

Antje Schmidt-Wigger  
Institute of Applied Information Science (IAI)  
Germany  
Presented at: TKE '99, 23-27.8.99, Innsbruck

# TERM CHECKING THROUGH TERM VARIATION

## Introduction

Terminology constitutes the major part of a technical document. The use of the right or preferred terms in a technical document is, for economical, legal or just for presentational reasons, of prime interest for the company. Controlling this use should be automated, especially in the application context of a Controlled Language. Term variation constitutes a valid clue for the realisation of such a control. By using general variant formation rules, variants of preferred terms can be detected automatically, without the need for exhaustive variant collections.

## Controlled Language and Terminology

During the last decade, industrial and research interest has grown in the field of Controlled Language (CL) development and applications (v.d.Eijk 1998). After concentrating in an earlier phase on a simple English to render translation of documents unnecessary, the current approaches focus on eliminating ambiguous, complex and redundant elements of natural language in order to allow better and quicker (human and automatic) reading, understanding and - as the ultimate and sometimes only goal - translation of text. An important step towards a successful integration of a CL in an industrial environment is the availability of automatic CL checkers for proof-reading/editing purposes. Special attention has to be paid for authors' interests as the users of such a tool.

Most CL approaches define a set of forbidden syntactic structures and words from the general vocabulary, together with the preferred structure or word. This allows the CL to be applicable to different subject fields, at least when describing a comparable text type. The rules for the specific terms of a particular subject field have been left to the terminologists, as the definition of preferred terms for a concept has always been their task (Wright 1997), following criteria comparable to those used for the CL as a whole: no ambiguity through homonyms, no redundancy through synonyms.

In theory, the integration of a well-defined terminology into a CL checker is an easy task, where for each rejected variant or synonym, the preferred term is retrieved and proposed as a correction to the author. In practice, the realisation is not as easy; checkers mostly identify term candidates<sup>1</sup> and reject those not known by the system. At best, similar terms in the lexicon are proposed, but if the lexicon does not cover the concept, nothing or nonsense is displayed (e.g. 'Terminology Suite' of XEROX). At worse, the technical author has to wait for a reaction from the terminology department to know which term he should use instead - if the error has not arisen simply from a lack in the system's dictionary (e.g. 'MaxIt' of SMART communications). But authors do not like to play the terminologist's role. They expect a CL checker to tell them how to write - if they are obliged to adapt their language to a standard, at least this standard has to be settled; if not, how can they trust the error messages to follow them even if they contradict their own ideas of language?

---

<sup>1</sup> See the abundant literature on term candidate recognition, e.g. in the present volume.

The obstacle to integrate a correction-driven terminology control in a CL checker is probably the huge amount of data which should be examined and judged by terminologists (if already the company is willing to pay for) to store all occurring variants in the lexicon. When confronted with users in an automated environment, the number of gaps becomes obvious. Not only should each concept of the company's activities be described, but in fact all possible variants of terms denoting this concept. For a human user or builder of a terminology, varying a term does not necessarily mean that a non-preferred variant has been used, as he/she is able to abstract away from e.g. orthographical differences.

- (1) *Vierventilmotor* four-valve engine
- (2) *4-Ventilmotor* 4-valve engine

If a machine is the user, even a slight difference in punctuation, spelling or word order can disturb the translation process, because the transfer entry cannot be retrieved. For Translation Memories (TM) which are based on string correspondences, such slight differences are manageable. But as it will be shown, term variation can be very complex, thus disturbing the similarity calculation enough to reject equal sentences only distinguished by such term variants.  
E.g. 'material characteristics'

- (3) *Kenndaten für Werkstoff* characteristic data for work stuff<sup>2</sup>
- (4) *Materialkennwerte* material's characteristic values

And when it comes to Machine Translation (MT)<sup>3</sup>, problems are even bigger, if the system in question does not incorporate a variant recognition process in its terminology recognition step. But mostly they limit themselves to inflectional morphology at this stage of processing.

As an alternative to a complete collection of term variants, we propose a method based on regular term variation. Here, terminologies need to contain only preferred terms - or, in case of an MT system, known terms. The non-preferred variants used by the technical authors are detected inside the texts by different transformations on the preferred terms. This procedure allows for avoiding redundancy in choice of wording; ambiguity is also tackled when a non-preferred term gets more than one equivalent proposal. As an important advantage to other checkers, the detection of variants is based on the variant relation itself. The authors thus always get a correction proposal together with the error message - an advantage helpful for the acceptance of a Controlled Language checker inside a company.

