IVAR UTNE

What Should be Included in a Commercial Word Data Base, and Why?

The original title, which should be considered as synonym, was:

User considerations and market strategy as basis for profiling content in computational word bases, with special regard to the Norwegian Term Bank's data base NOT and a multilingual Norwegian word list

Abstract

In the article I present definitions of the concepts quality and quality assurance, which are basis for proposing a general definition of quality of word bases. The quality of word bases is the specifications that customer and producer agree upon, and includes linguistic and other user relevant considerations. This is exemplified with the revision work of a term record format, and work with a word base of everyday language.

1 Introduction

In this paper I will propose guidelines that could be worth aiming at in order to compile word bases that the users wish, need, and will pay for. The presentation of the guidelines will be based on a presentation of the quality assurance concept, which is becoming very important in the industry and the service sector. This will be exemplified with some word bases, i.e. computer based word lists and terminological data bases. I will emphasize that the presented proposals are not a complete solution, but I hope they will present some ideas for further work.

The ideas will be based on experience with terminological data bases at the Norwegian Term Bank (NT, Norsk termbank), and with work on word lists based on everyday language at the Department of Scandinavian Languages and Literature (Nordisk institutt) in cooperation with the Norwegian Computing Centre for the Humanities (NAVFs edb-senter for humanistisk forskning). I have be working in or in close contact with these institutions for some years, and base

Proceedings of NODALIDA 1989, pages 358-371

the descriptions on the current term record format on documents produced by personnel at NT. Much of these activities, and especially those of terminological work, have been based solely on support from the industry and users outside the University.

Note that the term *producer* in this presentation always will mean *producer of products and services*, and the term *product* will always mean *product and service.*

2 Word Base Quality — Quality Assurance

In order to design word bases and to get language services contracts including compilation of lists as part of them, we need to have a strategy to succeed in the market.

The main points for such a *market strategy* could be described as transfer of products and services to a market with regard to:

(a1) Users' wishes and needs

(a2) Identification of new paying user groups and their wishes and needs

(a3) Compare user wishes and needs with the needs of the researchers to gain cooperation and/or coordination in order to give more resources to research and to improve the product

(a4) Prices, which will not be further developed here

According to this I will concentrate on the quality of the products, which is the most important premise on an effective market strategy.

Quality of products (and services) is according to the International Standardization Organization defined as:

> The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs. (ISO 8402-1986:2)

More shortly it could be stated as:

> Accordance with what specifications the customer and producer have agreed upon.

Another related concept is *quality assurance* (QA). According to the same standard QA is defined as:

> All those planned and systematic actions necessary to provide adequate confidence that a product or service will satisfy given requirements for quality. (ISO 8402-1986:4)

While ISO stresses the relation between producer and customer, Norwegian Petroleum Directorate (Oljedirektoratet) stresses saftey for personnel (and society) and Norwegian Society for Quality (NFK, Norsk Forening for Kvalitet) in some publications also expresses the importance of profit as result of effective production routines.

Norwegian Petroleum Directorate states aspects of QA in *Regulations concerning the licensee's internal control in petroleum activities on the Norwegian Continental Shelf with comments* (Oljedirektoratet 1985a:3):

> The regulations will be applicable to worker protection and the worker environment within the scope of the Act concerning worker protection and worker environment (the Working Environment Act).
>
> The regulations will also be applicable to protection against pollution in petroleum activities within the scope of the Act concerning pollution and refuse (the Pollution Act).

The QA requirements related to worker environment and public considerations is not yet precisely formulated in the legislation and regulations. An instructive presentation of QA and worker environment is Hellesøy 1988 (Norwegian text).

Norwegian Society for Quality states their views on profit for instance in:

> the title of a booklet called *Profit by quality* (The Norwegian original title is: *Lønnsomhet ved kvalitet*), NFK 1987a.
>
> the heading "Correct performed quality assurance increases the productivity" (The Norwegian original text is: "Riktig utført kvalitetssikring øker produktiviteten") in another booklet (The Norwegian original title is Kvalitet og Kvalitetsstyring which means Quality and Quality Control, NFK 1987b).

