Published in LSP & Professional Communication Volume 6, Number 1, April 2006 - ISSN 1601-1929
© DSFF / LSP Centre

Building a Controlled Language Lexicon for Danish

Margrethe H. Møller
& Ellen Christoffersen
University of Southern Denmark

1. Introduction

A controlled language (CL) is *"[…] an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style."* (Nyberg et al. 2003). CLs are used primarily for technical texts, e.g. user manuals, and although most CLs share a common core, they need to be tailored for each organisation or enterprise. The objective is to improve the quality of the texts. Ambiguity and complexity are reduced with a view to making texts easier to read, understand and translate. As an additional benefit, controlled texts will make the use of language technology more efficient. In translation memory systems, the number of hits (the leverage) is increased, and in machine translation systems, the quality of raw translations is improved. As the type of CL-rules needed for texts intended for human readers and texts intended for machine processing may differ, a distinction is made between human-oriented CLs and machine-oriented CLs.

So far, most CLs have been designed for English documentation, the most famous example being ASD Simplified Technical English (ASD-STE100$^{\text{TM}}$), previously known as AECMA Simplified English. ASD Simplified Technical English is a CL for aircraft-maintenance documentation which has become an international standard within the aerospace industry. In Denmark, Center for Sprogteknologi (Centre for Language Technology) at the University of Copenhagen has been working with controlled languages for two Danish enterprises using English as their corporate language, within the VID project (Henriksen et al. 2004).

In the project "Controlled Language for Danish Enterprises", we are investigating methods for designing controlled languages for Danish, and testing these methods on Danish enterprise texts. Another objective of the project is to develop Danish modules for a CL-checker. A CL-checker is a specialised grammar and style checker which may assist technical writers using the CL of the enterprise when producing and revising technical documents.

The present paper is a preliminary result of our work with texts supplied by industrial partners and focuses on problems related to building the CL-lexicon.

2. The CL-lexicon: One word – one meaning?

The CL-lexicon consists of approved and non-approved words, the latter carrying a reference to an approved word to be used by the technical writer. We talk about "words" and not "terms", because the controlled language may restrict not only specialized terminology, but also general vocabulary. We define words as single words as well as multiwords.

As in terminology work, the ideal is a one-to-one correspondence between words and concepts: "one word – one meaning" (Felber 1993:83; Nyberg et al. 2003:246). In controlled language, this means that synonyms and spelling variants (several words for one and the same concept) and homonyms (one word for several concepts) are banned. Typically, it also means that words may only be used as one part of speech, which is given in the CL-lexicon. Thus, according to ASD Simplified Technical English, the English word *check* may only be used as a noun, and not as a verb. Consequently, the sentence "Check the hydraulic system" must be rephrased into "Do a check of the hydraulic system".

When planning a CL-lexicon for an enterprise, it is important to take the text types and the intended readers into consideration. ASD Simplified Technical English was constructed for aircraft-maintenance documentation for readers who do not have English as their mother tongue, and who are easily confused by complex sentence structures and by the number of meanings and synonyms which English words can have. This motivates strict adherence to the principle of "one word - one meaning". On the other hand, in our project, we work with technical specifications and instructions in Danish intended for readers who are native speakers. Our project partners would like terminological consistency, but they would also like to adhere to the conventions – the "technical register" – of the subject areas and the text types in question. At the lexical level, this means that we may have to allow some synonyms and homonyms.

3. Creating a CL-lexicon

The lexicon of a CL-checker serves two purposes: that of an analysis lexicon and that of a CL-lexicon.

Serving the purpose of an analysis lexicon, it has to supply the parser with the linguistic information necessary to carry out grammatical analysis of texts, and it has to include all words which may appear in the texts to be analyzed.

Serving the purpose of a CL-lexicon, it has to supply the CL-checker with the information necessary to trigger lexical error messages, i.e. information about unapproved words, supplemented by references to approved words to be used instead, short definitions and examples.

The CL-lexicon may either include **all approved words** (and unapproved words, with a reference to an approved word to be used instead) in the documents of the enterprise in question, i.e. a positive list, or it may include **only the unapproved words** (with a reference to an approved word to be used instead), i.e. a negative list.

If a positive-list strategy is chosen, the CL-checker will display error messages a) when the technical writer uses unapproved words, and b) when he or she uses words that are not in the lexicon (unknown words).

If a negative-list strategy is chosen, the CL-checker will display error messages when the technical writer uses words which have been registered as unapproved, but not when he or she uses unknown words.