## **Project Context**

The proposed method has been successfully implemented for the German language in MULTILINT, a joint R&D project between our institute and a large German car manufacturer. The project has been realised in the after sales department, where service information and repair instructions are produced for the repair workshops. The technical authors responsible for the production of these instructions come exclusively from an engineering background, without any

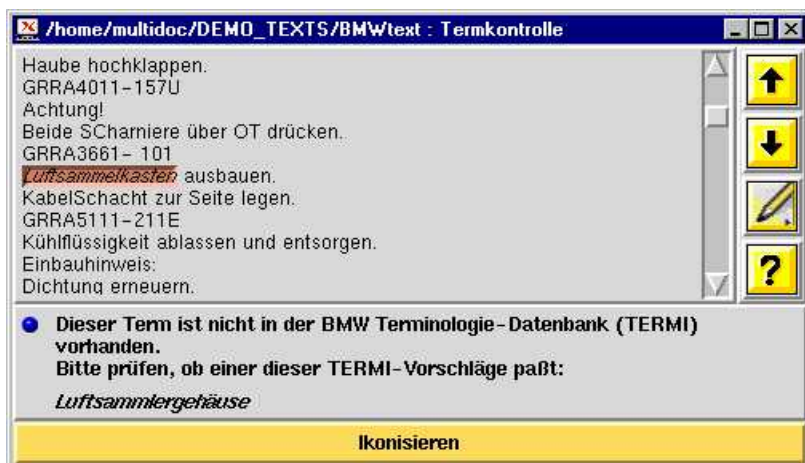
---

<sup>2</sup> For non-Germanophones, each example is accompanied by its transliteration into English. Transliteration has been preferred over real translation to underline, if possible, the subtle differences between the variants. Real, official translation (between quotes) has been added when synonyms allow for one common translation into English.

<sup>3</sup> The same stands for automatic indexing and retrieval.

fundamental training in authoring. Therefore, document quality is not always adequate and an amelioration on different linguistic levels is needed.

The overall goal of MULTILINT was thus to develop a prototype platform for language control (Reuther 1998) encompassing three different levels of linguistic quality: grammatical, stylistic and terminological correctness. While grammatical correctness are more or less universal, stylistic criteria have been designed especially for the text type(s) in question. Terminology control, however, always depends on a well defined terminology, a basis which was neither available from the industrial partner nor could be defined by the linguistic partner of the project. Therefore, a general methodology of terminology control has been developed and implemented during the project, which is valid independently of the size and quality of the underlying terminology.



The current version of the prototype, which is used daily with a group of about 30 authors, contains a basic, unstructured terminology list with about 15,000 entries. The authors ask the MULTILINT prototype for correction of their text. In response, they get for each different linguistic quality level a copy of their text with problematic text parts highlighted and linked with a correction proposal. The results of the term checker are generally considered by the authors sense- and useful. They adapt their choice of wording to the correction proposed by the system.

The underlying principles and programs are currently enlarged in description power and passed over to other application languages through MULTIDOC, a European R&D project also in the car manufacturing industry. Other applications are envisaged, showing reusability of the implementation for other subject fields.

### The study of term variation

Term formation has been studied deeply since the early seventies in the field of Languages for Special Purposes (LSP) when it still was mostly oriented towards the lexicon. Already then, term variation has been identified as an important obstacle for precision and linguistic economy (Lenders 1980). Studying term variation inside of LE applications, however, has only recently emerged in the linguistic community, supporting research for better indexing and retrieval results (Jacquemin 1991). In the field of CL, methods could be similar, but the range of forbidden variants is probably smaller than those used for indexing purposes. The relevant literature focuses on lexical term variation, especially on insertion and expansion phenomena where a term is modified and/or specified by further lexical items.

E.g. 'ribbed V-belt'<sup>4</sup>

<sup>4</sup> All examples have been found in the documentation of the project partner.

