

Building Consistent Terminologies

Presented at: COMPUTERM 1998, Montréal

Antje Schmidt-Wigger, MA.

Institut für Angewandte Informationsforschung
Martin-Luther-Straße 14,
D-66111 Saarbrücken
antje@iai.uni-sb.de

Abstract

Large terminologies tend to contain many inconsistencies on different linguistic levels. Mostly the multilingual inconsistencies are discussed for translation means, but monolingual inconsistencies introducing different variants of the same term cause already a large amount of problems during manual or automatic use of the terminology. The present article concentrates on monolingual inconsistencies in a specific collection of German terms. A classification is proposed of all variants encountered, based on their morphosyntactic properties. Automated solutions for the detection of these inconsistencies are suggested and a tentative implementation in a term checking tool is presented. The same tool can help users of the terminology to keep in line with it.

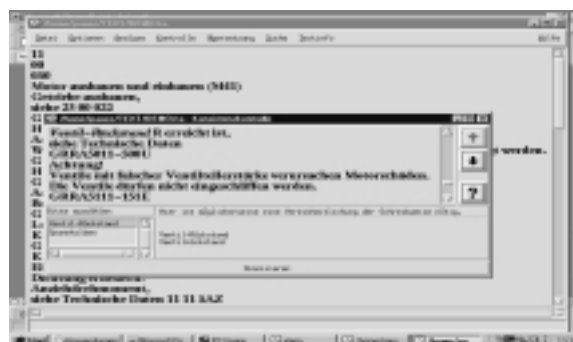
Introduction

Terminological resources are of prime interest for all companies where communication covers a wide range of topics, and partners are not next to each other while communicating. Terminology is also related to the increasing interest of automatic treatment and translation of texts. Therefore it is a major goal today in terminological work to unify the denominations of objects in a coherent way and to realise the thus settled normative standard in the text production process of the companies.

But large terminologies tend to contain many inconsistencies on different linguistic levels. These inconsistencies cause confusion for the writers and prevent a quick and easy automatic

postprocessing of the texts, i.e. automatic translation or indexing.

The present article tries to classify all inconsistencies encountered in a specific collection of German terms and to propose automated solutions for the detection of these inconsistencies¹. The same tools can help users of the terminology to keep in line with it.



1 Terminological Inconsistencies

One basic problem during the construction of a term collection is to collect just ONE terminological entry for EACH term. Thus while adding a new term to the base, it has to be ensured that it really is new, not representing in fact a variant of a term already contained in the base.

As a simple way to encompass the limits of string based search, fuzzy match applications are integrated in most term banks today (i.e. MultiTerm of TRADOS). But fuzzy matching on the word level is less efficient than on the sentence level, as every small difference between two words counts for more than a small difference between two sentences. And in addition, the morphologic structure of the words are not taken

¹ For other languages than German, research is abundant in the field, i.e. [Daille 1994] for French.

into account, which could result in ridiculous linking such as *Urinstinkt* and *Urin stinkt*ⁱ.

Thus, an attempt should be made to identify variants on the linguistic information contained in the new term and the terms already contained in the database.

An analysis of variant collections allows the following classes of morphosyntactic variants to be stated:

Class	Examples	%
Compounding vs. syntactic construction	<i>Autogenschweißen</i> ² <i>autogenes Schweißen</i> ⁱⁱ	3
Complete vs. hyphenated compounding	<i>Airbagmodul</i> <i>Airbag-Modul</i> ⁱⁱⁱ	17
Orthographic variants	<i>Amperemeter</i> <i>Ampèremeter</i> ^{iv}	8
'Fuge' vs. no 'Fuge'	<i>Ausgleichsschlauch</i> <i>Ausgleichschlauch</i> ^v	3
Derivational variants	<i>Anziehmoment</i> <i>Anzugsmoment</i> ^{vi}	8
Elliptic constructions	<i>Anziehdrehmoment</i> <i>Anziehmoment</i> ^{vi}	8
Real synonyms	<i>Abgasleitung</i> <i>Abgasrohr</i> ^{vii}	26
Full wording vs. abbreviation	<i>Abgasrückführung</i> <i>AGR</i> ^{viii}	26

The class of real synonyms could have been considered as the most prominent of this classification. But in the variant collection used for this paper, it represents only one quarter. The other classes of variants are also of importance.

