

Terms in Texts and the Challenge for Terminology Management

KLAUS-DIRK SCHMITZ

Cologne University of Applied Sciences

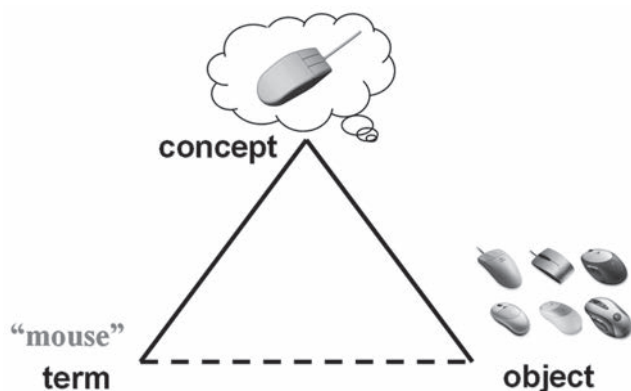
KEYWORDS: terminology management, term extraction, concept orientation, term autonomy, terminological entry, synonymy, ambiguity

1. BASIC THEORETICAL FOUNDATIONS

Terminology is defined as the “set of designations belonging to one special language” (ISO 1087–1:2000). This international definition refers more to the representation part of special language communication and ignores the conceptual view behind the designations. A more comprehensive definition is given in the German terminology standard DIN 2342:2011 where “terminology is the set or inventory of concepts and their representations in a specific subject field”. This definition not only includes the concept as an important aspect of terminology; it also shows a broader view to the concept representation side of terminology by not limiting it to terms or other language-related designations and by including symbols, icons, gestures or any other multimedia representations.

In order to explain the relation between concept and term, terminology theorists and researchers adapted the so-called “semiotic triangle” introduced by the American linguists C. K. Ogden and I. A. Richards (Ogden, Richards 1923: 11) to explain the relationship between concepts and terms (Figure 1). The triangle has undergone a long history of modifications and interpretations, and has also been attacked from several quarters as an over-simplification or misrepresentation of the complex relationships that exist between concepts and terms. Despite the criticisms that have been leveled against the triangle, its simplicity makes it an excellent tool for illustrating concept-term relationships to people who are just beginning to identify terms in texts and to create terminological data entries to document them. Other and more complex models are described in Arntz, Picht, Schmitz (2014: 41 ff.).

Figure 1. Terminological triangle (based on Ogden, Richards 1923)



Concepts are “cognitive representatives” (“gedankliche Vertreter” (Felber, Budin 1989: 69)) for objects, that arise out of the fact that humans recognize the common characteristics that exist in a majority of individual objects of the same type, and then store these characteristics and use them to impose order on the world of objects, in order to achieve mutual understanding when they communicate with other people. ISO 1087–1:2000 defines a concept as a “unit of knowledge created by a unique combination of characteristics” and DIN 2342:2011 describes a concept again more explicitly as a “unit of thinking made up of characteristics that are derived by categorizing objects having a number of identical properties.”

Both standards state in a note that concepts are not necessarily bound to specific languages, but the cultural, social and technical background of the human beings who generate the concepts and the environments in which the concepts are used affect the way they manifest themselves in any given situation. Regional differences within a language community (e.g. Germany and Austria) may lead to different conceptual orientations for the same term, whereas one cultural community where different languages are spoken (e.g. Switzerland) may allocate the identical concept represented by several terms in different languages.

Since concepts are mental or cognitive representations, we need definitions to explain and describe concepts. In most cases, definitions refer to other concepts (e.g. the superordinate concept) and mention specific characteristics that are unique and typical for the concept to be defined. On the basis of definitions and concept characteristics, it is possible to relate

concepts to each other and to construct terminological concept systems, taxonomies, ontologies, or other knowledge organization systems (SKOS). These concept systems represent the knowledge of a domain or sub-domain in a systematic way.