This means that QA implies existence of systematically controlled routines to ensure that:

(b1) For the *producer-customer* relation: A product will be completed according to what is agreed beforehand, which implies routines to secure that the order is unambiguous and that there exists routines during production to ensure that the production acts towards the agreed goal.

(b2) For the *employers*: That the producer has as effective routines as possible, to ensure low cost for both parts (employer).

(b3) For the *employees*: The personnel implied work takes place without any lack of sequrity and according to work environment requirements.

(b4) For the *society*: There shall be no considerable risks for workers' safety and pollution, and the national (here: Norwegian) language shall have preference for the sake of safety (A requirement stated by Norwegian Petroleum Directorate in Oljedirektoratet 1985b:4 may possibly be interpreted like this.)

In order to apply this on work with word bases we have to decide what is central aspects of word base quality and how customer and producer can agree on what alternative espressions or style and other non-linguistic specifications that are to be defined as goals.

The specifications for word base quality should consider at least language, user relevant information and user interface. The consideration may include different values, e.g. that cultural considerations and rhytmic language may be of low or no importance; more in detail:

(c1) Good and communicative *language*

Good style/performance (aesthetic and cultural considerations with possible consequences for economics)

- Rhytmic or harmonius language
- Cultural considerations, tradition of "good" written language
- Subcultural considerations, e.g. firm or other local standards

Effective information transfer (economic and administrative considerations) means:

- Requirements for the selection and formation of terms according to ISO 704-1987(E):12–13 are that the terms should be:
  - linguistically correct

    "the term should conform to the norms of language in question" (op.cit.), e.g. letters (*ks* instead of *x*) and inflection paradigms
  - accurate, or motivated

    "the term should reflect, as far as possible, the characteristics of the concept which are given in the definition" (op.cit.), e.g. *magnetic tape* which is defined as "carrier of a magnetic recording, having the form of a tape" (based on op.cit.)
  - concise

    i.e. preciseness
  - permit, if possible, the formation of derivatives

    "alcohol — alcoholic, alcoholism, alcoholize" (op.cit.)
  - monosemous, if the terms are considered for standardization

    there shold be only one standardized term for a concept, and only one concept for one term
- Native (e.g. Norwegian) versus foreign (e.g. English) expressions
  - a native expression may be preferable because of accuracy and because it is linguistically correct (in the native language)
  - a foreign expression may be preferable in international communication, especially for abbreviations, formulas and symbols, but also for subjects with international traditions like chemistry

- "noise free" or neutral expressions
  - * not stigmatizing in the actual subculture
  - * related to relevant tradition

(c2) *User relevant* information

Reference to standard/authority documents/publications, i.e. controlling/ prescriptive documents (publications from language councils or academies, standards)

New use of symbols and other expressions, which usually are not included in dictionaries

Additional consensus with subject/user defined groups

(c3) *User interface* in a data base system

Transportable program and data, including copy protection

Stratified selection of information related to the purpose

Simple and logical user dialogue

Frequent update, especially when language services interact in multipurpose projects

For the implied interest groups in the collection of QA definitions listed above this means for:

(d1) *Customers*:

To decide what will be the goals we have to clarify and coordinate the needs and wishes of the customers with the word base knowledge/expertise. Central interfaces may be listed like this:

- What the users want and need
- What the users want, but don't need, e.g.:
  - * alternative synonyms in the target language
- What the users need, but don't ask for, e.g.:
  - * consistent terminology without use of synonyms in the same language
  - * consequent use of substandards (-norms), such as British English without US forms
  - * defined subsets of Norwegian
  - * accommodation to existing standards and regulations

In order to decide the quality requirements of the product the customer and producer must be in dialogue:

- To choose and design expressions according to the linguistic requirements above, cf. (c1).
  To choose information categories (types)

  - * In word lists: the selection of languages, references, definitions etc.
  - * In thesauri: any registration of deleted subject words/concepts after revision of an old thesaurus
- – To design the information according to needs and subject knowledge
  - * In terminological data bases: should the additional synonyms, references, or context excerpts be left out
  - * In word lists: e.g. the selection of variants (style) within Norwegian-Nynorsk

(d2) *Employers*:

Backup routines

Effective tools

Reference information

Housekeeping of information for other projects and for research

(d3) *Employees*:

Effective and user friendly tools for automatizing, to get rid of boring work and have overview and control

Proper procedure descriptions

Easy access to relevant information

Unambiguous references

(d4) *Society*:

Consideration of culture values, according to the prevalent values

Public considerations, e.g. proper language for effective communication which may imply a precondition for safety and good health

In the next section I will apply these definitions and principles to the word base work at my institution. For a discussion of quality assurance for language work in general it is referred to Utne 1987 (Norwegian text).

3 Record Format for a Terminological Data Base

The terminological data bases at NT include mainly bilingual dictionaries and also hierarchically structured thesauri. Both these data types have been presented at *Symposium for datamatstøttet terminologi og leksikografi* in 1985 and 1987 (Utne 1986, Utne 1988), and revised descriptions are distributed from NT (free of cost). An excerpt from the present description (NT 1989) is included in the appendix.

The terminological work has usually been part of multipurpose projects aiming at cost effective document production. Language services have in most occasions been looked upon as a totality in this context. It is very unusual that the supporters give grants for development of dictionaries. For the bilingual dictionaries the exceptions have been projects for oil companies, mostly in the period 1984–86. The update of the data is partly financed by subscription on continuously updated copies. NT is maintenance contractor for a thesaurus developed by NT in 1985–86.

A radical restructuring of the data base format for the terminological data base was performed in 1988. A presentation of this restructuring while in process is presented in Ebeling and Utne 1988.

In the following I will examplify application of the quality assurance concept on the restructuring of this data base. This application is based on my point of view which is partly from outside. The process was in practice not planned and performed according to the principles presented below, but has in fact followed most of them. My presentation will be an application connected to a possible and realistic strategy.

Lots of considerations about quality of linguistic expressions are not dependant of this revision of format and are therefore not listed below. The *general leading principles* for the revision have been to gain better quality of:

(e1) *Language*, by:

Error free data to a larger extent

(e2) *User relevance*, by:

Unambiguous classification of greater part of the data

More flexible introduction of new categories/classification

More stress on standardization

(e3) *User interface*, by:

More flexible presentation, excerption and introduction of different and more fine grained data types

Improved distribution

This implied some more applied leading principles (without explaining the links to the general principles in detail):

(f1) More strictly constructed hierarchical structure

Goal: To make more flexible excerpts of subsets of data possible, i.e. to extract different combinations of fields and parts of fields, and introduce more unambiguous links between a term and its abbreviation, reference or its context

Solution: A more strictly hierarchical structuring inside the term record, which means that:

- Abbreviations, contractions, symbols and formulas are unambiguously bound to their full forms and not only to their concept
- References are unambiguously bound to their term (or abbreviation etc.), context-excerpts, definitions etc
- Contexts are unambiguously bound to their terms etc
- Comments are unambiguously bound to all relevant kinds of fields

(f2) More formal notation language

Goal: Be more consequent in registration of all kinds of information, i.e. more formal and restricted formats for the information types, to make it easier to extract expressions from the same page or pages in a specified source or to extract terms from a specified subject area

Solution: Defined vocabulary (e.g. titles) and syntax for references
Program tools according to the goal

(f3) A productive notation system to include user relevant information

Goal: Be more flexible in introducing of new information classes, i.e. the possibility of introducing new information classes according to simple and defined routines