- |                             |            |
|-----------------------------|------------|
| (5) <i>Keilriemen</i>       | V-belt     |
| (6) <i>Keilrippenriemen</i> | V-rib-belt |

The status of those variant types for a terminology control is not clear. On one hand, standardisation could forbid the modification of terms internally (i.e. (SAE 1995) by specifying the standard order for term building elements). On the other hand, expanded terms always represent hyponyms of the base terms and could therefore be considered new term candidates. But in both cases, one term is clearly not a lexical variant of the other, because new lexical material has been introduced.

Results should also be judged from the point of view of the specific users of the different applications. Technical authors as the users of a checking tool mostly are not skilled in linguistics, but willing to accept a proposal from the machine as a law. Therefore, precision is of highest importance, but obvious errors do not present a problem if the author can correct them himself. On the other hand, the author is not directly affected by a lack of recall (thus not creating an obstacle for acceptance), but it does affect translation. In information retrieval, on the other hand, users are very sensitive to recall problems because finding is their direct goal.

### The treatment of term variation

Term variation touches a wide spectrum of linguistic phenomena, from graphematical and orthographical to morphosyntactic differentiations. These variations in the strict sense can be enlarged when we consider also synonyms and quasi-synonyms as variants. Other very common variants are abbreviated forms, ranging from acronyms as the shortest, merely graphematically motivated forms, to the suppression of term elements or the use of hyperonyms. And it has to be stressed that all these regularities in variation can be combined during the writing process, taking into account the combinatorial and motivated characteristics of term formation.

E.g. 'service advisor'

- |  |                                |
|--|--------------------------------|
| (7) <i>Kundendienstmeister der Annahme</i> | (service master for reception) |
| => suppression + syntactic transformation  |                                |
| (8) <i>Kundendienstannehmer</i>            | (service receptionist)         |
| => graphematic transformation              |                                |
| (9) <i>Kundendienst-Annehmer</i>           | (service-receptionist)         |
| => abbreviation                            |                                |
| (10) <i>KD-Annehmer</i>                    | (S-receptionist)               |
| => synonymy                                |                                |
| (11) <i>Serviceberater</i>                 | (service advisor)              |

To capture the relations between the preferred terms and their variants, we propose to use a powerful morphological analysis abstracting away from inflection, compounding and derivation. In the concrete implementation, we use the German morphological analyser MPRO developed at our institute (Maas 1996), which produces about 98% reliable analyses on new input texts. Concerning the present work in a delimited domain, errors are even less common. The results of MPRO are presented in feature bundle format, including, among others, features for the original string (*ori*), the lemma (*lu*) and its morphological structure (*ls*), lexical category (*c,sc*), word number (*wnr*) and relevant morphological information (*nb,g,vtyp,per..*).

E.g. *Alle Ventildedern ausbauen.* ('Remove all valve springs.')

---

{ori=Alle,lu=alle,c=w,sc=quant,ehead={nb=plu},wnr=1},  
 {ori=Ventilfeder,lu=ventilfeder,ls=ventil#feder,c=noun,ehead={nb=plu,g=f},wnr=2},  
 {ori=ausbauen,lu=ausbauen,c=verb,vtyp=inf,wnr=3},  
 {ori=ausbauen,lu=ausbauen,c=verb,vtyp=fiv,per=1;3,nb=plu,hsns=ns,wnr=3},  
 {ori=.,lu=punct,c=w,sc=punct,wnr=4}

The calculation of the derivational base is driven by combinatorial criteria and information from the lexicon. This device allows for the linking of derivationally related terms:

E.g. *Anzugsdrehmoment* vs. *Anziehdrehmoment*  
 (attraction rotating moment) (attracting rotating moment) ('starting torque')

(12){ori=Anzugsdrehmoment,ls=an\_\$ziehen#drehen#moment}

(13){ori=Anziehdrehmoment,ls=an\_\$ziehen#drehen#moment}

If possible, a verbal form is chosen as the kernel of the derivational family. Nevertheless, it seems that other languages, e.g. French, have a more complex derivational morphology, i.e. with more verbalisations, a fact which presumably was an advantage for the treatment of German. Another advantage of the German language is its great lack of ambiguity concerning the lexical category of a given word. Because of capitalisation, ambiguity seldom occurs generally, leaving e.g. some adjectival forms and their verbalisations.

E.g. Metaphoric derivation for *links/linken* (left/to con)

(14) *den linken Stoßdämpfer*            the left shock absorber-OBJ

(15) *den Kunden linken*                to con the client<sup>5</sup>

In such cases, subject specific filter can be used to reduce the number of analyses produced by the morphological engine when compiling the terminological base, because no context is available there.