The term is defined in ISO 1087–1:2000 as a “verbal designation of a general concept in a specific subject field” or in DIN 2342:2011 as a “designation of a defined concept in a special language by a linguistic expression.” The term serves as the representation of the concept and we can write it down, say it out loud and use it for communication. We use the word “designation” as a superordinate concept when we talk about terms because there are also other ways to represent concepts, e.g. ones that aren’t necessarily made up of words, such as symbols, formulas, pictograms, gestures, etc.

Some terms consist of more than one word. These terms are called multi-word terms or compounds, e.g. “printer with single-sheet feed.” The way words combine to form terms varies from language to language. When dealing with terms, the linguistic side of terminology work comes into the game.

We have seen that the term is the verbal representation of the concept. In special or technical language, it is highly desirable that this relationship be unambiguous, even without contextual reference, which means that one term should be assigned to one concept, creating a condition called univocality. When this condition prevails, the meaning of terms is completely clear, even if the term appears without any explanatory context. Of course, this ideal situation is difficult to achieve or enforce. Two problems involving term–concept assignment recur frequently, even in technical and scientific texts:

Synonymy exists if two or more terms in a given language represent the same concept. Thus a synonym is a term used to designate the same concept as another term. Even though synonymy can compromise communication between experts, it occurs quite frequently in practice. This can happen especially in subject fields where many objects and concepts are still undergoing development. In these kinds of dynamic fields, competing terms are used in parallel until unambiguous terms are gradually established, either through a natural selection process or by conscious standardization. Even when people are quite aware of these problems, variants can remain in use for long periods of time, based on such factors

as natural regional variation or, quite intentionally, company or product-specific efforts to use terminological differences as one means of positioning a product in the market.

Homonymy involves the opposite situation from synonymy: here a term or several terms that have the same external form refer to several concepts¹. ISO 1087:2000 defines homonymy as a relation between designations and concepts in a given language in which one designation represents two or more unrelated concepts. Homonyms pose huge problems for technical communication. As a consequence, experts are constantly striving to avoid homonyms in technical subject fields. Nevertheless, when new concepts evolve, people like to form new terms for them by combining familiar existing terms or by adopting established terms from general language (e.g. mouse for a computer input device) or from other related subject fields (e.g. virus from medicine for malware).

2. CONCEPT-ORIENTED TERMINOLOGY MANAGEMENT

The results of any kind of terminological work have to be stored today in terminological databases (term bases) or terminology management systems. Although the access to and the retrieval of the content of the term bases will happen in most cases via the (search) term, the organizational principle of this terminological knowledge resource has to be the concept.

This is the basic difference between a lexicographical entry in a dictionary or lexicon and a terminological entry in a term base or terminology management system. While figure 2 shows the fundamental structure of a dictionary entry (e.g. Wiktionary) where the word (term) is the basis for organizing the linguistic information, figure 2 shows the structure of a prescriptive terminological entry in a term base (e.g. Wikipedia) with the main focus on the concept.

Figure 3 also demonstrates the prescriptive approach of terminology management implemented by standardizing organizations and companies to establish a consistent and unambiguous corporate language; only one term for each concept is the preferred term and other terms for the same concept are admitted or deprecated.

¹ The difference between homonymy and polysemy will not be discussed here; it is irrelevant for computerized terminology management and therefore for this paper.

Figure 2. Structure of a lexicographical entry

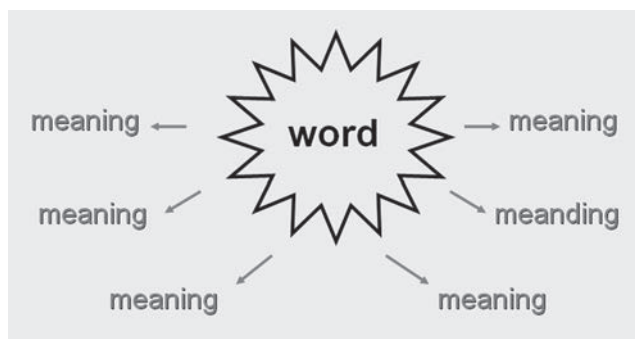
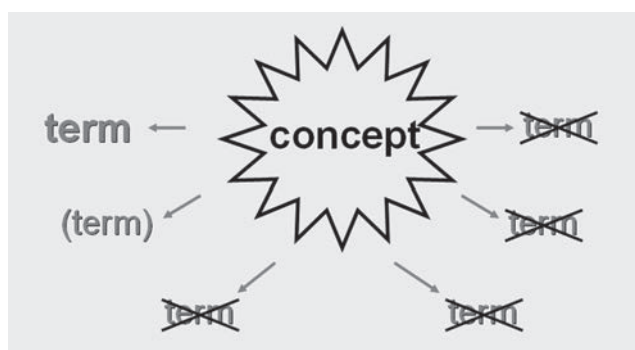