The ASD-STE CL-lexicon uses the positive-list strategy. However, it is a huge task to ensure completeness of the CL-lexicon, and error messages triggered by unknown words which should actually have been in the CL-lexicon as approved words, will annoy the user. Therefore, at least in a first stage, we have settled for a negative list in our project.

In addition to approved and unapproved words, the CL-lexicon may contain definitions and examples. The organization of the information may differ; in Figure 1 is an example from AECMA Simplified English (now ASD Simplified Technical English), cited from Nyberg et al. (2003:246).

Approved word	*prevent* (v)
Definition	To make sure that something does not occur
Example	*Attach the hoses to the fuselage to prevent their movement.*
Unapproved word	*preventive* (adj)
Approved alternative	*prevent* (v)
Unapproved example	*This is a corrosion preventive measure.*
Approved rewrite	*This prevents corrosion.*
Approved word	*right* (adj)
Definition	On the east side when you look north
Example	*Do a flow check of the pump in the right wing tank.*
Unapproved word	*right-hand* (adj)
Approved alternative	*right* (adj)
Unapproved example	*The fuel connector is in the right-hand wing.*
Approved rewrite	*The fuel connector is in the right wing.*

Figure 1. Examples of Simplified English: *prevent* vs. *preventive* and *right* vs. *right-hand*

Definitions and examples are short because they have to be quick and easy to read and must fit into the message window of a CL-checker.

4. Types of term variants and their treatment in the CL

Data for the CL-lexicon may be collected in various ways. Ideally, the enterprise has a terminological database where all relevant concepts have been defined, a unique, preferred term has been selected for each concept, and any synonyms, spelling and syntactic variants and deprecated terms have been registered. In that case, the linguist building the CL-lexicon may take over the information, perhaps shortening the definition into a format suitable for a CL-checker and adding unapproved examples and approved rewrites.

Less ideally, the linguist building the CL-lexicon may have to collect information from various sources, e.g. a corpus of texts of the type the enterprise wants to control, and a descriptive terminological database in which no distinction has been made between preferred terms and synonyms, and in which, perhaps, definitions are missing.

The enterprise's own experts (engineers, terminologists, translators) always constitute a very important source when it comes to verifying suggestions for synonyms and homonyms and deciding which terms should be approved or unapproved. That also involves deciding how restrictive the CL-lexicon should be.

4.1. Synonyms and homonyms

CL-lexicons are built by linguists and terminologists. Synonyms and homonyms may be difficult to identify for a non-subject-field-expert linguist. One method may be to study bilingual material, i.e. texts and their translations, or a bilingual term bank.

Gasser (2004:252f.) reports about using a term-extraction tool for bilingual term extraction from parallel texts (English-German) in order to identify possible synonyms in German: if several German terms were suggested for one English term, the German terms were potentially synonymous. The final decision as to whether they were actually synonyms was left to an expert of the subject domain.

In our project, we used a bilingual term bank to detect possible synonyms and homonyms in the Danish texts of one enterprise. The term bank had English as the pivot language, i.e. all definitions were given in English, and each concept had one English term and one or more Danish terms attached to it. The term bank was descriptive rather than prescriptive, i.e., although some Danish terms were marked as deprecated terms, no systematic attempt had been made to choose one preferred Danish term among several synonyms – probably because the term base was conceived for translation purposes, not for normative purposes. Consequently, there were many synonymous and a number of homonymous Danish terms in the term bank, see examples in Figure 2 and 3:

English term	Danish equivalents in the term bank
shaft end	*akseltap* *akselende*

Figure 2: Example of Danish synonyms in the term bank

English terms	Danish equivalents in the term bank
alarm signal	*alarm*
alarm unit	*alarm*

Figure 3: Example of Danish homonyms in the term bank

In these examples, the CL-lexicon could specify the Danish term *akseltap* as the approved term and *akselende* as an unapproved term. Also, the Danish term *alarm* could be restricted to the English *alarm-signal* meaning, and the Danish term *alarmanlæg* could be specified as the approved term in the English *alarm-unit* meaning.