But much ambiguity to be resolved lies in the decompositional problem of compound words of German. Here, MPRO attempts to maintain its excellent analysis quality by manually collecting false decompositions in an exclusion list.

To sum up, the use of a morphological analyser developed for general language texts allows for the correlation of morphologically related terms by the use of a device independent of a specific subject field.

The word-oriented device presented above has to be augmented by a set of general reordering rules allowing the recognition of compound terms vs. syntactic phrases. Without reordering, the use of a relational adjective in place of a compound non-head is already captured by the morphological analysis and a segmentation mechanism. The syntactic category is not used in term variant recognition<sup>6</sup>; it is only displayed here for demonstration purposes.

E.g. *Automatisches Getriebe* vs. *Automatik-Getriebe* ('automatic transmission')

(16){ori=automatisches, c=adj, ls=automatik}, {ori=Getriebe,c=noun,ls=treiben}

<sup>5</sup> Sorry, this is the only invented example.

<sup>6</sup> In the present implementation, we exclude inflected verbs, that means verbs outside of a compound, from the variant recognition process. This should be refined in further development.

(17){ori=Automatik-Getriebe,c=noun,ls=automatik#treiben,cs=noun#noun}  
=>{c=noun,ls=automatik},{c=noun,ls=treiben}

Equally recognisable, a type of non-reordering syntactic variants more and more common in technical documents is constructed on the 'English' compounding mechanism where compound parts are written separately, without a hyphen.

E.g. *Aggregateunterschutz* vs. *Aggregate-Unterschutz* vs. *Aggregate Unterschutz* ('splash guard')

(18){ori=Aggregateunterschutz,ls=aggregat#unter#schuetzen}  
=>{ls=aggregat},{ls=unter},{ls=schuetzen}

(19){ori=Aggregate-Unterschutz,ls=aggregat#unter#schuetzen}  
=>{ls=aggregat},{ls=unter},{ls=schuetzen}

(20){ori=Aggregate,ls=aggregat},{ori=Unterschutz,ls=unter#schuetzen}  
=>{ls=aggregat},{ls=unter},{ls=schuetzen}

To correlate preferred terms with their reordered syntactic variants, we are currently adapting the FASTR formalism (Jacquemin 1991) to the MPRO analyser and the specific German variant rules. This should demonstrate the feasibility of the presented approach also for examples much more complicated than the following, where a metarule is responsible for the inversion of the lexical elements and the deletion of the function word.

E.g. *Gebläseschalter* vs. *Schalter für Gebläse*  
(blower switch) (switch for blower)

(21){ori=Gebläseschalter,ls=blasen#schalten,c=noun,cs=noun#noun}  
=>{ls=blasen},{ls=schalten}

(22){ori=Schalter,ls=schalten,c=noun},{ori=für,c=w},{ori=Gebläse,ls=blasen,c=noun}  
=>{ls=blasen},{ls=schalten}

It can be shown in all these implementations that general rules allow for the distinction between terms and other parts of the sentences. They are not specific to the subdomain described nor to the treatment of terminology compared to general language noun phrases. They obey general language paraphrasing principles of category-switch and their dependent word-order and function word changes. In a way, the clusters of paraphrases are even language-independent, as these principles are valid for a large group of languages<sup>7</sup>.

A more controversial mechanism of term variant construction is the use of synonyms for parts of terms (Hamon 1998). In the present implementation, synonym relations are accounted for by replacing one lexeme being part of the term by a disjunction of its possible synonyms. We identified synonym group candidates mainly by the back-and-forth-translation method (Ehrlich 1998), but kept only relations between base words which occurred as discriminator for several composed synonym groups. The synonym group candidates have been controlled by experts of the domain, resulting in a number of 130 implemented synonym groups with an average of three to four synonyms per group.

E.g. *Drosseldüse* vs. *Drosselklappe* vs. *Drosselventil*  
(throttle nozzle) (throttle flap) (throttle valve)

(23){ori=Drosseldüse,ls=drosseln#duese}  
=>{ls=drosseln},{ls=duese,hahn;klappe;ventil}

---

<sup>7</sup> For English, e.g. parallel variant construction is exemplified through the valid translation of the German examples.