Figure 3. Structure of a (prescriptive) terminological entry



As defined in ISO 1087-1:2000 and reflected in the terminological meta model in ISO 16642:2003, a terminological entry has to contain all terminological data related to one concept. Therefore, terminological data modeling has to reflect the principle of **concept orientation** (see figure 3), thus allowing for the maintenance not only of all concept-related information but also of all terms in all languages with all term-related information within one terminological entry. Terminological entries designed according to the principle of term orientation (see figure 2), which we very often find in bilingual glossaries or dictionaries, are not appropriate for meticulous terminology management and will lead very soon to inconsistent terminology collections that are not very useful, especially if multilingual terminology management is required. If LSP lexicographical products – especially in printed form – have to be created, a term-oriented alphabetical view can be generated from a concept-oriented

terminological data base, since only the conceptual organization can guarantee an adequate collection, processing, revision and preparation of the domain-specific terms.

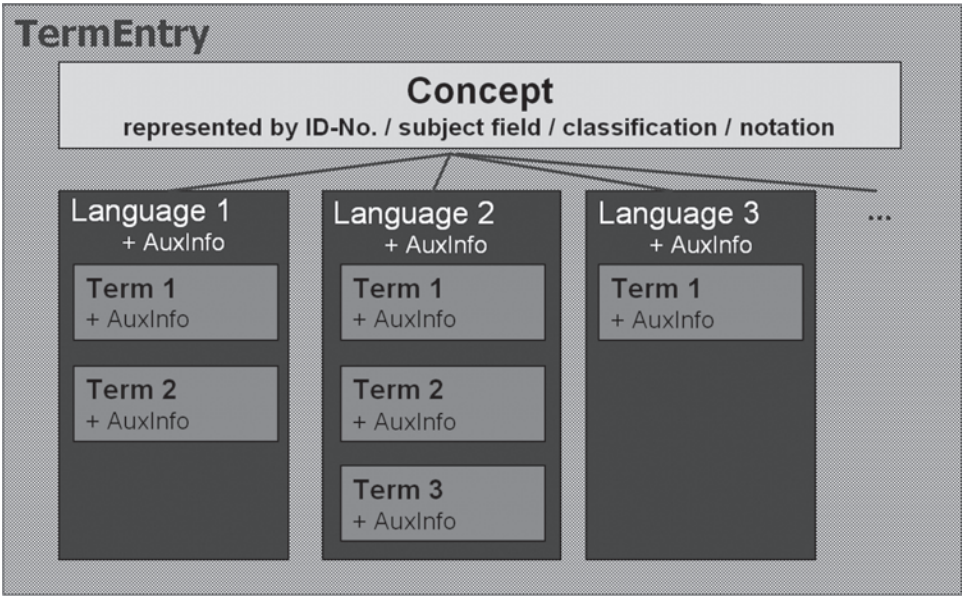
In addition – not in opposition – to concept orientation, the second important principle of terminological management is **term autonomy**. Term autonomy guarantees that all terms including synonyms, abbreviated forms and spelling variants can be documented with all necessary term-related data categories such as grammar, style, geographical restriction or context. This approach can be realized by designing the data model in a way that allows the user to create an unlimited number of term sections or term blocks containing individual terms and all additional data categories describing the term and its use.