In the following, we will trace specific problems of all these classes and we will propose possible solutions for their recognition.

1.1 Compounding vs. syntactic construction

Compound words can always be paraphrased by syntactic constructions by placing the non-head of the compound in another syntactic constituent where it is in the same dependency relationship with the head noun as in the compound. Thus the paraphrase would again result in a nominal phrase, where the non-head is realized as an

² All the examples (except two) are directly taken from the term collections of a car manufacturing company.

adjective or inside a prepositional or genitive phrase.

Autogenschweißen = *autogenes Schweißen*ⁱⁱ
Nockenwellenanordnung = *Anordnung der Nockenwelle*^{ix}

1.2 Complete vs. hyphenated compounding

Complete compounding and hyphenated compounding of the same word are very often found in technical documentation. The two variants are clearly synonymous, and only definition can decide for one standardised form.

Airbagmodul = *Airbag-Modul*ⁱⁱⁱ

But it seems that some characteristics of the compound parts (i.e. etymology, word form, syntactic category) influence the possibility of one or the other variant, and a hyphen is always helpful for long or ambiguous compounds.

Staubecken = *Stau-Becken* vs. *Staub-Ecken*^x

In this context, a new word formation mechanism has to be described, which begins to represent a sensible portion of German word creation [Königer 1997]: build on the English example, non head nouns are often put beside their head noun as a separated word without any relation marker at all.

Aggregateunterschutz = *Aggregate-Unterschutz* = *Aggregate Unterschut*^{xi}

1.3 Orthographic variants

Orthographic variants are not very common in German. The writing of a word is defined in the DUDEN lexicon which stands as law for all German writing, is used in school and is referred to in cases of legal conflicts. Thus orthographic variants are mostly found for foreign words which just have entered the German linguistic community. Here most problems are caused by consonants in Latin based words, which are written in the source language in one way, but which could be written in German in the assimilated way.

aufklipsen = *aufclipsen*^{xii}
Automatikgetriebe = *Automatic-Getriebe*^{xiii}

1.4 'Fuge' vs. no 'Fuge'

In German, the parts of a compound can be linked by different elements called 'Fugen'. Historically related to inflectional affixes for case or number, the 'Fuge' today does not carry

semantic information. But the choice is normally driven by the non-head, which chooses by analogy only one sort of 'Fuge' when it enters a compound. But for some words, the 'Fuge' information is not very decisive and speakers/writers do arbitrarily choose one or the other form.

Ausgleichschlauch = Ausgleichsschlauchⁱ

1.5 Derivational variants

Derivational variants are found in head and in non-head position. The synonymy between derivationally related variants is based on the different referential capacities of a derivational affix. However, synonymy between two words related by derivation should only be taken into account when the difference is located in the non-head, as head differences are of greater importance for the meaning of the whole word.

Entlüfterventil = Entlüftungsventil^{iv}

Another characteristic of derivational systems is that different derivational affixes can denote the same prototypical meaning [Streiter and Schmidt-Wigger 1995]. Here, a proposition of synonymy is much easier and can also be proposed for head variants.

das Einsatzhärten = die Einsatzhärtung^v

1.6 Elliptic constructions

In elliptic constructions, a – syntactically and semantically – necessary word or word part is omitted, but the process of understanding is not disturbed by this: the omitted word can be reconstructed by the reader on the basis of his world knowledge and the context the elliptic construction is standing in.

*Rücklaufsperrventil = Sperrventil^{vi}
Ausdrückwerkzeug = Ausdrücker^{vii}*

It could also be a reason of ellipsis that a term was not completely assimilated by a writer; he can even produce confusions of relation regarding the constituents of the term by turning around the order of the elements.

Frischlufklappenmotor = Frischklappenluftmotor^{viii}

1.7 Real Synonyms

The most frequent variants found are based on real synonyms, completely different words for one object. In terms of automatic control of the use and consistency of a terminology, this possibility leads to the greatest difficulties, because it

cannot be discovered by linguistic analysis of the word only. Each synonymous relation has to be stated in a list.