The term bank often had one Danish term for what should, according to the English definitions in the term bank, be two or more concepts. Thus, the Danish term *fejlmelding* (Eng. lit. ‘fault message’) had three English equivalents according to the type of signal, see Figure 4:

English terms and definitions	Danish equivalents in the term bank
fault reading (message in text or code in a display)	*fejlmelding*
fault indication (indication by means of indicator light)	*fejlmelding*
fault signal (signal sent via transmitters, contacts, relays etc. to external controllers or systems)	*fejlmelding*

Figure 4: Example of potential Danish homonyms in the term bank

This could be seen as homonymy in Danish, where the CL-principle of "one meaning – one word” would require three distinct terms in Danish. In many cases, however, a more precise description of the problem is that Danish uses a hypernym (a superordinate word) where English uses a hyponym (a subordinate word). If a language – or an enterprise or text type – does not need a certain distinction, i.e. several hyponyms, but prefers a hypernym, this cannot be said to violate the principle of "one meaning - one word”, and no attempt should be made in the Danish CL-lexicon to coin three distinctive terms for the three meanings of the Danish term *fejlmelding*.

In some cases, both the Danish term and the English term are homonymous, e.g. the Danish term *belastning* (Eng. *load*) covers three different concepts relevant to the texts produced by the enterprise in question, see Figure 5:

English terms and definitions	**Danish equivalents in the term bank**
load (the amount of work assigned to a machine or mechanical system)	*belastning*
load (the power absorbed from an electric circuit, i.e. the power output of an electric machine)	*belastning*
load (the weight supported by or the mechanical force applied to sth.)	*belastning*

Figure 5: Example of English and Danish homonyms in the term bank

These three concepts can be assigned to three different subject areas: mechanical engineering, electrical engineering and physics. As Felber notes, the one-to-one-relationship between terms and concepts should apply within a subject area (Felber 1993). In most cases, technical texts are a mixture of several subject areas, and in practice, it would hardly be possible to create new terms in order to disambiguate the three meanings of *belastning* or *load*. So, when homonyms come from different subject areas, they will often have to be accepted in the CL, and they should not trigger an error message from the CL-checker – in other words, the terms in question should simply be ignored by the CL-checker.

4.2. Other term variants

In addition to synonyms in the ordinary sense, there are a number of other types of alternative designations to concepts in technical texts – we will refer to them here as ‘term variants’ – which will be discussed below. These are spelling variants, syntactic variants and various types of compression.

Generally, human readers will have no problems with these types of term variants, although they may be annoyed by those variations which are in conflict with Danish spelling rules (see below), so a human-oriented CL may allow them.

But if the texts are intended not only for human readers, but also for computer-aided translation by means of machine translation systems or translation memory systems, spelling variants and syntactic variants and compressed forms will reduce the efficiency of the systems. In a machine translation system, all variants will have to be entered in the lexicon to be recognized, and compressed forms will be difficult for the system to interpret correctly. In a translation memory system, variants will reduce the hit rate of the system. Therefore, a machine-oriented CL should preferably choose one variant as the preferred one in order to ensure maximum lexical consistency in source texts.

Because these term variants are built of the elements also constituting the approved term, they can often be recognized automatically (Schmidt-Wigger 1999). Thus, if they are formed in a regularized way and the CL-checker has rules to recognize them, they will not have to be included in the CL-lexicon.

4.2.1. Spelling variants

According to Danish spelling rules, compounds can be written in one word, in exceptional cases with a hyphen. Compounds written with a space as in English constitute a frequent error type in Danish texts, see the following example from an electronics text:

> *buskommunikation*
> - term variant with hyphen: *bus-kommunikation*
> - term variant with space: **bus kommunikation*

In a CL-checker which is able to use syntactic analysis to rule out *bus* as a possible verb in the context, **bus kommunikation* can be recognized as an erroneous variant of the approved term *buskommunikation* and trigger an error message.

4.2.2. Syntactic variants

In Danish texts, compound terms may be varied according to several patterns, e.g. the following ones:

> *tætningsdiameter* (nominal compound, Eng. *seal diameter*)
> - term variant with genitive attribute: *tætningens diameter*
>   (Eng. lit. ’the of-the-seal diameter’)
> - term variant with prepositional modifier: *diameteren på tætningen*
>   (Eng. lit. ’the diameter of the seal’)
>
> *atmosfæretryk* (nominal compound, Eng. lit. ’athmosphere pressure’)
> - term variant with adjectival attribute: *atmosfærisk tryk*
>   (Eng. lit. ’athmospheric pressure’)

As mentioned above, these types of variants will not disturb the human reader, but they will reduce the efficiency of a translation system. Consequently, they should trigger an error message in a machine-oriented CL. It is not necessary to include them in the CL-lexicon, if the CL-checker has rules which can recognize them.