Term autonomy is represented in the terminological meta model (ISO 16642:2003) by the fact that each term section is only allowed to have exactly one term. Several terms in one language for the same concept (synonymous designations) will be organized by using several term sections each containing exactly one term and additional data categories documenting this term. In application scenarios where the terminology management system plays the role of the terminological knowledge base for a number of applications and programs, term autonomy is really essential; e.g. a quality control program that checks the correct use of terms in texts, has to have access to the term base where preferred, admitted and deprecated terms are stored in the same terminological entry (see figure 3), but in different term sections.

The principles of concept orientation and term autonomy are reflected in the model of a terminological entry shown in figure 4. Terminology management programs designed according to this model are able to give a warning when the user tries to enter information related to only one concept into different terminological entries (see Schmitz 2011: 242 f.).

Concept orientation and term autonomy are prerequisites for domain- or company-specific terminological data collections (see also Schmitz, Straub 2010: 38 ff.). Synonymous terms such as *USB stick*, *USB memory stick*, *USB flash drive*, *USB memory key*, *memory stick*, *pendrive*, *thumbdrive*, or only *key* as a short form, are stored in one terminological entry with only one English definition and concept relations to other concept entries; each term can be documented with attributes specifying that e.g. *USB flash drive* is the preferred term for documentation and *USB key* for product labelling, and all other terms shall not be used within any docu-

Figure 4. Model of the terminological entry reflecting concept orientation and term autonomy



ments of this company. Because of the concept orientation, any user consulting the term base searching for e.g. *pendrive* will find the concept entry (with the definition), will see that *pendrive* is not the recommended term, and will find the preferred term *USB flash drive*.

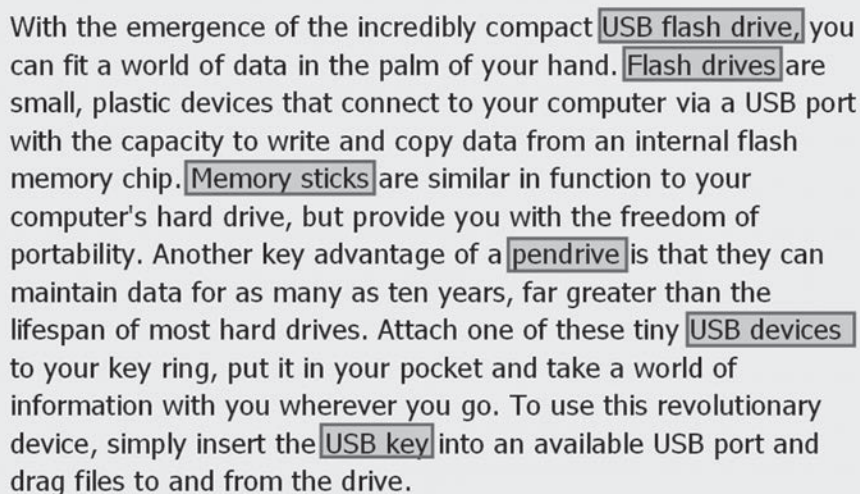
On the other hand, searching for *key* will result in several hits in the same database, because the term *key* is a designation for several concepts (homonym or polysemous word, e.g. *key* as part of the keyboard) and therefore stored in several entries with different definitions and concept relations.

3. TERMS IN TEXTS

Although terms in terminology management systems are organized in concept-oriented entries, they are used in technical documents as elements of the linguistic surface of the text. Depending on the morphological and syntactical features of a given language, terms appear in inflected form indicating e.g. plural, genitive, or past tense. The connection from a term in a document to the corresponding concept can only be established by the reader comprehending the message of the text and grasping the meaning behind the term.

In special or technical communication, it is highly recommended to use only one term for a single concept (no synonyms) and to use only terms that are not ambiguous and do not represent two or more concepts (no homonyms). But this is not always easy to achieve and the world is full of documents that do not meet the requirements of a consistent and univocal usage of terminology. Figure 5 shows a sample document with 6 terms, some of them are multi-word terms, that (may) represent the same concept. For understanding the message of the text, the reader has to cope with the question, if all terms represent the same concept or have different meanings. Some non-experts in the domain the document deals with may be confused, embarrassed, or even overstrained.