(24){ori=Drosselklappe,ls=drosseln#klappe}  
=>{ls=drosseln},{ls=klappe;duese;hahn;ventil}  
(25){ori=Drosselventil,ls=drosseln#ventil}  
=>{ls=drosseln},{ls=ventil;duese;hahn;klappe}

Some of the synonym groups identified during our research are in fact also quite independent of the subject field, as they introduce in different forms generic information such as 'process', 'tool' or 'thing' necessary for the instantiation of a subject specific item in the textual context (Reinhardt et al. 1992). The weak semantic attribution of these 'support nouns' can also be demonstrated by replacing them with an equivalent derivational suffix<sup>8</sup>.

E.g. *Steuergerät* vs. *Steuereinheit* vs. *Steuerung*  
(control device) (control unit) (control)

(26){ori=Steuergerät,ls=steuern#geraet}  
=>{ls=steuern},{ls=geraet;einheit;Ø}  
(27){ori=Steuereinheit,ls=steuern#einheit}  
=>{ls=steuern},{ls=einheit;geraet;Ø}  
(28){ori=Steuerung,ls=steuern}  
=>{ls=steuern},{ls=Ø}

By including those synonym based variants, linguistically driven methods have a great advantage over string-based methods, which would easily be able to capture the first, but run into difficulties for the last variant group. For demonstrating this difference, we ran some tests with FindLink (of Connex), a fuzzy match device allowing for parametrising on a minimal matching score, length difference of matching strings and number of displayed matches in the database. Term variants have been compared with the complete term list database, containing also the preferred variant of the terms in question. But for finding even very near variants, errors occur in parallel.

E.g. to find (29), it also finds (30).

(29) *Ausschwingvorgang* vs. *Ausschwingungsvorgang*  
(decaying process) (decay process)  
(30) *Automatikgetriebe* vs. *Automatikgetriebeöl*  
(automatic transmission) (automatic transmission oil)

### Problematic cases

As for all generalising procedures, term variant recognition also comes up with undesired linking of terms. This could be due to specific, accidental problems, but also to general tendencies of the morphosyntax of the language described.

One general problem is the semantic content derivational morphemes add to the word they create. One cannot ignore the difference they introduce: term variants need to carry corresponding semantics, as they denote per definition the same concept. But it would be too strict to limit variant linking to words of the same derivational class, given the heterogeneous meanings of derivational suffixes.

E.g. *Steckverbinder* vs. *Steckverbindung* are synonyms in the instrumental meaning.  
(plug connector) (plug connection)

---

<sup>8</sup> This elliptic variant formation process sheds light on the head inheritance principle of word formation. Indeed, the current implementation of FASTR does not foresee the possibility of head elision or the division of semantic information between the derivational base and its suffixes.

- (31){ori=Steckverbinder,ls=stecken#verbinden,s=instr}  
 (32){ori=Steckverbindung,ls=stecken#verbinden,s=result;instr}

Beside the prototypical meaning, suffixes in all languages tend to cover other apparent meanings which cannot be predicted for each word in each application context or sublanguage. One famous example is the agent-instrument dichotomy of the correspondent suffix. Other problems are linked with suffixes of the 'process' family. They are often involved in resultative or even agentive derivations. A solution could be to allow derivational difference only for non-heads: the basic semantics of the concept is triggered by the meaning of the head. The semantics of the non-head are not decisive for the whole concept, because each of the participants of a situation could define the situation as being a part of the concept in question.

E.g. *Halteklammern* vs. *Halterklammern*  
 (retaining clips) (retainer clips)

- (33){ori=Halteklammern,ls=halten#klammer,ss=process#thing,s=thing}  
 (34){ori=Halterklammern,ls=halten#klammer,ss=agent#thing,s=thing}

In the present implementation, the problem is approached using a two-fold strategy: 1) single words are not presented as term variants because of a lack of context evidence and 2) some general semantic information of the head on agentivity and on quality/state derivation calculated by the morphological analyser on lexical and derivational information limits possible linking.

E.g. not linked are:

- (35){ori=Kompressor,ls=komprimieren,s=agent} (compressor)  
 (36){ori=Kompressibilität,ls=komprimieren,s=state} (compressability)  
 (37){ori=Komprimierung,ls=komprimieren,s=process} (compression)

Despite the acceptable results, we wish to refine this approach in the future.