Anhebung = Steigerung^{xix}

Possibilities of calculation lie on the paradigmatic level. If two words often occur in the same context or together with the same other words, they probably stand in one semantic relation to each other, which can be synonymous (or hyponymous, hyperonymous, or contradictory).

Another possibility of calculation lies on the multilingual level [Ehrlich 1998]. If two words are translated into one or more languages in the same way, they are probably synonyms. Otherwise, the translations are homonymic themselves, or the difference between the two words cannot be expressed in the target language.

1.8 Full wording vs. abbreviation

A very common – and legal – way to construct variants is through abbreviation. This process is discussed here because it represents a large part of variant production, but should not be considered as a source of inconsistency. Abbreviations are used when text should become shorter, i.e. on machine plates, in tables and figures, on screen menus etc. Technical language makes great use of abbreviations, also because terms become longer and longer and more complicated through the technical progress, but also through the terminology work itself.

Ampere = Amp.^{xx}

Acronyms are the most common type of abbreviations. They are written in capital letters and constructed on the basis of the first letters of the parts of multi-word units or compounds.

Automatisches Testsystem = ATS^{xxi}

Thus they can be calculated automatically for the terms, but produce a high degree of homonymy in a terminological base. In fact, this is the problem for acronyms. To avoid the problem, one could use second and third letters of some parts to distinguish between the different abbreviated words. It could also be interesting to produce a pronounceable acronym by this way [International 1995].

2 Cyclic term construction

Through the compounding process in German, a term can easily be used to construct a new term,

i.e. for subspecifying a hyponym or for designing a process the term is involved in. In this case, all of the variants described above could theoretically be used, including the abbreviated form or even the separated, 'English' compounding.

Schiebehebedach vs. Stahlschiebe-Hebedach^{xxii}

When checking the consistency of a term bank, tools also have to be sensitive to this cyclic nature of term construction and check that a non-approved term variant is not used inside a more complex term.

3 Multilingual Inconsistencies

Multilingual inconsistencies are linked to the one-to-many and many-to-one translation problem. They occur when a rigid, consistent terminology is only developed for the source language, and the target equivalents are added by different translators at various times and places through the translation process. Multilingual inconsistencies can be avoided when a terminological base is constructed from scratch for all languages involved. Then, translation relations for terms included in the terminology bank should only present one-to-one relations, except for some clearly specified homonyms.

Multilingual inconsistencies, however, can be useful while searching for variant candidates. As described in the section 'Real Synonyms', they can help to propose synonyms and also to identify homonyms which should be assigned only one meaning. Also they often help to find out candidates for hyponymous relationships.

4 An integrated tool for checking terminological inconsistencies

On the basis of the nature and form of variants described above, we propose an integrated solution for checking terminological inconsistencies. The proposed approach has been partly implemented in the MULTILINT prototype [Haller 1996], a toolbox for the technical author which checks his texts wrt. correct orthography and grammar, stylistic guidelines and the consistent use of abbreviations and terminology.

The kernel for all checking tools is a powerful morphologic analyser [Maas 1996] which is able to abstract away from:

- formatting differences (small vs. capitalised letters, hyphen ...)
- orthographic differences (stem variants for some consonants ...)
- inflectional differences (suffixes, German Umlaute,..., 'Fuge')
- derivational differences (suffixes ...)

For all terms in MULTILINT, the official variant is stored together with its abstract form resulting from the morphological analysis. An incoming text is analyzed towards its abstract form and then matched with the abstract forms of the base.

TEXT: *Anziehungsmoment* = {*ls=an_**\$ziehen#moment*}

BASE: *Anzugsmoment* = {*ls=an_**\$ziehen#moment*}

After checking, non-approved variants are highlighted in the text and the official term is displayed to the writer.

The same control can be applied to the texts internally. Here, it does not control the use of a specific terminology, but the consistent use of the terms found in the text. In this way, the consistency check can also be applied if no terminology is available/obligatory for the writer.

For enabling the checking tool to detect all variants described above, further developments are necessary: The checking of syntactic variants, elliptic compounds and cyclic compounding involves further transformation of the abstract form (i.e. by metarules [Jacquemin and Royauté 1994]). For the detection of real synonyms, a slot for links should be opened in the terminological base. Another slot would contain the allowed abbreviations.