In lists and indexes, the adjectival pre-modifier is often put behind the head word:

> In running text: *metalimprægneret kul* (Eng. lit. ’metal-impregnated coal’)
> Term variant in list: *kul, metalimprægneret*

Also this type of term variant should be recognized by CL-checker rules, but it should not result in an error message unless the term is unapproved. A prerequisite for the correct treatment of this problem by the CL-checker would be that such elements are marked up as list or index elements in the text, and that the CL-checker is designed to use markup information.

4.2.3. Compression of terms: abbreviations, acronyms and codes

Term variants can be created by compressing long terms in various ways, e.g. by creating acronyms such as *DDT, A (Amp), V (Volt)* or by leaving out letters or syllables as in *lab (laboratory)* and *stagflation (stagnation + inflation).* Sager (1997:37) refers to this technique as compression, the purpose being to create short forms for frequent terms or to create new exclusive terms for long terms which might not be understood as terminological units.

In our texts, there are many examples of compression, e.g.

- types of material: *Krom-nikkel-molybdæn-stål* (Eng. lit. ‘chrome-nickel-molybdenum-steel’) - compressed form: *Cr-NiMo-stål,*
- components: *O-ringstætning med fast medbringer* (Eng. lit. ‘o-ring shaft seal with fixed seal driver’) - compressed form: *Type A.*

If the compressed form is not generally known, it is usually introduced together with the long form in the beginning of the text or the paragraph, and subsequently the compressed form is used.

In a controlled language, it would not be reasonable to enforce the principle of "one word - one meaning" by prescribing either the long term or the compressed term. Therefore the CL-checker should accept both, but it should be a general rule of the CL to write the long term in brackets the first time the compressed form is used.

Furthermore, the CL-checker should be able to recognize strings such as *DDT* and *Cr-NiMo* as acronyms and issue an error message if an acronym is not in the CL-lexicon and therefore potentially wrong.

4.2.4. Compression of terms: head words

Another way of compressing long terms is to mention them via their head word when they are repeated in a text. In other words, a superordinate term (a hypernym) is used as a substitute for a subordinate term (a hyponym). Göpferich (1998) refers to this type of compression as “spontane Abkürzungen” (‘spontaneous abbreviations’) and notes that in running text, exact terms like e.g. *Wasserpumpenzange mit Rillengleitgelenk* (Eng. lit. ‘water pump tongs with grooved joint’) are often too differentiated and difficult to handle.

If the modifiers included in the exact term are not necessary in the context, they will, according to Göpferich, hamper communication and reading speed. Therefore, they are often replaced by short compound words or, more frequently, head words.

Head words are often polysemous, but in actual texts, they are mostly disambiguated in context – e.g. by means of a headline.

As Göpferich notes, the short forms are often used in texts oriented towards man/machine-interaction, e.g. in manuals, and therefore, the principle that product-related concepts, e.g. the components of a car, should always be referred to by the same term, is rarely followed in practice (Göpferich 1996:388-390).

In a controlled language for the text types and the target groups we work with, it would not make sense to prescribe the consistent usage of the exact, long term. However, it should be a general rule of the CL to use the head word, only if it is clear from the context which concept is meant. This is an example of a rule which could probably not be checked by a CL-checker.

4.2.5. Term variants consisting of an approved term and a support noun

Most terms can be combined with generic nouns, which Reinhardt et al. (1992) refer to as support nouns.

We have identified different types of support nouns, e.g.:

- nouns indicating that the term is a superordinate term in a concept hierarchy, e.g. the Danish noun *type* (Eng. *type*) as in *tætningstype* (Eng. *seal type*)
- nouns relating to systems and components, e.g. the Danish noun *enhed* (Eng. *unit*) as in *kommunikationsenhed* (Eng. *communication unit*)
- nouns relating to dimensions, e.g. the Danish noun *størrelse* (Eng. *size*) as in *beholderstørrelse* (Eng. *container size*).

In a complete CL-lexicon, all possible combinations of approved terms and support nouns plus any unapproved variants should be registered. However, this would be very inefficient. Instead, the CL-checker should have grammar rules which can strip off the support noun and check the term as such. Thus, if the Danish *bus-kommunikation* (with a hyphen) is an unapproved term and *buskommunikation* is approved, then the support-noun compound *bus-kommunikationsenhed* (Eng. *bus communication unit*) should trigger an error message recommending *buskommunikationsenhed* instead.

