c) documentation for terminology work (which basically means the coding of reference data in terminology work and terminography and linking these data to the respective full information in other files or databases);
d) terminology & documentation - T&D.

Figure 3: Terminology & documentation



T&D is the interdisciplinary integration of the methods of terminology science and those of documentation (coined in parallel to I&D - information & documentation). It is difficult to imagine, how T&D could be neglected in any activity related to knowledge processing.

4.2 Relation between terminology and documentation

During the seventies it became obvious that bibliographical data and documentation methods are indispensable in nearly all kinds of terminological activity. Terminology work and terminography cannot do without information & documentation (I&D) methods, if its aim is to produce high quality and reliable data. But efficient I&D cannot do without terminological methods and data either. Terminology science and information science are closely linked in their practical applications. On the one hand terminography can be regarded as a special kind of handling of factual information on the micro-structural level of knowledge. Terminological records (being 'documents' from the formal point of view) need a macro-structure in order to remain manageable in large quantities - as it is the case with bibliographic records. As a rule documentation languages, such as classification schemes or documentation thesauri, indicate the subject-field or sub-field to which a given concept belongs.

The information on the sources from which terminological data have been taken are of particular importance in a terminological entry/record. This information is of particular relevance for the evaluation of the reliability of the terminological data in question. All terminology data banks, therefore, need and in fact are designed to handle documentary data proper (i.e. references to sources, viz. documents). Documentation for terminology work provides the source information in the (coded or not coded) form required in terminological entries/records.

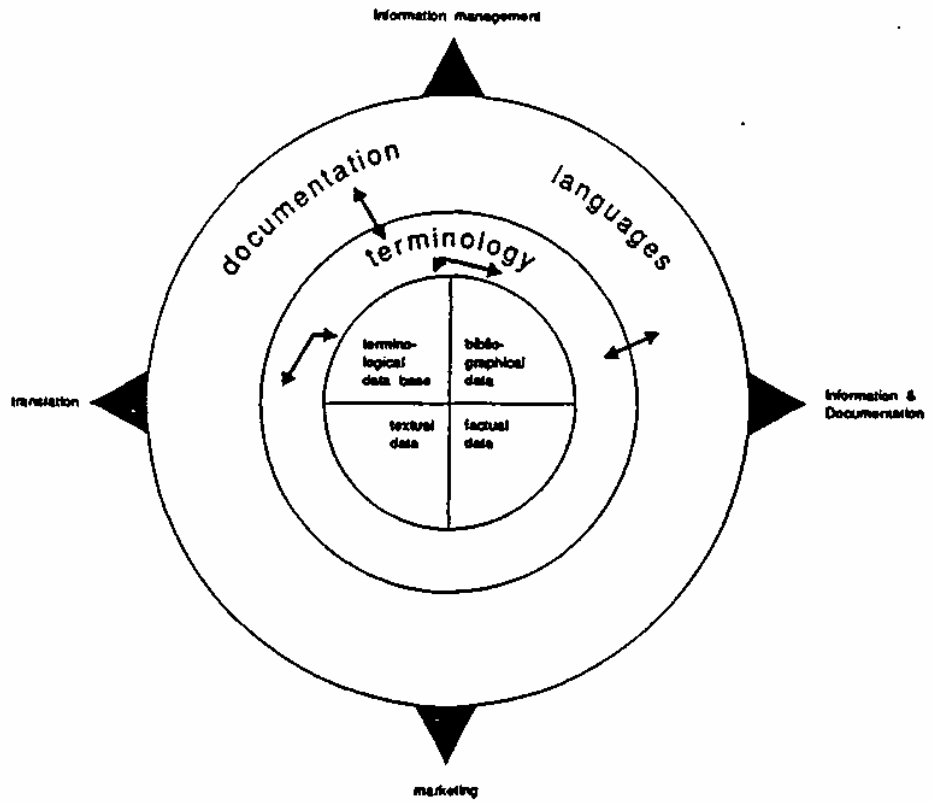
4.3 Documentation languages

Moreover, terminological methods are applied in the process of preparing a documentation language or its adaptation to a given situation. Nevertheless, the terminology of a subject field or sub-field and a documentation language concerning the same domain should not be

- confused, although they may look very similar from the linguistic point of view,
- merged, because they have distinct functions, which make them mutually incompatible (although complementary in their use).