Solution: Supplementary information not traditionally included in dictionaries, is introduced because of the usefulness for the target groups. This means for instance that there is introduced a distinction between general abbreviations, project specific abbreviations, symbols, classification codes and formulas. Some of the categories are:

- International standardized symbol; shown, referred to an illustration, and/or described
- Chemical/mathematical formulas, e.g. NH_4HSO_3 for *ammonium hydrogensulfite*
- International classification codes, e.g. according to UN (United Nations), EC (European Community), CAS (Chemical Abstract Services) and widely used national standards
- Trade names for products of the concept
- References to official documents, e.g. standards, registers, laws and regulations
- Hazard classification, e.g. fire, poison
- References to figures which illustrate the concepts
- Area of application (not uniquely for this data base), for housekeeping, collecting subsets and for indicating meaning.

(f4) Existence of programs and other routines for error checks

Goal: Make update more free from errors, i.e. facilitate more throughout error checks

Solution: Program tools according to the goal

(f5) Standardization

Goal: As it has been a goal for years, one concept is one record

Solution: Synonyoms in the same language are collected in the same record.

Goal: As it has been a goal for years, there should be only one preferred full form term for each language.

Solution: This is called main term, and the others are called synonyms or deprecated terms.

Goal: The format should also include registration of alternative standards, like former main term, and main terms in other standards.

Solution: Introduction of fields for approval date and scope of a standard. This includes also out of date standards.

(f6) Transportable program and data

Goal: Program and data files produced for different machines and media

Goal: Routines for copy protection

Solution: Program tools for both these goals.

4 Word Lists

The development of word lists of everyday language was initiated by the researchers at the University about 20 years ago. The project is a cooperation between NT and Norwegian Computing Centre for the Humanities. Further development is financed by sale. Customers are mostly software houses, institutions and firms (graphic industry and newspapers) which are able to include the lists in existing software or to develop standalone programs.

The quality of these lists is partly based on general needs for spelling lists, special and general needs for lists with hyphenation marks (with special reference to different levels) and our estimation of user needs for other lists with grammatical information and lists of word parts. The development of spelling lists has been based on our own assumption that a frequency based and general correspondance vocabulary would suit the users' needs best. The development of lists with grammatical information and word composita is based on point (a3) in market strategy above, i.e. combination of researchers' and customers' needs. The development of lists with substandards also accommodates needs and wishes for accommodation to substandards and more personal ways of writing.

The existing word lists include spelling lists based on general need and a list with hyphenation-marks based on special needs:

- *Spelling word lists* for Norwegian-Bokmål and for Norwegian-Nynorsk, including a collection of high frequency word forms without any further coding.
- Word lists with *hyphenation marks*, based on special need, but also with general application. The marks express different levels, so that the different types of marks borders between:
  - *Compounds*, like data+base (which is written as one word in Norwegian)
  - *Pre- and suffixes* and the rest of the word like ex=plos=ion (Norwegian: eks=plo=sjon)
  - *Inflection morphemes*, like the definite plural in bil//ene (= the car//s)

The further development includes partly further refinement of the lists above and development of new word list types partly based on a combination of what is asked for in new products for text processing or data base tools with dialogues in natural language, and work with machine-aided translation with what is asked for, cf. combination of customer needs and wishes with the researchers' as market strategy. That means lists containing:

- *Word composita*, e.g. parts of latin and greek loan words.
  - The machine aided system may use lists like this to translate loan words which have identical parts, but related inflectional paradigms.
  - In text processing systems this may be used as part of hyphenation programs, and combined with supplementary rules partly also as part of spell checkers.
- Word lists including *grammatical information*, like part of speech and inflectional paradigms.
  - The most important part of a machine aided translation.
  - In text processing systems and especially in dialogue based data base tools this may be used in programs that are based on simpler syntax analysis or calculations of possible part of speech for words in a text string.