5. Concluding remarks

When mentioned in connection with CL, the principle of “one meaning – one word” seems well-motivated. From a theoretical point of view, it is easy to understand that this principle will reduce lexical ambiguity in texts and make them easier to understand. In practice, however, there are difficulties. First, “one meaning” may be a superordinate concept in one language, enterprise or text type whereas in another language, enterprise or text type, in which more fine-grained distinctions are needed, “one meaning” may be a number of subordinate concepts. Second, homonyms in a text may come from different subject areas, in which case there may be no natural and acceptable way to assign different terms to the different concepts.

As noted before, a machine-oriented CL requires more restrictions than a human-oriented CL. When building a CL-lexicon, one goal must be to strike the balance between eliminating lexical ambiguities as far as possible and necessary and attaining a result that seems both acceptable and relevant to the enterprise and the readers in question. Rephrasing the sentence "Check the hydraulic system" into "Do a check of the hydraulic system" results in stilted language and may not be necessary for the purposes of a human-oriented CL. Likewise, eliminating homonyms from different subject areas may be difficult. On the other hand, the easy task, and the one which is most readily understood and accepted by enterprises, is to eliminate synonyms.

Bibliography

ASD-STE 100 TM
ASD Simplified Technical English Specification ASD-STE 100 TM:
A Guide for the Preparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language. Issue 3. January 2005.
Brussels: InfoVision.

Felber (1993)
Helmut Felber: *Allgemeine Terminologielehre und Wissenstechnik: Theoretische Grundlagen.* TermNet, Wien 1993.

Gasser (2004)
Yvonne Gasser: ”Wege zu einer standardisierten Unternehmensterminologie”, in: Susanne Göpferich / Jan Engberg (Hrsg.) *Qualität Fachsprachlicher Kommunikation*, Gunter Narr Verlag Tübingen, 2004 (pp. 235-259)

Göpferich (1998)
Susanne Göpferich: *Interkulturelles Technical Writing*, Gunter Narr Verlag Tübingen, 1998

Göpferich (1996)
Susanne Göpferich / Peter A. Schmitt:
”Begriff und adressatengerechte Benennung: Die Terminologiekomponente beim Technical Writing”, in: Hans P. Krings (Hrsg.): *Wissenschaftliche Grundlagen der technischen Dokumentation*, Gunter Narr Verlag Tübingen, 1996 (pp. 369-402)

Henriksen et al. (2003)
Lina Henriksen, Bart Jongejan and Bente Maegaard:
Kontrolleret sprog
Vid-rapport nr. 1, Center for Sprogteknologi, September 2003
(http://www.cst.dk/vid/public/index.html)

Nyberg et al. (2003)
Eric Nyberg, Teruko Mitamura and Willem-Olaf Huijsen:
Ch. 14 ”Controlled language for authoring and translation”
in: Harold Somers (ed.): *Computers and translation: a translator’s guide*
John Benjamins B.V., 2003

Reinhardt et al. (1992)
Werner Reinhardt, Claus Köhler and Gunter Neubert:
Deutsche Fachsprache der Technik
Hildesheim: Georg Olms Verlag , 1992

Schmidt-Wigger (1996)
Antje Schmidt-Wigger: ”Term Checking through Term Variation”, in: Peter Sandrini (ed.): *Terminologi and knowledge engineering: proceedings / TKE ’99*, TermNet 1999.

Wright-Budin (1997)
Sue Ellen Wright, Gerhard Budin (eds.): *Handbook of Terminology Management, Vol. I*, John Benjamins B.V. 1997.

***

Published in LSP & Professional Communication Volume 6, Number 1, April 2006 - ISSN 1601-1929
© DSFF / LSP Centre

ABSTRACT

Building a Controlled Language Lexicon for Danish

Margrethe H. Møller
& Ellen Christoffersen
University of Southern Denmark

A controlled language (CL) is a set of writing rules for the technical texts of an enterprise specifying constraints on lexicon, grammar and style with the purpose of reducing ambiguity, thus making texts easier to understand and process - by human readers and/or by translation systems. So far, most CLs have been designed for English documentation. The present paper is a preliminary result of our work within the project "Controlled Language for Danish Enterprises" and focuses on problems related to building the controlled-language lexicon, e.g. on whether it is possible to enforce the principle of “one word – one meaning”.

***