Figure 5. Sample document with terms that may represent the same concept



With the emergence of the incredibly compact **USB flash drive**, you can fit a world of data in the palm of your hand. **Flash drives** are small, plastic devices that connect to your computer via a USB port with the capacity to write and copy data from an internal flash memory chip. **Memory sticks** are similar in function to your computer's hard drive, but provide you with the freedom of portability. Another key advantage of a **pendrive** is that they can maintain data for as many as ten years, far greater than the lifespan of most hard drives. Attach one of these tiny **USB devices** to your key ring, put it in your pocket and take a world of information with you wherever you go. To use this revolutionary device, simply insert the **USB key** into an available USB port and drag files to and from the drive.

But not only human readers of technical documents will have problems to deal with texts containing synonymous and homonymous terms. Computer-assisted text processing tools will have similar or even more problems.

In many organizational environments and application scenarios of terminology work, the extraction of terminology from existing textual material is recommended. Typical scenarios are the preparatory terminology work for large translation projects with several translators, before the translation starts and each translator has to do (probably the same) ad-hoc terminology work, and the initial feeding of a new term base with com-

pany or subject specific terminology in order to identify the basic necessary set of concepts and terms.

Terminology extraction comprises tasks for extracting terminological information, mainly terms itself, from textual material. Textual material can be a set of monolingual documents, a pair of parallel texts either produced in both languages, a source language text together with its translation, or a text corpus with a structured and systematically collected set of sentences.

Human term extraction by domain experts or experienced terminologists is the most time consuming and expensive method, but probably leads to the best results. Computer-assisted term extraction programs can handle texts in (almost) all languages if they use only statistical methods. If their term identification algorithm is based on linguistic methods, the results of term extraction will be much better, especially for multi-word terms and phrases, but (commercial) linguistic-based term extraction tools are only available for “major” languages such as English, French, German, Spanish and few others.

Term extraction tools offer common functionalities known from concordance programs (e.g. WordSmith): they identify the words of a textual document, create word frequency statistics, display a KWIC index (Key Word In Context), and display the results sorted in alphabetic order or by frequency. Since words appear in texts in inflected forms, linguistic-based term extraction tools can reduce the text form of a word to its basic canonical form; this is needed for real word statistics and reliable term candidate lists, but requires linguistic knowledge about the morphology of the respective language.

Since terminology is always related to domain-specific language, term extraction tools should be able to filter out and ignore function words (e.g. articles, conjunctions, prepositions etc.) as well as general language words; for this feature, most of the tools use so-called stop word lists that are language dependent and can be complemented by the user. But sometimes it is not so easy to decide if a word is a general language word or a special language term.

Although term extraction tools may be very helpful in specific application scenarios, the following issues have to be taken into account:

- The result of a term extraction process is a list of term candidates; this list must be checked and “cleaned” by a terminologist.

- Term extraction tools provide just a list of terms (sometimes with context examples) and no other terminological information; it can be seen as a to-do-list for the terminologist who has to enrich the terminological entries with all other necessary information and who has to intellectually check and combine e.g. synonyms (different term candidates) to concept-oriented entries.
- Many term extraction tools have problems to exactly identify multi-word terms, noun phases, or verbal phases, especially if they are part of elliptical constructions or composed of discontinuous elements.
- The more linguistic knowledge is integrated into term extraction programs, the better are the results, especially for identifying inflected word forms and reducing them to the canonical form, but the applicability is limited to only “major” languages.

Although term extraction is an important linguistic-based working procedure for terminology management, the results are only useful if the concept-oriented and ontological aspect of terminology is taken into consideration. Extracted terms have to be allocated intellectually to the respective concepts, and synonyms, spelling variants and abbreviated forms as well as homonyms have to be identified and ordered adequately into concept-oriented terminological entries in term bases.

5. CONCLUSION

Terminology work and terminology management has to deal with concepts and terms. For terminology tasks such as term creation and term extraction, LSP linguistics provide appropriate means to coin, select and identify terms, but for any kind of terminology management it is extremely important not to lose track of the concept part of terminology. Not only terminological concept systems but also all types of term bases and terminological data collections – seen as knowledge organization systems – have to follow a concept-oriented data modelling and working procedure approach. This approach is the necessary basis to guarantee that the (linguistic) knowledge of a subject field or a company is not accidentally scattered by linguistic features of the terms, but organized and managed by the ontological part of the concepts and the relations between concepts.