And as a combination with work to systematize *substandards* in the written Norwegian languages:

> Word lists with grammatical information that classifies the words and word forms (often inflectional paradigms) as moderate (as different as possible) and radical (approaching to each other) within the two official Norwegian languages. This is of importance to writers who need language checks to profile their writing closer to such

a substandard. This implies a great variety of possible written languages which computer systems should be able to control. The exact definitions of each such language are not objectively stated, but are to some extent a matter of personal or user groups decisions. Table 1 and 2 in Utne 1989 (paper at *Nordiske datalingvistikdage 1989*) present some examples of this diversity. Some of the examples are repeated in Table 1 below. A further explanation of the language situation is presented in Utne 1989.

Language		porridge	line	problems	boys
Mod.	Norw.-Bokmål:	grøt	linje	problemer	gutter
Rad.	Norw.-Bokmål:	graut	linje	problem	gutter
Rad.	Norw.-Nynorsk:	graut	linje	problem	gutar
Mod.	Norw.-Nynorsk:	graut	line	problem	gutar

Table 1. Spelling and inflection in Norwegian
(Mod. = Moderate, Rad. = Radical)

To some extent this substandard works as if there are slightly different written sublanguages inside each of the two official Norwegian languages. There is also some tradition for unofficial written standards, e.g. one more moderate than Norwegian-Nynorsk called Conservative Norwegian-Nynorsk, another more moderate than Norwegian-Bokmål called Conservative Norwegian-Bokmål (Norwegian: Riksmål) and a third one between the two official languages which is sometimes called Pan-Norwegian (Norwegian: Samnorsk). Of these three Conservative Norwegian-Bokmål has the widest use with its use in at least one of the most widespread newspapers, in lots of books and publications every year.

The list containing a diversity a form alternatives within each of the language and also to some extent other unofficial language standards can be considered as a multilingual dictionary. This total word base concept, which at the time beeing contains between 20 and 30 000 entries (which can be inflected according to different alternatives) is the base of such a multilingual Norwegian word base. This base will be developed both for research, included machine aided translation, and for commercial applications.

Other possibilities not worked out in detail yet are for instance:

Synonyms and words with related meaning

Deprecated words and expressions, and their substitutions

Lots of deprecated expressions in Norwegian-Bokmål have common or near related preferred words both in Norwegian-Bokmål and Norwegian-Nynorsk.

5 Conclusion

Through this discussion of commercial word bases I have formulated what are the concepts quality and quality assurance, and proposed a general definition of quality of word bases. While quality is the part of the product, quality assurance is the procedures to secure quality and also consider the interests of employer, employee and the society. The quality of word bases is the coordinated specification that customer and producer agree upon, and includes linguistic and other user relevant considerations.

The examplifications from the work at the University of Bergen concern the revision of a term record format and the work with a word base of every day language. The work with format revision emphasized language from the view of errors cheks, user relevance, and user interface. The work with word bases emphasized an ongoing work with a multilingual Norwegian word base which includes form variants and also unofficial languages. In the work with this word base there are combined interests between research and profit.

References

Ebeling, Jarle and Ivar Utne. 1988: New Record Format at The Norwegian Term Bank. *Nordisk tidsskrift for fagspråk og terminologi* (Nordic Journal of L.S.P. and Terminology), vol 6, no 1:7–14. ISSN 0108-77891

Hellesøy, Odd H. 1988: Kvalitetsstyring av arbeidsmiljø? (= Quality control of worker environment?). *Nordisk Ergonomi* Vol 6, no 1:11–16.

ISO (International Standardization Organization) 704-1987 (E): *Principles and methods of terminology.*

ISO (International Standardization Organization) 8402-1986 (E): *Quality — Vocabulary.*

NFK (Norsk Forening for Kvalitet) 1987a: *Lønnsomhet ved kvalitet.* Booklet.