As is the case with terminologies, documentation languages are also organized and internally structured according to the topics/subjects of the subject-fields in question. They are 'theme systems' (consisting of theme concepts) reflecting views on a given subject-field corresponding to a given purpose. Several different documentation languages may, therefore, be applied side by side in a given complex information system for different purposes.

Figure 4: Interrelation between terminology and documentation languages



4.4 Combination of terminology and documentation languages

The combination of terminology and documentation languages for application in
- text management
- speed learning
- indexing and abstracting
- information retrieval (especially in full-text databases)
- specialized language teaching/learning

and last but not least in
- MT

offers promising solutions to the problem of subdividing large quantities of different types of (whether linguistic or non-linguistic textual) data on the one side and of tying together sets of such data - as they occur in specialized texts - under the umbrella of 'themes', which may vary according to the view on the data (by creators or users of data).

As structure is at least as important as the representations of the individual entities of both terminologies and documentation languages, the above-mentioned combination offers a - in principle - language-independent approach to information processing of textual data in the broad sense as described above. Thus the difficulties of an approach oriented towards language-pairs can be avoided.

5 Terminology in MT systems

It seems more and more obvious to MT users and system developers to focus on terminological problems, since MT is primarily geared towards the translation of specialized texts, with its terminology and all its extra-linguistic and above all extra-textual dimensions.

Moreover, it is increasingly recognized today that for reasons of cost-efficiency terminological data should be recorded for multiple purposes, viz. multi-functionally. Everybody active in terminology today knows that the terminology problem cannot be solved without the cooperation of producers and different users of terminologies. No MT user can solve it alone without cooperation with external partners. To deny this fact could result in a deception of the MT customer who expects the applicability of a MT system on the spot in his/her work environment (comprising also the respective SPLs).

6 Standardization and MT

According to ISO/IEC Guide 2 "standardization" is the ***"activity of establishing, with regard to actual or potential problems, provisions for common and repeated use, aimed at the achievement of the optimum degree of order in a given context".*** Note 2 to this definition reads ***"Important benefits of standardization are improvements of the suitability of products, processes and services for their intended purposes, [.....] and facilitation of technological cooperation".*** In a broad sense standards facilitate communication and the exchange of goods and services by means of rules in order to avoid unnecessary efforts (thus increasing effectiveness and efficiency). As a rule standardization, viz. the preparation of standards at national, regional or international levels, is a cooperative and quite democratic activity based on the state-of-the-art of development of technology and methodology, and

creates rules by consensus. Thus standardization is oriented towards feasibility, practicality and appropriateness rather than an ideal state.

In the social sciences and humanities there exist deep-rooted sentiment against standardization, partly due to social and intellectual attitudes. In technology the benefits of standardization are widely accepted. Applied to terminology, standards enhance

- the consistency and coherence of large amounts of terminological data
- an adequate quality of the data
- the compatibility of data categories and data structures
- the exchangeability of data
- the possibility to re-use the data for other purposes
- the comparability between data in different languages (including transparence with regard to justified differences) etc.

Standards prepared by technical committees, such as ISO/IEC JTC 1 "Information technology", ISO/TC 37 "Terminology (principles and co-ordination)", ISO/TC 46 "Documentation" of the International Standards Organization (ISO) should be taken into account by MT developers in order to improve quality standards.

ISO/TC 37 has repeatedly suggested
- the standardization of additional terminology-related requirements from the MT point of view
- the standardization of suitable lexicographical principles and methods
- joint cooperative efforts with regard to the preparation of reliable terminological data and to take into account standards with regard to record formats, exchange formats, database design, interfaces, conversion routines, text-based terminology extraction methods etc. in the field of terminology. If existing standards prove inadequate or insufficient for MT purposes, they can be amended or supplemented.

Conclusion

An integration of terminology management into MT systems would contribute to consistency, enhanced validity of MT knowledge bases, and their performance in text analysis (of the source text) as well as text synthesis (of the target text) for specialized text types.

Possible items of standardization that would (directly or indirectly) have a positive impact on MT:

- Extension of ISO DIS 12200 (SGML-based Terminology Interchange Format) for machine-translation dictionaries
- Link to lexicographical data interchange format (LIF)
- Machine-translation terminological - lexicographical data element directory
- Phraseology and collocation data element directory
- Translation memory requirements and specifications
- MT-oriented SGML-based text analysis
- Computer-assisted software documentation for TDBs
- Object-oriented data in MT (e.g. encyclopedic knowledge data, facts, propositions)
- Basic semantic repository for meta data catalogues
- Application of documentation languages in machine-translation

etc.