A second general problem lies in homonymous elements inside synonym groups. Those words have to be identified carefully to avoid too large and heterogeneous groupings. The present implementation allows for the splitting up of synonymous groups so that a homonymous word can be defined as being part of more than one synonym group.

E.g. *Leitung* ('cable' vs. 'pipe') being homonymic:

- (38) *Batteriekabel* vs. *Batterieleitung* ('battery cable')  
 => {ls=batterie}, {ls=leiten;kabel}  
 (39) *Ablaufleitung* vs. *Ablaufrohr* ('drain pipe')  
 => {ls=ab\_\$laufen}, {ls=leiten;rohr;stutzen}  
 (40) *Druckkabel* vs. *Druckrohr*  
 (pressure cable) (pressure pipe)  
 => {ls=drucken}, {ls=leiten;kabel}  
 => {ls=drucken}, {ls=leiten;rohr;stutzen}

Another general problem is home-made: at the moment, all function words carry the same lexical information and their semantic content is not taken into consideration. This is helpful in cases of e.g. different prepositions for the same relation, but harmful when these prepositions are opposed, as in the following example:

E.g. *mit/ohne* (with/without)

- (41) *Anhängelast mit Bremse* ('towed load with brake)  
(42) *Anhängelast ohne Bremse* (towed load without brake)

In the following example, the uniform treatment of function words mis-encounters in addition an accidental homonym. And indeed, accidental linkings are mostly due to the encounter of more than one erroneous match.

E.g. *Steuern* (taxes / controlling)

- (43) *vor Steuern* (before-tax profit)  
(44) *nach Betätigung* (after actuation)

But in the end, erroneous links between a preferred term in the lexicon and elements in the text can often be repaired when the textual elements have in fact to be considered as term candidates missing in the lexicon. Once they are stored, their use is documented and the term checker does not mark them as errors anymore. As a reminder: the principle presented for terminology control is oriented towards the authors and can handle terminological lists of different quality, but it does not replace a substantial, regular terminological work on the lists themselves.

## Conclusion

In the application field of Controlled Language, paraphrasing principles can be used for term variant recognition, a valid approach for terminology control. These general variant formation rules are actually independent of the subject field involved; only for specific synonym groups, restrictions from the subject field have to be taken into account. Taking advantage over term candidate extraction methods, control of term variation directly can be introduced into an authoring environment where production of unambiguous and precise information in strict time limits is a challenge for quality measurement.

## Bibliography

EHRlich, U. (1998): Automatic Extraction of a Unique Terminology Based on a Multilingual Corpus and Dictionary. In: Proceedings of LREC. Granada: ELRA.

EIJk, van der, P. (1998): Controlled Languages in Technical Documentation. In: ELSNews 7.1, Feb.1998. Edinburgh: ELSNET.

HAMON, T. (1998): A Step towards the Detection of Semantic Variants of Terms in Technical Documents. In: Proceedings of COLING. Montréal.

JACQUEMIN, C. (1991): Transformations des noms composés. Doctoral thesis. Paris VII.

LENDERS, G. (1980): Erscheinungsweisen der Synonymie in der terminologischen Lexik technischer Fachsprachen, untersucht am Wortschatz der Elektrotechnik/Elektronik. Doctoral thesis. Dresden.

MAAS, D. (1996): MPRO - Ein System zur Analyse und Synthese deutscher Wörter, In: HAUSSER, R. (ed.): Linguistische Verifikation, Sprache und Information, Tübingen: Max Niemeyer Verlag.

REUTHER, U. (1998): Controlling Language in an Industrial Application. In: Proceedings of the Second International Workshop on Controlled Language Applications CLAW'98. Pittsburgh.

REINHARDT, W., KÖHLER, C., NEUBERT, G. (1992): Deutsche Fachsprache der Technik. Hildesheim: Olms.

SAE International (1995): SAEJ 1930 - surface vehicle recommended practice. Technical report, issued 9.1991, revised 9.1995. Warrendale, PA : The Engineering Society For Advancing Mobility Land Sea Air and Space.

WRIGHT, S.E. (1997): Chapter 2.2.1: "Terminology Standardization: Management Strategies", In: WRIGHT, S.E./BUDIN, G. (eds.): Handbook of Terminology Management, Vol.I. Basic Aspects of Terminology Management. Amsterdam/Philadelphia: John Benjamins.