BIBLIOGRAPHY

- Arntz R., Picht H., Schmitz K.-D. 2014: *Einführung in die Terminologearbeit*. 7., vollständig überarbeitete und aktualisierte Auflage, Hildesheim: Georg Olms Verlag.
- DIN 2342: 2011 *Begriffe der Terminologielehre*, Berlin: Beuth.
- Felber H., Budin G. 1989: *Terminologie in Theorie und Praxis*, Tübingen: Narr.
- ISO 1087-1:2000 *Terminology work – Vocabulary – Part 1. Theory and application*, Geneva: ISO.
- ISO 16642:2003 *Computer applications in terminology – Terminological markup framework*, Geneva: ISO.
- Ogden C. K., Richards I. A. 1923: *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, London: Kegan, Paul, Trench, Trubner.
- Schmitz K.-D. 2011: Managing Terms in Terminology Management. – *Magyar Terminológia = Journal of Hungarian Terminology* 4/2, 238–245.
- Schmitz K.-D., Straub D. 2010: *Successful Terminology Management in Companies – Practical tips and guidelines*. Stuttgart: TC and more. (Updated German issue will be available in 2015.)

TERMINAI TEKSTUOSE IR IŠŠŪKIS TERMINIJOS TVARKYBAI

Mokslinėse publikacijose, nacionaliniuose ir tarptautiniuose standartuose aiškinama, kad terminologijos teorija ir terminijos tvarkyba nagrinėja sąvokas ir terminus. Sąvokos apibrėžiamos kaip žinių ar mąstymo vienetai ar apskritai kaip mintyse esantys vaizdiniai, kuriuos galima apibūdinti apibrėžtimis. Geriausia, kai apibrėžtys siejasi su kitomis sąvokomis ir įvardija būdingas apibrėžiamos sąvokos ypatybes. Remiantis apibrėžtimis ir sąvokų ypatybėmis, galima sąvokas susieti ir sudaryti jų sistemas, atspindinčias tam tikros srities ar posričio žinias. Iš sąvokų yra sudėliota bet kuri žinių organizavimo sistema – tezasauras, ontologija, taksonomija, terminologinė sąvokų sistema ar tiesiog terminologinių duomenų bazė.

Profesinėje komunikacijoje sąvokoms įvardyti vartojami nusistovėję terminai. Žinoma, techniniuose tekstuose jie atsiranda nepakeistu ir nuo nieko nepriklausančiu būdu, jų formą lemia morfologiniai tam tikros kalbos bruožai, jie pateikiami tam tikrame (socio)lingvistiniame kontekste, į kurį būtina atsižvelgti imantis bet kokios terminologinės veiklos. Terminijos tvarkybai yra labai svarbu neišleisti iš akių terminijos sąvokinės pusės. Ne tik terminologinės sąvokų sistemos, bet ir visi terminų bazių bei terminologinių duomenų rinkinių tipai, laikomi žinių organizavimo sistemomis, turi remtis į sąvokos vietą orientuotu duomenų modeliavimu ir darbo tvarka. Terminų, įskaitant rašybos variantus, sutrumpintas formas ir sinonimus, teisingas priskyrimas atitinkamoms sąvokoms, ypač automatiškai išrenkant terminus, yra sunkus terminijos tvarkybos uždavinys, kurį išspręsti gali padėti specialistai ar terminologai. Tik laikantis į sąvoką orientuoto požiūrio galima užtikrinti, kad tam tikros srities ar įmonės kalbinės žinios būtų struktūrinamos ir tvarkomos, remiantis sąvokomis ir jų ryšiais, o ne tik šiek tiek papildomos lingvistiniais terminų bruožais.

Gauta 2015-08-03

Klaus-Dirk Schmitz
Cologne University of Applied Sciences
Institute of Translation and Multilingual Communication
Ubierring 48, 50678 Cologne, Germany
E-mail klaus.schmitz@fh-koeln.de