NFK (Norsk Forening for Kvalitet) 1987b: *Kvalitet og kvalitetsstyring.* Booklet.

NT (Norsk termbank) 1989: *"NOT" User's Guide.* Ver. 1989-05-23. Booklet.

Oljedirektoratet (Norwegian Petroleum Directorate) 1985a: *Forskrift om rettighetshavers internkontroll i petroleumsvirksomheten på norsk kontinentalsokkel med kommentarer. Regulations concerning the licensee's internal control in petroleum activeties on the Norwegian Continental Shelf with comments.* ISBN 82-7257-183-8

Oljedirektoratet (Norwegian Petroleum Directorate) 1985b: *Forskrift om sikkerhet m.v. til lov om petroleumsvirksomhet. Regulations concerning safety in exploration, exploration drilling and recovery of petroleum deposits, etc.* ISBN 82-7257-186-2

Utne, Ivar. 1986: Terminologidata på mikromaskin ved Norsk termbank. *Symposium om datorstödd terminologi och lexikografi i Helsingfors den 13 og 14 december 1985*:52–58. Centralen för Teknisk Terminologi. Helsingfors 1986.

Utne, Ivar. 1987: Nye perspektiver på språklig kvalitet — introduksjon til kvalitetssikring for språklig arbeid. (New perspectives on language quality — Introduction to quality for language Work.) *Nordisk tidsskrift for fagspråk og terminologi* (Nordic Journal of L.S.P. and Terminology), vol. 5, no 1:14–21. (Norwegian text with English summary.) ISSN 0108-77891

Utne, Ivar. 1988: Terminologi, arbeidsinstrukser og lagerstyring — om kodeuttrykk i fagspråk. *Nordiske Datalingvistikdage og Symposium for datamatstøttet leksikografi og terminologi 1987.* Proceedings:273–285. Institut for Datalingvistik, Handelshøjskolen i København. Copenhagen.

Utne, Ivar. 1989: Machine aided translation between the two Norwegian languages Norwegian-Bokmål and Norwegian-Nynorsk. *Nordiske datalingvistikdage 1989.* Reykjavík. In press.

Strømgaten 53
N-5007 BERGEN
Norway

Appendix

From *"NOT" User's Guide*:2, ver. 1989-05-23

The Record Format

The terms in this data base are organized in concept records, which consist of one Norwegian section, one English section and one section that is common to both languages.

The Norwegian section consists of the following fields:

N	hvd	=	Norwegian main term (recommended for use)
N	syn	=	Norwegian synonym to the same concept (to be avoided)
N	fra	=	Norwegian synonym not recommended for use (must not be used)
N	krt	=	Norwegian contraction (used for reasons of space, as for instance in screen pictures, drawings, signs etc.)
N	frk	=	Norwegian abbreviation
N	def	=	Norwegian definition
	kon	=	context of term
	kom	=	comment to term
	ref	=	reference of term
	fig	=	reference to figure

The English section contains the following fields:

E	hvd	=	English main term
E	syn	=	English synonym to the same concept (to be avoided)
E	fra	=	English synonym not recommended for use (must not be used)
E	krt	=	English contraction (used for reasons of space, as for instance in screen pictures, drawings, signs etc.)
E	frk	=	English abbreviation
E	def	=	English definition
	kon	=	context of term
	kom	=	comment to term
	geo	=	geographical distribution of term
	ref	=	reference of term
	fig	=	reference to figure

The common section consists of information about the concept as a whole:

	brk	=	area of application
	sbl	=	international standardized symbol
	fml	=	chemical/mathematical formula
.	nr	=	national and international numbers
.	nvn	=	trade name
.	kom	=	comment to the concept as a whole
	kon	=	context of symbol, formula etc.
	kom	=	comment to symbol, formula etc.
	ref	=	reference of symbol, formula etc.
	fig	=	reference to figure of symbol, formula etc.

It is possible to introduce new fields at all levels.