During checking, non-approved variants and words not found in the terminology should be collected together with their context. Statistic memory than can trace the importance of each variant, helping the terminologist to correct his base towards the real world use of term variants or to add new terms to his base. By this way, a term bank could also be built from scratch out of machine readable corpora of the subject field.

Internal consistency of the base itself should be controlled and corrected wrt. the following rules:

- Each term should occur only once.
- A term should not be a linguistic variant of another term.

- The construction of terms should be consistent towards rules for hyphenating.
- Tendencies of analogy should be followed.
- An acronym should be constructed on the basis of the compound parts.
- An abbreviation should not be homonymic, if possible.
- Non-approved variants should not be used to construct a new term.

If the term bank should also be used in translation, a parallel term bank has to be added for each target language, where for each source language term there exists a parallel target language term. The parallel term bank has to be submitted to the same consistency constraints as the source term bank.

5 Conclusion

The experimental implementation has shown that useful, relevant results can be achieved by the proposed approach. The users of the prototype, the technical writers and editors, got aware of the problem and are willing to adapt their documents to the underlying terminology. But they also got aware of the inconsistencies the terminology in fact contained.

The advantages of the proposed approach lie in the possibility of applying the consistency check with or without an existing terminology to a text, or to a terminology directly. It thus can be used independently of the terminological situation inside the company.

References

- [Daille 1994] Beatrice Daille 1994. Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres lexicales, PhD Thesis, URL: http://talana.linguist.jussieu.fr/Presentation/pub_the_se.html.
- [Ehrlich 1998] Ute Ehrlich 1998. Automatic Extraction of a Unique Terminology Based on Multilingual Corpus and Dictionary. In: Proceedings of 1. International Conference on Language Retrieval and Evaluation, Granada, p. 691-696.
- [Haller 1996] Johann Haller. 1996. MULTILINT, A Technical Documentation System with Multilingual Intelligence. In *Translating and the Computer 18*, London. Aslib, The Association for Information Management, Information House.
- [International 1995] SAE International. 1995. SAEJ 1930 - surface vehicle recommended practice. Technical report, The Engineering Society For Advancing Mobility Land Sea Air and Space, Warrendale, PA, issued 9.1991, revised 9.1995.
- [Jacquemin and Royauté1994] C. Jacquemin and J. Royauté 1994. Retrieving terms and their variants in a lexicalised unification-based framework. In *17th International ACM-SIGIR Conference on Research and Development in Information Retrieval*.
- [Königer 1997] Paul Königer. 1997. Dynamik technisch geprägter Sprache. In Rüdiger Weingarten, editor, *Sprachwandel durch Computer*. Westdeutscher Verlag, Opladen.
- [Maas 1996] Heinz-Dieter Maas. 1996. MPRO - Ein System zur Analyse und Synthese deutscher Wörter. In Roland Hausser, editor, *Linguistische Verifikation, Sprache und Information*. Max Niemeyer Verlag, Tübingen.
- [Streiter and Schmidt-Wigger 1995] Oliver Streiter and Antje Schmidt-Wigger. 1995. Patterns of derivation. In *TMI-95*. URL: <http://www.iai.uni-sb.de/cat2/docs.html>.

ⁱ Primary instinct vs. The urine stinks.

ⁱⁱ Autogenous welding

ⁱⁱⁱ Airbag module

^{iv} Ammeter

^v Differential hose

^{vi} Tightening torque

^{vii} Exhaust pipe

^{viii} Exhaust gas recirculation

^{ix} Camshaft arrangement

^x Storage reservoir vs. Dust corner

^{xi} Aggregate protective plate

^{xii} to clip on

^{xiii} Automatic transmission

^{xiv} Bleeder valve vs. Bleeding valve

^{xv} Case hardening

^{xvi} Anti-drainback valve vs. Non-return valve

^{xvii} Pushing tool vs. Pusher

^{xviii} Fresh air door motor

^{xix} Increase

^{xx} Ampere

^{xxi} Automatic test system

^{xxii} Slide/tilt sunroof vs Steel slide/tilt sunroof